Cupid's Invisible Hand:

Social Surplus and Identification in Matching Models

Alfred Galichon[*] **Bernard Salanié**[†]

June 1, 2021[‡]

Abstract

We investigate a model of one-to-one matching with transferable utility and general unobserved heterogeneity. Under a separability assumption that generalizes Choo and Siow (2006), we first show that the equilibrium matching maximizes a social gain function that trades off exploiting complementarities in observable characteristics and matching on unobserved characteristics. We use this result to derive simple closed-form formulæ that identify the joint matching surplus and the equilibrium utilities of all participants, given any known distribution of unobserved heterogeneity. We provide efficient algorithms to compute the stable matching and to estimate parametric versions of the model. Finally, we revisit Choo and Siow's empirical application to illustrate the potential of our more general approach.

[*] Economics and Mathematics Departments, New York University, and Economics department, Sciences Po; e-mail: ag133@nyu.edu.

[†] Department of Economics, Columbia University; e-mail: bsalanie@columbia.edu.

[‡]This paper builds on and very significantly extends our earlier discussion paper Galichon and Salanié (2010), which is now obsolete. The authors are grateful to Pierre-André Chiappori, Eugene Choo, Chris Conlon, Facundo Danza, Gautam Gowrisankaran, Leon Guzman lizardo, Jim Heckman, Sonia Jaffe, Robert McCann, Jean-Marc Robin, Aloysius Siow, the editor and referees and many seminar participants for very useful comments and discussions. Part of the research underlying this paper was done when Galichon was visiting the University of Chicago Booth School of Business and Columbia University, and when Salanié was visiting the Toulouse School of Economics. Galichon thanks the Alliance program for its support, and Salanié thanks the Georges Meyer endowment. Galichon's research has received funding from NSF DMS-1716489, and ERC grants FP7–313699 and CoG–86627.

Introduction

Since the seminal contribution of Becker (1973), many economists have modeled the marriage market as a matching problem. When utility is perfectly transferable, each potential match generates a marital surplus. The distributions of tastes and of desirable characteristics determine equilibrium shadow prices, which in turn explain how partners share the marital surplus in any realized match. This insight is not specific to the marriage market: it characterizes the "assignment game" of Shapley and Shubik (1972), i.e. models of matching with transferable utilities. Family economics makes extensive use of this class of models; we refer the reader to the recent book by Chiappori (2017). Matching with transferable utilities has also been applied to competitive equilibrium in good markets with hedonic pricing (Chiappori, McCann, and Nesheim, 2010), to trade (e.g., Costinot and Vogel, 2015) to the labour market (Tervio (2008) and Gabaix and Landier (2008)) and to industrial organization (Bajari and Fox (2013), Fox (2018), Fox, Yang, and Hsu (2018)) among other fields. Our results apply in all of these contexts; however for concreteness, we will stick to the marriage metaphor in our exposition of the main results.

While Becker presented the general theory, he focused on the special case in which the types of the partners are one-dimensional and are complementary in producing surplus. As is well-known, the social optimum then exhibits *positive assortative matching*: higher types pair up with higher types. Moreover, the resulting configuration is stable, and it is in the core of the corresponding matching game. This sorting result is both simple and powerful; but its implications are also at variance with the data, in which matches are observed between partners with quite different characteristics. To account for a wider variety of matching patterns, one solution consists of allowing the matching surplus to incorporate latent characteristics—heterogeneity that is unobserved by the analyst. Choo and Siow (2006) have shown how it can be done in a way that yields a highly tractable model in large populations, provided that the unobserved heterogeneities enter the marital surplus quasi-additively, and that they are independent and identically distributed as standard type I extreme value terms. Choo and Siow (2006) used their model to evaluate the effect of the legalization of abortion on gains to marriage; and they applied it in Siow and Choo (2006) to Canadian data to measure the impact of demographic changes. It has also been used to study increasing returns in marriage markets (Botticini and Siow (2011)), to compare the preference for marriage versus cohabitation (Mourifié and Siow, 2021) and to estimate the changes in the returns to education on the US marriage market (Chiappori, Salanié, and Weiss, 2017). A continuous version of Choo and Siow's logit framework has been developed by Dupuy and Galichon (2014) to understand the affinities between continuous characteristics personality traits on the marriage market, using Dagsvik's theory of extreme value processes. Ciscato, Galichon, and Goussé (2020) used this approach to compare same-sex and different-sex couples.

We revisit here the theory of matching with transferable utility in the light of Choo and Siow's insights. Three assumptions underlie their contribution: the unobserved heterogeneities on the two sides of a match do not interact in producing matching surplus; they are distributed as iid type I extreme values; and populations are large. We maintain the first "separability" assumption, and the last one which is innocuous in many applications. Choo and Siow's distributional assumption, on the other hand, is very special; it generates a multinomial logit model that has quite specific restrictions on cross-elasticities. We first show that this distributional assumption can be completely dispensed with, and that the Choo-Siow framework can be extended to encompass much less restrictive assumptions on the unobserved heterogeneity. Our second contribution is to spell out a complete empirical approach to identification, parametric estimation, and computation in this class of models. Our third contribution is to revisit the original Choo and Siow (2006) dataset on marriage patterns by age, making use of the new possibilities allowed by our extended framework. We shall defer to Section 1.3 the precise description of each step of our paper.

There are other approaches to estimating matching models with unobserved heterogeneity; see the handbook chapter by Graham (2011, 2014) and the surveys by Chiappori and Salanié (2016) and Chiappori (2020). For markets with transferable utility, Fox (2010, 2018) has proposed pooling data across many similar markets and relying on a "rank-order property" that is valid when unobserved heterogeneity is separable and exchangeable—which excludes the nested logit, mixed logit, and other models considered in our paper. Bajari and Fox (2013) applied this approach to spectrum auctions. Fox, Yang, and Hsu (2018) focus on identifying the complementarity between unobservable characteristics. Gualdani and Sinha (2019) study partial identification issues in nonparametric matching models.

The literature on markets with non-transferable utility has evolved separately, with some interesting similarities—in particular with Menzel (2015)'s investigation of large NTU markets, building on a model of Dagsvik (2000). Many papers have modeled school assignment, where preferences on one side of the market are highly constrained by regulation (see Agarwal and Somaini (2020) for a recent review.) Agarwal (2015) estimates matching in the US medical resident program; his work relies on the assumption that all hospitals agree on how they rank candidates.

Notation and terminology In the following, $X \sim P$ will denote that random variable X has probability distribution P. We use **bold** type to denote vectors and matrices. Under perfectly transferable utility, the stable matching maximizes the social surplus over the set of feasible matchings (Shapley and Shubik, 1972); we sometimes use the terms "social optimum" or "equilibrium" to denote the stable matching. For simplicity, we also use "joint surplus" and "joint utility" interchangeably. We hope that this creates no confusion.

1 Framework and Roadmap

We study in this paper a bipartite, one-to-one matching market with transferable utility. We maintain throughout some of the basic assumptions of Choo and Siow (2006): utility transfers between partners are unconstrained, matching is frictionless, and there is no asymmetric information among potential partners. We call the partners "men" and "women", as we have in mind an application to the heterosexual marriage market; our results are not restricted to a marriage context, however.

1.1 The setting

Following Choo and Siow, we assume that the analyst can only observe which *group* each individual belongs to. Each man $i \in \mathcal{I}$ belongs to one group $x_i \in \mathcal{X}$; and, similarly, each woman $j \in \mathcal{J}$ belongs to one group $y_j \in \mathcal{Y}$. We will say that "man i is in group x" and "woman j is in group y." There is a finite number of groups; they are defined by the intersection of the characteristics which are observed by all men and women, and also by the analyst. On the other hand, men and women of a given group differ along some dimensions that they all observe, but which do not figure in the analyst's dataset.

Like Choo and Siow, we assume that there is an (uncountably) infinite number of men in any group x, and of women in any group y. We denote n_x the mass of men in group x, and m_y the mass of women in group y. Since the problem is homogenous, we can assume that the total mass of individuals is normalized to one, that is $\sum_x n_x + \sum_y m_y = 1$. Hence, n_x and m_y are not to be thought as numbers of individual of each types, but as masses. We will often use the notation $\boldsymbol{r} = (\boldsymbol{n}, \boldsymbol{m})$ for the vector that collects the "margins" of the problem.

A *matching* is the specification of who matches with whom. It is *feasible* if each individual is matched to 0 or 1 partner. It is *stable* if no individual who has a partner would prefer to be single, and if no two individuals would prefer forming a couple to their current situation.

Our data can only describe matchings at the group level—that is, the mass distribution of matched pairs across groups. Let μ_{xy} be the mass of the couples where the man belongs to group x, and where the woman belongs to group y. The (group-level) feasibility constraints state that the mass of married individuals in each group cannot be greater than the mass of individuals in that group, which is denoted $\boldsymbol{\mu} \in \mathcal{M}\left(\boldsymbol{r}\right)$, where $\mathcal{M}\left(\boldsymbol{r}\right)$ (or $\mathcal{M}$ in the absence of ambiguity) is defined by:

$$\mathcal{M}\left(\boldsymbol{n}, \boldsymbol{m}\right) = \left\{ \boldsymbol{\mu} \geq 0 : \forall x \in \mathcal{X},\ \sum_{y \in \mathcal{Y}} \mu_{xy} \leq n_x\ ;\ \forall y \in \mathcal{Y},\ \sum_{x \in \mathcal{X}} \mu_{xy} \leq m_y \right\} \tag{1.1}$$

With mild abuse, we will call each element of $\mathcal{M}$ a *feasible matching*. For notational convenience, we shall denote $\mu_{x0} = n_x - \sum_{y \in \mathcal{Y}} \mu_{xy}$ the corresponding mass of single men of group x and $\mu_{0y} = m_y - \sum_{x \in \mathcal{X}} \mu_{xy}$ the mass of single women of group y. We also define the sets of marital choices that are available to male and female agents, including singlehood:

$$\mathcal{X}_0 = \mathcal{X} \cup \{0\}\,, \; \mathcal{Y}_0 = \mathcal{Y} \cup \{0\}\,,$$

and we denote

$$\mathcal{A} = (\mathcal{X} \times \mathcal{Y}) \cup (\mathcal{X} \times \{0\}) \cup (\{0\} \times \mathcal{Y})$$

the set of marital arrangements.

1.2 Separability

Every match between a man i and a woman j generates a *joint surplus*, which is the excess of the sum of their utilities when married over the sum of their utilities when single. As shown in Chiappori, Salanié, and Weiss (2017), an important assumption made implicitly in Choo and Siow is that the joint surplus created when a man i of group x marries a woman j of group y does not allow for interactions between their unobserved characteristics, conditional on (x, y). This leads us to assume:

Assumption 1 (Separability)**.** *There exist a matrix* $\mathbf{\Phi}$ *and random terms* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{\eta}$ *such that*

(i) the joint utility from a match between a man i *in group* $x \in \mathcal{X}$ *and a woman* j *in group* $y \in \mathcal{Y}$ *is*

$$\tilde{\Phi}_{ij} = \Phi_{xy} + \varepsilon_{iy} + \eta_{xj}, \tag{1.2}$$

(ii) the utility of a single man i *is* $\tilde{\Phi}_{i0} = \varepsilon_{i0}$

(iii) the utility of a single woman j *is* $\tilde{\Phi}_{0j} = \eta_{0j}$

where, conditional on $x_i = x$, the random vector $\boldsymbol{\varepsilon}_i = (\varepsilon_{iy})_{y \in \mathcal{Y}_0}$ has probability distribution $\boldsymbol{P}_x$, and, conditional on $y_j = y$, the random vector $\boldsymbol{\eta}_j = (\eta_{xj})_{x \in \mathcal{X}_0}$ has probability distribution $\boldsymbol{Q}_y$. The variables

$$\max_{y \in \mathcal{Y}_0} |\varepsilon_{iy}| \quad \textit{and} \quad \max_{x \in \mathcal{X}_0} |\eta_{xj}|$$

have finite expectations under $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ respectively.

While separability is a restrictive assumption, it allows for "matching on unobservables": a match between a man of group x and a woman of group y may occur because this woman has unobserved characteristics that make her attractive to men of group x, and/or because this man has a strong unobserved preference for women of group y. What separability does rule out, however, is sorting on unobserved characteristics on both sides of the market, e.g. some unobserved preference of man i for some unobserved characteristics of woman j.

Note that we did not constrain the distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ to belong to the extreme value class. We extend the logit framework of Choo and Siow (2006) in several important ways: we allow for different families of distributions, with any form of heteroskedasticity, and with any pattern of correlation across partner groups. We will demonstrate the use of these extensions on an application in Section 6.

To summarize, a man i in this economy is characterized by his full type $(x_i, \boldsymbol{\varepsilon}_i)$, where $x_i \in \mathcal{X}$ and $\boldsymbol{\varepsilon}_i \in \mathbb{R}^{\mathcal{Y}_0}$; the distribution of $\boldsymbol{\varepsilon}_i$ conditional on $x_i = x$ is $\boldsymbol{P}_x$. Similarly, a woman j is characterized by her full type $(y_j, \boldsymbol{\eta}_j)$ where $y_j \in \mathcal{Y}$ and $\boldsymbol{\eta}_j \in \mathbb{R}^{\mathcal{X}_0}$, and the distribution of $\boldsymbol{\eta}_j$ conditional on $y_j = y$ is $\boldsymbol{Q}_y$.

1.3 Objectives and a roadmap

While the paper's final goal is to develop inference tools for matching problems with transferable utility and separable unobserved heterogeneity, this will require several intermediate steps.

First, we show how given separability, the two-sided matching problem re-

solves into a collection of one-sided problems of lower complexity.

Second, we provide new results on discrete choice (one-sided) models. One-sided discrete choice problems will play a key role in our analysis. Section 2 provides new results on this class of problems. We introduce a convex function which we call the *generalized entropy of choice*. Theorem 1 shows that this function is the value of an optimal transport problem, for which numerous computational methods have been developed. Theorem 2 then proves that given the choice probabilities and the distribution of errors, the underlying mean utilities are identified by the gradient of the generalized entropy of choice. These results should be useful beyond the setting of this paper.

Third, we show how the stable matching solves a convex optimization problem. This is done in Section 3.1, and formally stated in Theorem 3.

Fourth, we use convex duality to identify the matching surplus. Identification consists of recovering the matching surplus based on the observation of the matching patterns; this is the "inverse problem" to the computation of the stable matching given the surplus. We show in Section 3.2 that these two problems are conjugate of each other in the sense of convex duality. As a consequence, the matching surplus is identified from the matching patterns given any distribution of errors (Theorem 4).

Fifth, we propose new computational methods for the equilibrium and estimation problems. The convexity of all of our objects allows for a number of efficient computational methods to compute the stable matching and/or recover the joint surplus. Section 4 shows how this can be done by gradient descent, coordinate descent, and linear programming. In particular, coordinate descent generates a very efficient "IPFP" algorithm for variants of the logit model; we prove its convergence in Theorem 5.

Taken together, these results allow us to develop **a comprehensive set of tools for the parametric estimation of the matching model.** We allow for parameters both in the matching surplus and in the distribution of the random utility. Section 5 first investigates the properties of maximum likelihood estimation in that setting (Section 5.1). We present an alternative method based on matching observed moments of the distributions of matched pairs in Section 5.2. This is attractive as unlike maximum likelihood, it retains global convexity and has an intuitive interpretation. Finally, **we test our approaches** in Section 6, where we fit several instances of separable models to the Choo and Siow (2006) dataset.

We have tried to keep the exposition intuitive in the body of the paper; all proofs can be found in Appendix A. Appendix B specializes our results to several common specifications of unobserved heterogeneity. The paper is complemented by several online appendices where we discuss the assumptions that are relaxed or maintained in the paper (Appendix C); we provide complementary results with equilibrium predictions (Appendix D); we provide complementary estimation results (Appendix E); we give pseudo-code for our IPFP algorithm and give simulation results for this and other algorithms (Appendix F); and we provide additional details on the application of Section 6 (Appendix G). Finally, we provide Python and R code to estimate this class of models at https://bsalanie.github.io/.

2 Social surplus and identification in the one-sided case: discrete choice models

As shown by Chiappori, Salanié, and Weiss (2017), separability reduces the two-sided matching problem to a collection of one-sided discrete choice problems that are only linked through a surplus-splitting formula. Men of a given group x match with women of different groups, since each man i has idiosyncratic $\boldsymbol{\varepsilon_{i\cdot}}$ shocks. But as a consequence of the separability assumption, if a man of group x matches with a woman of group y, he would be equally well-off with any other woman of this group[1].

We now state this result more rigorously:

Proposition 1 (Splitting the surplus)**.** *Under Assumption 1, there exist* $\boldsymbol{U} = (U_{xy})$ *and* $\boldsymbol{V} = (V_{xy})$ *for* $(x, y) \in A$*, with* $U_{x0} = V_{0y} = 0$*, such that at any stable matching* (μ_{xy})*,*

(i) A man i of group x marries a woman of group $y^* \in \mathcal{Y}$ iff y^* maximizes $U_{xy} + \varepsilon_{iy}$

[1] Provided of course that she in turn ends up matched with a man of group x.

over $y \in \mathcal{Y}_0$. If the maximum is achieved at $y = 0$, this man remains single. Man i's utility $\tilde{u}_i$ is the value of the maximum.

(ii) A woman j of group y marries a man of group $x^* \in \mathcal{X}$ iff x^* maximizes $V_{xy} + \eta_{xj}$ over $x \in \mathcal{X}_0$. If the maximum is achieved at $x = 0$, this woman remains single. Woman j's utility $\tilde{v}_j$ is the value of the maximum.

(iii) $U_{xy} + V_{xy} \geq \Phi_{xy}$ for all $(x, y) \in \mathcal{A}$, with equality if $\mu_{xy} > 0$.

Before we solve the two-sided matching problem, we need to derive results on one-sided discrete choice problems. Since these results are of independent interest, we step back from the matching problem and consider the classic problem of an individual who chooses from a set of alternatives $y \in \mathcal{Y}_0 = \mathcal{Y} \cup \{0\}$ whose utilities are $U_y + \varepsilon_y$. We assume that the vector $\boldsymbol{\varepsilon} = (\varepsilon_y)_{y \in \mathcal{Y}_0}$ has a distribution $\mathbb{P}$; without loss of generality, we impose $U_0 = 0$ and we denote $\boldsymbol{U} = (U_1, \ldots, U_{|Y|})$.

2.1 Social surplus in discrete choice models

We first show that the ex-ante indirect surplus can be expressed as a sum of two terms: the weighted sum of the mean utilities, and a *generalized entropy of choice* which stems from the unobservable heterogeneity. We will provide two useful characterizations of the generalized entropy function, one as the convex conjugate of the ex-ante indirect utility, and the other one as the solution to an optimal transport problem (see Galichon, 2016, for an introduction). To the best of our knowledge, these results are new.

The average utility of the agent is

$$G(\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{P}} \max_{y \in \mathcal{Y}_0} (U_y + \varepsilon_y) \tag{2.1}$$

where the expectation is taken over the random vector $\boldsymbol{\varepsilon} = (\varepsilon_0, \ldots, \varepsilon_{|\mathcal{Y}|}) \sim \boldsymbol{P}$. The function G is known as the *Emax operator* in the discrete choice literature[2].

[2] The Emax operator is available in closed-form in classical instances like McFadden's generalized extreme value class (McFadden, 1978). In other cases, one needs to use numerical integration; see Train (2009) and references therein.

Note that as the expectation of the maximum of linear functions of the (U_y), G is a convex function of $\boldsymbol{U}$. Now consider a large population of individuals who face the same mean utilities and draw independent $\boldsymbol{\varepsilon}_i$ from $\mathbb{P}$. Let $Y_i^* \in \mathcal{Y}_0$ denote the optimal choice of individual i; then

$$G(\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{P}}\left(U_{Y_i^*} + \varepsilon_{i,Y_i^*}\right) = \sum_{y\in\mathcal{Y}} \mu_y U_y + \mathbb{E}_{\boldsymbol{P}}\left(\varepsilon_{i,Y_i^*}\right), \tag{2.2}$$

where μ_y is the proportion of individuals who choose alternative y.

2.2 Generalized entropy of choice

Our analysis gives a prominent role to a classical concept in convex analysis: the *Legendre-Fenchel transform* of G. Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{|\mathcal{Y}|})$; we define

$$G^*(\boldsymbol{\mu}) = \sup_{\tilde{\boldsymbol{U}}=(\tilde{U}_1,\dots,\tilde{U}_{|\mathcal{Y}|})} \left(\sum_{y\in\mathcal{Y}} \mu_y \tilde{U}_y - G(\tilde{\boldsymbol{U}})\right) \tag{2.3}$$

whenever $\sum_{y\in\mathcal{Y}} \mu_y \leq 1$, and $G^*(\boldsymbol{\mu}) = +\infty$ otherwise. Note that the domain of G^* is the set of $\boldsymbol{\mu}$ that can be interpreted as vectors of choice probabilities of alternatives in $\mathcal{Y}$. As the supremum of a set of linear functions of $\boldsymbol{\mu}$, G^* is a convex function.

We will see in Example 2.1 that in the logit setting, $-G^*$ is the usual entropy function. This motivates the following definition:

Definition 1. *We call the function $-G^*$ the* generalized entropy of choice*.*

The theory of convex duality implies that since G is convex, it is reciprocally the Legendre-Fenchel transform of G^*:

$$G(\boldsymbol{U}) = \sup_{\tilde{\boldsymbol{\mu}}=(\tilde{\mu}_1,\dots,\tilde{\mu}_{|\mathcal{Y}|})} \left(\sum_{y\in\mathcal{Y}} \tilde{\mu}_y U_y - G^*(\tilde{\boldsymbol{\mu}})\right). \tag{2.4}$$

Assume that $\boldsymbol{\mu}$ attains the supremum in (2.4). Then

$$G(\boldsymbol{U}) + G^*(\boldsymbol{\mu}) = \sum_{y\in\mathcal{Y}} \mu_y U_y;$$

and comparing with (2.2) shows that

$$G^*\left(\boldsymbol{\mu}\right) = -\mathbb{E}_{\boldsymbol{P}}\left(\varepsilon_{iY_i^*}\right). \tag{2.5}$$

Therefore $-G^*\left(\boldsymbol{\mu}\right)$ is just the average heterogeneity that is required to rationalize the conditional choice probability vector $\boldsymbol{\mu}$. The following result goes beyond formula (2.5) and allows us to provide a useful characterization of the generalized entropy of choice. It shows that it can be computed by solving an optimal transport problem.

Theorem 1 (Characterization of the generalized entropy of choice)**.** *Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_{|\mathcal{Y}|})$ with $\sum_{y\in\mathcal{Y}} \mu_y \leq 1$, and denote $\mu_0 = 1 - \sum_{y\in\mathcal{Y}} \mu_y$. Let $\mathcal{M}\left(\boldsymbol{\mu}, \boldsymbol{P}\right)$ denote the set of probability distributions π of the random joint vector $(\boldsymbol{Y}, \boldsymbol{\varepsilon})$, where $\boldsymbol{Y} \sim (\mu_0, \boldsymbol{\mu})$ is a random element of $\mathcal{Y}_0$, and $\boldsymbol{\varepsilon} \sim \boldsymbol{P}$ is a random vector of $\mathbb{R}^{|\mathcal{Y}_0|}$.*

Then $-G^*(\boldsymbol{\mu})$ is the value of the optimal transport problem between the distribution $(\mu_0, \boldsymbol{\mu})$ of $\boldsymbol{Y}$ and the distribution $\boldsymbol{P}$ of $\boldsymbol{\varepsilon}$, when the surplus is given by $\varepsilon_{\boldsymbol{Y}}$. That is,

$$-G^*(\boldsymbol{\mu}) = \sup_{\pi\in\mathcal{M}(\boldsymbol{\mu},\boldsymbol{P})} \mathbb{E}_\pi\left(\varepsilon_{\boldsymbol{Y}}\right). \tag{2.6}$$

Since a discretized version of problem (2.6) can be solved by efficient linear programming algorithms, it provides us with a practical solution to the computation of generalized entropy for quite general distributions of unobserved heterogeneity. Several applications of this result to useful classes of distributions are given below in Section 2.4.

2.3 Identification of discrete choice models

The generalized entropy of choice function $-G^*$ is our gateway to identifying the mean utilities. Let us first give the intuition of our result. Assume that the distribution $\mathbb{P}$ is known and that it generates functions G and G^* that are continuously differentiable – this is the case, in particular, when the distribution has a density with full support. By the Daly-Zachary-Williams theorem[3], we know that the derivative of the average maximized utility of an agent with respect to U_y is equal to the probability that this agent chooses the corresponding alternative y, that is

$$\frac{\partial G}{\partial U_y}(\boldsymbol{U}) = \mu_y. \tag{2.7}$$

This is simply an application of the envelope theorem to (2.1). We can also use it on (2.3); this gives

$$\frac{\partial G^*}{\partial \mu_y}(\boldsymbol{\mu}) = U_y \tag{2.8}$$

where U_y achieves the maximum in (2.3). By the Fenchel duality theorem[4], these two sets of conditions are equivalent. As a consequence, for any fixed distribution of $\boldsymbol{\varepsilon}$ (which determines the shape of G and G^*), the mean utilities $\boldsymbol{U}$ are identified from $\boldsymbol{\mu}$, the observed matching patterns of the agents.

Theorem 2 (Identifying the mean utilities)**.** *Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{|\mathcal{Y}|})$ with $\sum_{y \in \mathcal{Y}} \mu_y \leq 1$; $U_0 = 0$; and $\boldsymbol{U} = (U_1, \ldots, U_{|\mathcal{Y}|})$. If $\boldsymbol{P}$ has full support and is absolutely continuous with respect to the Lebesgue measure, the following statements are equivalent:*

1. *for every $y \in \mathcal{Y}$, $\mu_y = \dfrac{\partial G}{\partial U_y}(\boldsymbol{U})$*
2. *for every $y \in \mathcal{Y}$, $U_y = \dfrac{\partial G^*}{\partial \mu_y}(\boldsymbol{\mu})$*
3. *there exists a scalar function $u(\boldsymbol{\varepsilon})$, integrable with respect to $\boldsymbol{P}$, such that $(u, \boldsymbol{U})$ are*

[3] Williams (1977) and Daly and Zachary (1978).

[4] See e.g. Hiriart-Urruty and Lemaréchal (2001, p. 211).

the unique minimizers of the dual problem to (2.6), that is of:

$$\begin{aligned} -G^*(\boldsymbol{\mu}) = \min_{\bar{U},\bar{u}} \quad & \int \bar{u}(\boldsymbol{\varepsilon})\, d\boldsymbol{P}(\boldsymbol{\varepsilon}) - \sum_{y\in\mathcal{Y}} \mu_y \bar{U}_y \\ s.t. \quad & \bar{u}(\boldsymbol{\varepsilon}) - \bar{U}_y \geq \varepsilon_y \quad \forall y \in \mathcal{Y}, \forall \boldsymbol{\varepsilon} \in \mathbb{R}^{\mathcal{Y}_0} \\ & \bar{U}_0 = 0. \end{aligned} \tag{2.9}$$

Since the functions G and G^* are convex, they are differentiable almost everywhere. Our assumption on $\boldsymbol{P}$ makes them continuously differentiable. This is not essential to our approach[5], but it makes for simpler formulæ and numerical computations.

Part 1 of Theorem 2 is well-known in the discrete choice literature, and we only restate it for completeness. Parts 2 and 3 do not seem to have appeared before our paper. The only related prior results we could find are the inversion formulæ of Hotz and Miller (1993) and Arcidiacono and Miller (2011) for dynamic discrete choice models; but their scope is much more restricted since they only apply to multinomial logit and to GEV models, respectively. In contrast, parts 2 and 3 provides a constructive method to identify U_y based on the conditional choice probabilities $\boldsymbol{\mu}$, as the solution to a convex optimization problem (part 2) which is in fact an optimal transport problem (part 3). The intuition behind part 3 is simply that each observed choice probability μ_y must be matched to the values of idiosyncratic preference shocks $\boldsymbol{\varepsilon}_i \sim \boldsymbol{P}$ for which y is the most preferred choice. The $\boldsymbol{U}$ are the shadow prices that support this matching. Chiong, Galichon, and Shum (2016) apply our approach to dynamic discrete-choice models.

2.4 Examples

Example 2.1 (Logit and nested logit)**.** *The nested logit model is a well-known generalization of the ubiquitous (multinomial) logit model. Consider a two-layer nested logit model where alternative* 0 *is alone in a nest and each other nest* $n \in \mathcal{N}$ *contains alternatives* $y \in \mathcal{Y}(n)$*. The correlation of alternatives whithin nest* n *is proxied by* $1 - \lambda_n^2$ *(with* $\lambda_0 = 1$

[5] It holds in all popular specifications, including the multinomial logit model of course.

for the nest made of alternative 0*). Calculations detailed in Appendix B.2 show that*

$$G(\boldsymbol{U}) = \log\left[1 + \sum_{n\in\mathcal{N}}\left(\sum_{y\in\mathcal{Y}(n)}\exp\left(\frac{U_y}{\lambda_n}\right)\right)^{\lambda_n}\right], \tag{2.10}$$

$$G^*(\boldsymbol{\mu}) = \mu_0\log\mu_0 + \sum_{n\in\mathcal{N}}\left(\lambda_n\sum_{y\in\mathcal{Y}(n)}\mu_y\log\mu_y + (1-\lambda_n)\,\mu_n\log\mu_n\right) \tag{2.11}$$

where $\mu_0 := 1 - \sum_{y\in|\mathcal{Y}|}\mu_y$ *and* $\mu_n := \sum_{y\in\mathcal{Y}(n)}\mu_y$.

Moreover, $U_y = \lambda_n\log\left(\mu_y/\mu_0\right) + (1-\lambda_n)\log\left(\mu_n/\mu_0\right)$.

In particular, when $\lambda_n = 1$ *for every nest* n*, we recover the multinomial logit model:*

$$G(\boldsymbol{U}) = \log\left(1 + \sum_{y\in\mathcal{Y}}\exp(U_y)\right) \tag{2.12}$$

$$G^*(\boldsymbol{\mu}) = \mu_0\log\mu_0 + \sum_{y\in\mathcal{Y}}\mu_y\log\mu_y. \tag{2.13}$$

along with $\mu_y = \exp\left(U_y\right)/\left(1+\sum_{y'\in\mathcal{Y}}\exp(U_{y'})\right)$ *and* $U_y = \log\left(\mu_y/\mu_0\right)$.

Example 2.2 (Random coefficients multinomial logit and pure characteristics model)**.** *Now consider the random coefficient logit model which underlies much of empirical industrial organization (Berry, Levinsohn, and Pakes, 1995). In this case,* $\boldsymbol{\varepsilon} = \boldsymbol{Ze} + T\boldsymbol{\eta}$*, where* $\boldsymbol{e}$ *is a random vector on* $\mathbb{R}^d$ *with distribution* $\mathbf{P}_e$*;* $\boldsymbol{Z}$ *is a* $|\mathcal{Y}_0| \times d$ *matrix;* $T > 0$ *is a scalar parameter, and* $\boldsymbol{\eta}$ *is a vector of* $|\mathcal{Y}|$ *extreme value type-I (Gumbel) random variables that is independent from* e*. Appendix B.3 shows that* $-G^*(\boldsymbol{\mu})$ *is a solution to the regularized optimal transport problem*

$$-G^*(\boldsymbol{\mu}) = \min_{U_0=0,\boldsymbol{U}\in\mathbb{R}^{\mathcal{Y}}}\left[\int T\log\sum_{y\in\mathcal{Y}_0}\exp\left(\frac{U_y + (\boldsymbol{Ze})_y}{T}\right)d\boldsymbol{P}_e(\boldsymbol{e}) - \sum_{y\in\mathcal{Y}}\mu_y U_y\right] \tag{2.14}$$

and the vector $\boldsymbol{U}$ *that attains the minimum in (2.14) is the solution to the identification problem.*

The case $T = 0$ yields the pure characteristics model of Berry and Pakes (2007) described at greater length in Appendix B.4. Then $\boldsymbol{\varepsilon} = \boldsymbol{Ze}$; and

$$-G^*(\boldsymbol{\mu}) = \min_{U_0=0,\boldsymbol{U}\in\mathbb{R}^{\mathcal{Y}}} \int \max_{y\in\mathcal{Y}_0} \left\{ (\boldsymbol{Ze})_y + U_y \right\} d\boldsymbol{P}_\epsilon(\epsilon) - \sum_{y\in\mathcal{Y}} \mu_y U_y \tag{2.15}$$

is the solution to the power diagram problem (see Galichon, 2016, Chapter 5).

3 Social surplus and identification in the two-sided case: matching models

We now return to matching models. Proposition 1 shows that a man i of group x can be modeled as choosing a partner by maximizing $(U_{xy} + \varepsilon_{iy})$ over $y \in \mathcal{Y}_0$ (continuing with our convention that $U_{x0} = 0$). Building on our results on one-sided discrete choice, we define G_x to be the corresponding Emax function:

$$G_x(\boldsymbol{U}_{x\cdot}) = E_{\boldsymbol{P}_x} \max_{y\in\mathcal{Y}_0} (U_{xy} + \varepsilon_{iy})$$

and the Legendre-Fenchel transform

$$G_x^*(\boldsymbol{\nu}) = \max_{\boldsymbol{U}\in\mathbb{R}^{\mathcal{Y}}} \left(\sum_{y\in\mathcal{Y}} \nu_y U_y - G_x(\boldsymbol{U}) \right)$$

for $\sum_{y\in\mathcal{Y}} \nu_y \leq 1$ (and $G_x^*(\boldsymbol{\nu}) = +\infty$ otherwise). Given group numbers $\boldsymbol{n} = (n_x)$, the *aggregate welfare of men* is

$$G(\boldsymbol{U}, \boldsymbol{n}) = \sum_{x\in\mathcal{X}} n_x G_x(\boldsymbol{U}_{x\cdot}); \tag{3.1}$$

for $\boldsymbol{\mu} = (\mu_{xy})_{x\in\mathcal{X},y\in\mathcal{Y}}$, we denote its Legendre-Fenchel transform by

$$G^*(\boldsymbol{\mu}, \boldsymbol{n}) = \sup_{\boldsymbol{U}\in\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}} \left(\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mu_{xy}U_{xy} - G(\boldsymbol{U}, \boldsymbol{n}) \right)$$

which is (minus) the generalized entropy of choice of all men. Standard calculations show that

$$G^*(\boldsymbol{\mu}, \boldsymbol{n}) = \sum_{x\in\mathcal{X}} n_x G_x^*(\boldsymbol{\mu_{x\cdot}}/\boldsymbol{n_x}).$$

We define $H_y(\boldsymbol{V}_{\cdot y})$ as the Emax function on women's side. Given group numbers $\boldsymbol{m} = (m_y)$, the aggregate welfare of women is $H(\boldsymbol{V}, \boldsymbol{m})$; the generalized entropy of choice of women of group y and of all women are the respective Legendre-Fenchel transforms of H_y and H.

3.1 Social surplus, equilibrium and entropy of matching

It has been known since Shapley and Shubik (1972) that under perfectly transferable utility, the stable matching maximizes the social surplus over the set of feasible matchings. Theorem 3 provides a simple analytical expression for the value of the optimal social surplus. We start with an intuitive derivation of this result.

The social surplus $\mathcal{W}$ is simply the sum of the aggregate welfare of men and the aggregate welfare of women:

$$\mathcal{W} = G(\boldsymbol{U}, \boldsymbol{n}) + H(\boldsymbol{V}, \boldsymbol{m}) = \sum_{x\in\mathcal{X}} n_x G_x(\boldsymbol{U_{x\cdot}}) + \sum_{y\in\mathcal{Y}} m_y H_y(\boldsymbol{V_{\cdot y}}). \tag{3.2}$$

Let $\boldsymbol{\mu} = (\mu_{xy})_{x\in\mathcal{X},y\in\mathcal{Y}}$ be the stable matching that corresponds to $(\boldsymbol{U}, \boldsymbol{V} = \boldsymbol{\Phi} - \boldsymbol{U})$. Summing the expressions (2.4) over x gives

$$G(\boldsymbol{U}, \boldsymbol{n}) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mu_{xy}U_{xy} - G^*(\boldsymbol{\mu}, \boldsymbol{n});$$

and similarly,

$$H\left(\boldsymbol{V},\boldsymbol{m}\right)=\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}V_{xy}-H^{*}\left(\boldsymbol{\mu},\boldsymbol{m}\right).$$

As a result, the value of the social welfare can be expressed as

$$\mathcal{W}=\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}\Phi_{xy}+\mathcal{E}(\boldsymbol{\mu},\boldsymbol{n},\boldsymbol{m}) \tag{3.3}$$

where we have defined the *generalized entropy of matching* by

$$\mathcal{E}(\boldsymbol{\mu},\boldsymbol{n},\boldsymbol{m}):=-G^{*}(\boldsymbol{\mu},\boldsymbol{n})-H^{*}(\boldsymbol{\mu},\boldsymbol{m}). \tag{3.4}$$

To simplify the exposition, we will make sure that the G,H,G^{*} and H^{*} are continously differentiable everywhere.

Assumption 2 (Full support)**.** *For all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, the distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ have full support and are absolutely continuous with respect to the Lebesgue measure.*

Theorem 3 shows that the values of the optimum social welfare and the stable matching patterns emerge from the solution to simple convex optimization problems:

Theorem 3 (Social surplus at equilibrium)**.** *Under Assumptions 1 and 2, for any $\boldsymbol{\Phi}$ and $\boldsymbol{r}=(\boldsymbol{n},\boldsymbol{m})$ the stable matching $\boldsymbol{\mu}$ maximizes the social gain over all feasible matchings $\boldsymbol{\mu}\in\mathcal{M}(\boldsymbol{r})$, that is*

$$\mathcal{W}\left(\boldsymbol{\Phi},\boldsymbol{r}\right)=\max_{\boldsymbol{\mu}\in\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}}\left(\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}\Phi_{xy}+\mathcal{E}(\boldsymbol{\mu},\boldsymbol{r})\right). \tag{3.5}$$

Equivalently, $\mathcal{W}$ is given by its dual expression

$$\begin{aligned}\mathcal{W}\left(\boldsymbol{\Phi},\boldsymbol{r}\right) &= \min_{\boldsymbol{U},\boldsymbol{V}\in\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}}\left(G(\boldsymbol{U},\boldsymbol{n})+H(\boldsymbol{V},\boldsymbol{m})\right) \qquad (3.6)\\ s.t. &\quad U_{xy}+V_{xy}\geq\Phi_{xy}\ \forall x\in\mathcal{X},y\in\mathcal{Y}.\end{aligned}$$

The optimal solution $\boldsymbol{\mu}$ *to* (*3.5*) *and the optimal solution* $(\boldsymbol{U}, \boldsymbol{V})$ *to* (*3.6*) *are related by*

$$\mu_{xy} = \frac{\partial G}{\partial U_{xy}}(\boldsymbol{U}, \boldsymbol{n}) = \frac{\partial H}{\partial V_{xy}}(\boldsymbol{V}, \boldsymbol{m}). \tag{3.7}$$

The proof of this result is given in Appendix A. It calls for a few remarks.

Remark 1. The right-hand side of equation (3.5) gives the value of the social surplus when the matching patterns are $\boldsymbol{\mu}$. Its first term $\sum_{xy} \mu_{xy} \Phi_{xy}$ reflects "systematic preferences" on group characteristics, while the second term $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{r})$ reflects the effect of idiosyncratic preferences. The market equilibrium trades off matching on group characteristics and matching on unobserved characteristics. If the first term dominates, then one recovers the linear programming problem of Shapley and Shubik (1972). If on the contrary, available data were so poor that unobserved heterogeneity dominates ($\boldsymbol{\Phi} \simeq 0$), then the analyst should observe something that looks like random matching. Information theory tells us that entropy is a natural measure of statistical disorder; and as we will see in Example 3.1, in the simple case analyzed by Choo and Siow the "generalized entropy of matching" $\mathcal{E}$ is just the usual notion of entropy, which is why we chose this term.

Remark 2. The dual problem (3.6) explains the *destination* of the surplus shared at equilibrium between men and women: $n_x G_x(\boldsymbol{U_{x\cdot}})$ is the total amount of utility going to men of type x, while $m_y H_y(\boldsymbol{V_{\cdot y}})$ is the total amount of utility going to women of type y. In contrast, the primal problem (3.5) accounts for the *origin* of the surplus: Φ_{xy} originates from the part of the surplus that comes from the interaction between observable characteristics in pair xy, while $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{m})$ originates from unobservable heterogeneities in tastes.

Remark 3. Equations (3.7) are the first-order conditions of (3.6). They can be rewritten as the equality between the demand of men of group x for women of group y, and the right-hand side is the demand of women of group y for men of group x. In equilibrium these numbers must both equal the number of matches between these two groups, μ_{xy}.

Remark 4. A wealth of comparative statics results and testable predictions can be deduced from Theorem 3; we explore some of them in Appendices D.1 and D.2.

Characterizing individual and systematic utilities. We can now offer a characterization of equilibrium utilities, both at the individual level and aggregated over observable groups.

Proposition 2 (Individual and group surpluses)**.** *Let* $(\boldsymbol{U}, \boldsymbol{V})$ *solve* (3.6)*, and* $U_{x0} = V_{0y} = 0$*. Under Assumptions 1 and 2,*

(i) A man i *of group* x *who marries a woman of group* y^* *obtains utility*

$$U_{xy^*} + \varepsilon_{iy^*} = \max_{y \in \mathcal{Y}_0} \left(U_{xy} + \varepsilon_{iy} \right).$$

(ii) The average utility of men of group x *is*

$$u_x = G_x(\boldsymbol{U_{x\cdot}}) = \frac{\partial \mathcal{W}}{\partial n_x}(\boldsymbol{\Phi}, \boldsymbol{r}).$$

(iii) Parts (i) and (ii) transpose to the other side of the market with the obvious changes; and $U_{xy} + V_{xy} = \Phi_{xy}$ *for all* x, y*.*

3.2 Identification

Ideally, we would want to identify nonparametrically both the matrix $\boldsymbol{\Phi}$ and the distributions of the error terms. This is clearly out of reach since we only observe the matching patterns $\boldsymbol{\mu}$. We focus in this section on the case when the distributions of the error terms are (assumed to be) known. Section 5 will turn to parameric inference.

Since Proposition 2 allowed us to decompose the matching problem into two collections of discrete choice problems, we can now use Theorem 2 in order to identify the matching surplus matrix $\boldsymbol{\Phi}$ as s function of the corresponding stable matching $\boldsymbol{\mu}$:

Theorem 4. *Under Assumptions 1 and 2:*

1. *$\boldsymbol{U}$ and $\boldsymbol{V}$ are identified from $\boldsymbol{\mu}$ by*

$$\boldsymbol{U} = \frac{\partial G^*}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}) \text{ and } \boldsymbol{V} = \frac{\partial H^*}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}) \tag{3.8}$$

2. *The constraint in (3.6) is always saturated: $U_{xy} + V_{xy} = \Phi_{xy}$ for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. As a result, the matching surplus $\boldsymbol{\Phi}$ is identified by*

$$\Phi_{xy} = -\frac{\partial \mathcal{E}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{r}), \tag{3.9}$$

which gives for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$\Phi_{xy} = \frac{\partial G_x^*}{\partial \boldsymbol{\mu}_{y|x}}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}\right) + \frac{\partial H_y^*}{\partial \mu_{x|y}}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{y}}\right), \tag{3.10}$$

where $\mu_{xy} = \mu_{y|x} n_x = \mu_{x|y} m_y$.

Combining Theorem 2 and 4 shows that all of the quantities in Theorem 3 can be computed by solving simple convex optimization problems.

Example 3.1 (The Choo and Siow Specification)**.** *Assume that $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ are the distributions of centered i.i.d. standard type I extreme value random variables. Then the generalized entropy is*

$$\mathcal{E} = -\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}_0}} \mu_{xy} \log \mu_{y|x} - \sum_{\substack{y \in \mathcal{Y} \\ x \in \mathcal{X}_0}} \mu_{xy} \log \mu_{x|y}, \tag{3.11}$$

which is a standard entropy[6]*.*

Average utilities are linked to matching patterns by $u_x = -\log \mu_{0|x}$ and $v_y = -\log \mu_{0|y}$, and surpluses are related to matching patterns by

$$\Phi_{xy} = 2 \log \mu_{xy} - \log \mu_{x0} - \log \mu_{0y}. \tag{3.12}$$

[6] The connection between the logit model and the classical entropy function is well known; see e.g. Anderson, de Palma, and Thisse (1988).

This is Choo and Siow (2006)'s identification result, which may be more familiar as

$$\mu_{xy} = \sqrt{\mu_{x0}\mu_{0y}}\exp(\Phi_{xy}/2). \tag{3.13}$$

Define

$$F(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\Phi}, \boldsymbol{r}) := \sum_{x\in\mathcal{X}} n_x \left(u_x + e^{-u_x} - 1\right) + \sum_{y\in\mathcal{Y}} m_y \left(v_y + e^{-v_y} - 1\right) + 2\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \sqrt{n_x m_y} e^{\frac{\Phi_{xy}-u_x-v_y}{2}} \tag{3.14}$$

As a sum of exponentials and of linear functions, it is a globally strictly convex function of $(\boldsymbol{u}, \boldsymbol{v})$*. As proved in Appendix A, the social welfare* $\mathcal{W}(\boldsymbol{\Phi}; \boldsymbol{r})$ *equals its minimum value and at the minimum,*

$$\begin{aligned}
\mu_{x0} &= n_x \exp(-u_x)\\
\mu_{0y} &= m_y \exp(-v_y)\\
\mu_{xy} &= \sqrt{n_x m_y}\exp\left((\Phi_{xy} - u_x - v_y)/2\right).
\end{aligned}$$

4 Computation

We present two methods to compute the equilibrium: min-Emax (based on gradient descent), and IPFP (based on coordinate descent). In Appendix F, we benchmark them and present a third one: linear programming based on simulated draws.

4.1 Min-Emax method

Theorem 3 gave two expressions for the social surplus. Program (3.5) solves for the equilibrium matching patterns $\boldsymbol{\mu}$. Alternatively, program (3.6) solves for the $\boldsymbol{U}$ and $\boldsymbol{V}$ utility components. Since the generalized entropy $\mathcal{E}$ is concave and the functions G and U are convex, these two programs are globally convex, with linear inequality constraints. Under Assumption 2, none of the constraints $\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{n}, \boldsymbol{m})$ in the first program bind at the optimum since all μ_{x0} and μ_{0y} are positive; and by part (ii) of Proposition 4, the constraints $\boldsymbol{U} + \boldsymbol{V} \geq \boldsymbol{\Phi}$ in the second program are all saturated at the optimum. Therefore by Theorem 3, we can obtain the equilibrium matching patterns by solving the globally concave unconstrained maximization problem (3.5), and we can obtain the $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices by solving its dual, the globally convex unconstrained minimization problem

$$\min_{\boldsymbol{U} \in \mathbb{R}^{\mathcal{X}\mathcal{Y}}} \left(G(\boldsymbol{U}, \boldsymbol{n}) + H(\boldsymbol{\Phi} - \boldsymbol{U}, \boldsymbol{m}) \right). \tag{4.1}$$

Since $G = \sum n_x G_x$, where G_x is the average value of the maximum utility of men of group x, we call the method based on (4.1) the *min-Emax* method. Problem (4.1) has dimension $|\mathcal{X}| \times |\mathcal{Y}|$, is unconstrained, and has a very sparse structure: it is easy to see that the Hessian of the objective function contains a large number of zeroes. It only requires evaluating the G_x and H_y, which is often available in closed-form; when not, we will show later (in Appendix F.2) how to use simulation and linear programming to approximate the problem. As (4.1) is globally convex, a descent algorithm converges nicely under weak conditions[7]; each of its iterations consists of updating $\boldsymbol{U}$ so as to reduce the excess demand of x for y for instance by decreasing U_{xy}, or equivalently increasing the price $V_{xy} = \Phi_{xy} - U_{xy}$ of women of group y for men of group x. Solving (4.1) therefore replicates a Walrasian tâtonnement process; we need not be concerned about its convergence since global convexity guarantees it[8].

In some cases, such as the Choo and Siow specification, the sparse structure of the problem can be exploited very easily to reduce the dimensionality further. The function F of (3.14) only has $|X| + |Y|$ arguments, rather than the $|X| \times |Y|$ of G and H. This speeds up the search for a minimum considerably—see Appendix F.

[7] As would other algorithms—see Boyd and Vandenberghe (2004).

[8] Anorher way to see it is that the demand for partners satisfies the global substitutes property.

4.2 IPFP

In some applications, the number of groups $|\mathcal{X}|$ and $|\mathcal{Y}|$ is large and solving for equilibrium by minimizing (4.1) may not be a practical option. We develop here an algorithm that extends the Iterative Projection Fitting Procedure (IPFP); it can provide a very efficient solution if the generalized entropy $\mathcal{E}$ is easy to evaluate.

The idea that underlies the algorithm is that the average utilities (u_x) and (v_y) of the groups of men and women play the role of prices that equate demand and suppply. Accordingly, we adjust the prices alternatively on each side of the market. First we fix the prices (v_y) and we find the prices (u_x) such that the demands of women for partners clear the markets for men of each group, in the sense that $\sum_{y\in\mathcal{Y}}\mu_{xy}+\mu_{x0}=n_x$ for each $x\in\mathcal{X}$. Then we fix these new prices (u_x) and we find the prices (v_y) such that the demands of men for partners clear the markets for women of each group $y\in\mathcal{Y}$; and we iterate. This is a *coordinate descent* procedure. As its name indicates, the Iterative Projection Fitting Procedure was designed to find projections on intersecting sets of constraints, by projecting iteratively on each constraint[9]. We describe the algorithm in full detail in Appendix F, and we prove its convergence there.

Theorem 5. *Under Assumptions 1 and 2, the IPFP algorithm converges to the solution* $\boldsymbol{\mu}$ *of* (3.5) *and to the corresponding average utilities* $\boldsymbol{u}$ *and* $\boldsymbol{v}$*.*

In the case of the multinomial logit Choo-Siow model of Example 3.1 for instance, we show in Appendix F.1.4 that the algorithm boils down to

$$\begin{cases} \mu_{x0}^{(2k+1)} = \left(\sqrt{n_x+\frac{A_x^2}{4}}-\frac{A_x}{2}\right)^2 \text{ with } A_x=\sum_{y\in\mathcal{Y}}\exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{0y}^{(2k)}} \\ \mu_{0y}^{(2k+2)} = \left(\sqrt{m_y+\frac{B_y^2}{4}}-\frac{B_y}{2}\right)^2 \text{ with } B_y=\sum_{x\in\mathcal{X}}\exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}^{(2k+1)}} \end{cases} \tag{4.2}$$

We tested the performance of our proposed algorithms on an instance of the Choo and Siow model; we report the results in Appendix F. The IPFP algorithm is extremely fast

[9] It is used for instance to impute missing values in data (and known for this purpose as the RAS method.)

compared to standard optimization or equation-solving methods. The min-Emax method of (4.1) is slower but it still works very well for medium-size problems, and it is applicable to all separable models.

5 Parametric Inference

We assume in this section that all observations concern a single matching market; we briefly discuss approaches that use several markets in Appendix E.3. While the formula in Theorem 3 (i) gives a straightforward estimator of the systematic surplus function $\boldsymbol{\Phi}$, with multiple payoff-relevant observed characteristics x and y it is likely to result in large standard errors when matching patterns are estimated from data on a finite number of matches. In addition, we do not know the distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$. Both of these remarks point to the need for a parametric model in most applications. Such a model would be described by a family of joint surplus functions $\Phi^{\boldsymbol{\lambda}}_{xy}$ and distributions $\boldsymbol{P}^{\boldsymbol{\lambda}}_x$ and $\boldsymbol{Q}^{\boldsymbol{\lambda}}_y$ for $\boldsymbol{\lambda}$ in some finite-dimensional parameter space Λ.

In matching markets, the sample may be drawn from the population at the individual level or at the household level. In the former case, each man or woman in the population is a sampling unit; in the latter, all individuals in a household are sampled. Household-based sampling is the norm in population surveys and we will assume it here: our sample consists of a predetermined number H of households, some of which consist of a single man or woman and some of which consist of a married couple. Such a sample will have $\hat{S} = \sum_x \hat{N}_x + \sum_y \hat{M}_y$ individuals, where $\hat{N}_x$ (resp. $\hat{M}_y$) denotes the number of men of group x (resp. women of group y) in the sample. Since sampling is at the household level, for any given value of H the numbers $\hat{\boldsymbol{N}}$ and $\hat{\boldsymbol{M}}$ of men and women of each group the sample are random: if for instance we happen to draw many households with single men, then the number of men in the sample will be large.

We will denote $\hat{n}_x = \widehat{N}_x/\hat{S}$ and $\hat{m}_y = \widehat{M}_y/\hat{S}$ the respective empirical frequencies of types of men and women. We group them in $\hat{\boldsymbol{r}} = (\hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$; and we let $\hat{\mu}_{xy}$ denote the observed number of matches between men of group x and women of group y, which satisfy the usual margin equations

$$\begin{cases} \sum_{y\in\mathcal{Y}} \mu^{\boldsymbol{\lambda}}_{xy} + \mu^{\boldsymbol{\lambda}}_{x0} = \hat{n}_x \\ \sum_{x\in\mathcal{X}} \mu^{\boldsymbol{\lambda}}_{xy} + \mu^{\boldsymbol{\lambda}}_{0y} = \hat{m}_y \end{cases} \tag{5.1}$$

We assume that this dataset is drawn from a population where matching was generated by the parametric model above, with true parameter vector $\boldsymbol{\lambda}_0$. Recall the expression of the social surplus:

$$\mathcal{W}(\boldsymbol{\Phi}^{\boldsymbol{\lambda}}, \hat{\boldsymbol{r}}) = \max_{\boldsymbol{\mu}\in\mathcal{M}(\hat{\boldsymbol{r}})} \left(\sum_{x,y} \mu_{xy} \Phi^{\boldsymbol{\lambda}}_{xy} + \mathcal{E}^{\boldsymbol{\lambda}}(\boldsymbol{\mu}, \hat{\boldsymbol{r}}) \right).$$

Let $\boldsymbol{\mu}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{r}})$ be the stable matching for parameters $\boldsymbol{\lambda}$ and margins $\hat{\boldsymbol{r}}$. We have shown in Section 4 how it can be computed efficiently. We now focus on statistical inference on $\boldsymbol{\lambda}$. We propose three methods: maximum likelihood, a moment matching method, and a minimum distance estimator.

5.1 Maximum Likelihood estimation

Estimation requires that we first compute the optimal matching with parameters $\boldsymbol{\lambda}$ for given populations of men and women. To do this, we take the numbers $\hat{n}_x$ and $\hat{m}_y$ as fixed; that is, we impose the constraints (5.1). The simulated number of households

$$H^{\boldsymbol{\lambda}} \equiv \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mu^{\boldsymbol{\lambda}}_{xy} + \sum_{x\in\mathcal{X}} \mu^{\boldsymbol{\lambda}}_{x0} + \sum_{y\in\mathcal{Y}} \mu^{\boldsymbol{\lambda}}_{0y} = \sum_{x\in\mathcal{X}} \hat{n}_x + \sum_{y\in\mathcal{Y}} \hat{m}_y - \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mu^{\boldsymbol{\lambda}}_{xy}$$

depends on the values of the parameters. Let $\hat{\mu}_{x0}$ (resp. $\hat{\mu}_{0y}$) be the number of single men (resp. women) of observed characteristics x (resp. y) in the sample; and $\hat{\mu}_{xy}$ the number of (x, y) couples[10]. It is easy to see that the log-likelihood of this sample can be written as

$$\log L(\boldsymbol{\lambda}) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \hat{\mu}_{xy} \log \frac{\mu^{\boldsymbol{\lambda}}_{xy}}{H^{\boldsymbol{\lambda}}} + \sum_{x\in\mathcal{X}} \hat{\mu}_{x0} \log \frac{\mu^{\boldsymbol{\lambda}}_{x0}}{H^{\boldsymbol{\lambda}}} + \sum_{y\in\mathcal{Y}} \hat{\mu}_{0y} \log \frac{\mu^{\boldsymbol{\lambda}}_{0y}}{H^{\boldsymbol{\lambda}}}.$$

[10] By construction, $\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \hat{\mu}_{xy} + \sum_{x\in\mathcal{X}} \hat{\mu}_{x0} + \sum_{y\in\mathcal{Y}} \hat{\mu}_{0y} = H$.

The maximum likelihood estimator $\hat{\boldsymbol{\lambda}}^{MLE}$ given by the maximization of $\log L$ is consistent, asymptotically normal, and asymptotically efficient under the usual set of assumptions.

5.2 Moment-based estimation in semilinear models

Maximum likelihood estimation allows for joint parametric estimation of the surplus function and of the unobserved heterogeneity. However, the log-likelihood may have several local extrema and it may be hard to maximize. We now introduce an alternative method, which is computationally very efficient but can only be used under two additional conditions. First, the distribution of the unobservable heterogeneity must be parameter-free—as it is in Choo and Siow (2006) for instance; or at least we conduct the analysis for fixed values of its parameters. Second, the parametrization of the $\boldsymbol{\Phi}$ matrix must be linear in the parameter vector:

$$\Phi^{\boldsymbol{\lambda}}_{xy} = \sum_{k=1}^{K} \lambda_k \phi^k_{xy} \tag{5.2}$$

where the parameter $\boldsymbol{\lambda} \in \mathbb{R}^K$, and $\tilde{\boldsymbol{\phi}} := (\boldsymbol{\phi}^{\mathbf{1}}, \ldots, \boldsymbol{\phi}^{\boldsymbol{K}})$ are K known linearly independent *basis surplus vectors*. If the number of basis surplus vectors is rich enough, this can approximate any surplus function. The *moment-matching estimator* of $\boldsymbol{\lambda}$ we propose in this section simply matches the moments predicted by the model with the empirical moments; that is, it solves the system

$$\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \hat{\mu}_{xy}\phi^k_{xy} = \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \mu^{\lambda}_{xy}\phi^k_{xy} \text{ for all } k. \tag{5.3}$$

Then the moment-matching estimator is

$$\hat{\boldsymbol{\lambda}}^{MM} := \arg\max_{\boldsymbol{\lambda}\in\mathbb{R}^K} \left(\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \hat{\mu}_{xy}\Phi^{\boldsymbol{\lambda}}_{xy} - \mathcal{W}\left(\boldsymbol{\Phi}^{\boldsymbol{\lambda}}, \hat{\boldsymbol{r}}\right) \right). \tag{5.4}$$

Since $\mathcal{W}$ is convex in $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^{\boldsymbol{\lambda}}$ is linear in $\boldsymbol{\lambda}$, the objective function in this program is globally concave. Moreover, equation (3.5) shows that the derivative of $\mathcal{W}$ with respect to Φ_{xy} is the corresponding μ_{xy}. It follows that the first-order conditions associated with (5.4) are (5.3). Appendix E shows how to derive a specification test from this program.

We show in Galichon and Salani (2021) that in the case of the Choo and Siow (2006) model, the moment matching estimator can be reformulated as a generalized linear model and estimated by a Poisson regression with two-sided fixed effects.

5.3 Minimum distance estimation

Finally, one can use (3.9) as the basis for a minimum distance estimator. That is, we write a mixed hypothesis as

$$\exists \boldsymbol{\lambda}, \quad \boldsymbol{D}^{\boldsymbol{\lambda}} \equiv \boldsymbol{\Phi}^{\boldsymbol{\lambda}} + \frac{\partial \mathcal{E}^{\boldsymbol{\lambda}}}{\partial \boldsymbol{\mu}};$$

and we choose $\hat{\boldsymbol{\lambda}}$ to minimize $\|\boldsymbol{D}^{\boldsymbol{\lambda}}\|^2_{\boldsymbol{\Omega}}$ for some positive definite matrix $\boldsymbol{\Omega}$. If we make the efficient choice $\boldsymbol{\Omega} = \left(V\boldsymbol{D}^{\boldsymbol{\lambda}}\right)^{-1}$, the minimized value of the squared norm follows a $\chi^2(p)$ if the model is well-specified, where $p = |X| \times |Y| - \dim(\boldsymbol{\lambda})$.

This is a particularly appealing strategy if the distributions $\mathbb{P}_x$ and $\mathbb{Q}_y$ are parameter-free and the surplus matrix $\boldsymbol{\Phi}^{\boldsymbol{\lambda}}$ is linear in the parameters, as the minimum distance estimator can then be implemented by linear least-squares.

6 Empirical Application

We tested our methods on Choo and Siow's original dataset, which they used to evaluate the impact of the *Roe vs Wade* 1973 Supreme Court abortion ruling on marriage patterns and on both genders' marriage market surpluses. A detailed description of the data can be found in Appendix G. Choo and Siow (2006) exploited two waves of surveys: one from the years 1970 to 1972, and one for 1980 to 1982. They distinguished those states in which abortion was already liberalized (the "reform states") from those where the Supreme Court ruling implied major legal changes. Our focus here is not on reexamining the effect of the ruling. We aim to test their chosen specification (a fully flexible surplus $\mathbf{\Phi}$ and iid type I EV errors) against some of the many other specifications that our analysis allows for. To do this, we select one of their subsamples. We chose to work with the 1970s wave, when couples married younger. This allows us to focus on the age range 16 to 40 with little loss[11]. We use the "non-reform states" subsample, which has 224,068 observations representing 13.3m individuals.

Our Proposition 4 implies that if we let the surplus $\mathbf{\Phi}$ be non-parametric as in Choo and Siow (2006), all separable models achieve an exact fit to the data. In that sense, there is no way to choose between say a nested logit model and a Random Scalar Coefficients model. To circumvent this issue, we proceed in two steps. First, we keep Choo and Siow's choice of error distribution but we fit several hundred parametric models of surplus to the data, using the semilinear model described in 5.2. We use the Bayesian Information Criterion (BIC) to select a set of basis functions (ϕ^k_{xy}), as described in Appendix G.3. We then fit alternative specifications to the data, using this set of basis functions and different distributions for the error terms.

6.1 Heteroskedastic Logit Models

We focused on specifications that allow for parameterized distributions of the error terms $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$. These parameters cannot be estimated by moment matching, which can only be used to estimate the coefficients of the basis functions for given values of the distributional parameters. One could maximize the resulting profile log-likelihood. Alternatively, the moment-matching equalities can be imposed as constraints in an MPEC approach. We have found that in practice, maximizing the log-likelihood over all parameters (distributional and coefficients of basis functions) worked well. This is the approach we use in the rest of this

[11] Choo and Siow (2006) allowed for marriage from ages 16 to 75. Our sample is 12% smaller.

section[12].

We explored several ways of adding heteroskedasticity to our benchmark model, while maintaining the scale normalization that is required in this two-sided discrete choice problem[13]. As reported in Appendix G.3, adding heteroskedasticity across genders barely improves the fit, and deteriorates the BIC value. On the other hand, we found that introducing heteroskedasticity on both gender and age does improve the value of the BIC. Our preferred model in this class replaces the term $\varepsilon_{iy} + \eta_{xj}$ with $\sigma_x \varepsilon_{iy} + \tau_y \eta_{xj}$, with $\sigma_x = \exp(\sigma_1 x)$, and $\tau_y = \exp(\tau_0)$. This still quite parsimonious model yields a noticeable improvement in the fit: +37.5 points of loglikelihood, and +25.2 points on BIC. The two distributional parameters are precisely estimated.

Our estimates give $\tau_y = 0.47$ and a σ_x that increases from 0.19 at age 16 to 5.29 at age 40; or, to focus on more likely ages at marriage for men in the early 1970s[14], from 0.28 at age 18 to 0.72 at age 25. This large relative variation directly impacts the shares of surplus that each partner can expect to get in a match. Simple calculations show that in this heteroskedastic version of the Choo and Siow (2006) model, the average share of the man in an (x, y) match is

$$\frac{u_x}{u_x + v_y} = \frac{\sigma_x \log \mu_{0|x}}{\sigma_x \log \mu_{0|x} + \tau_y \log \mu_{0|y}}.$$

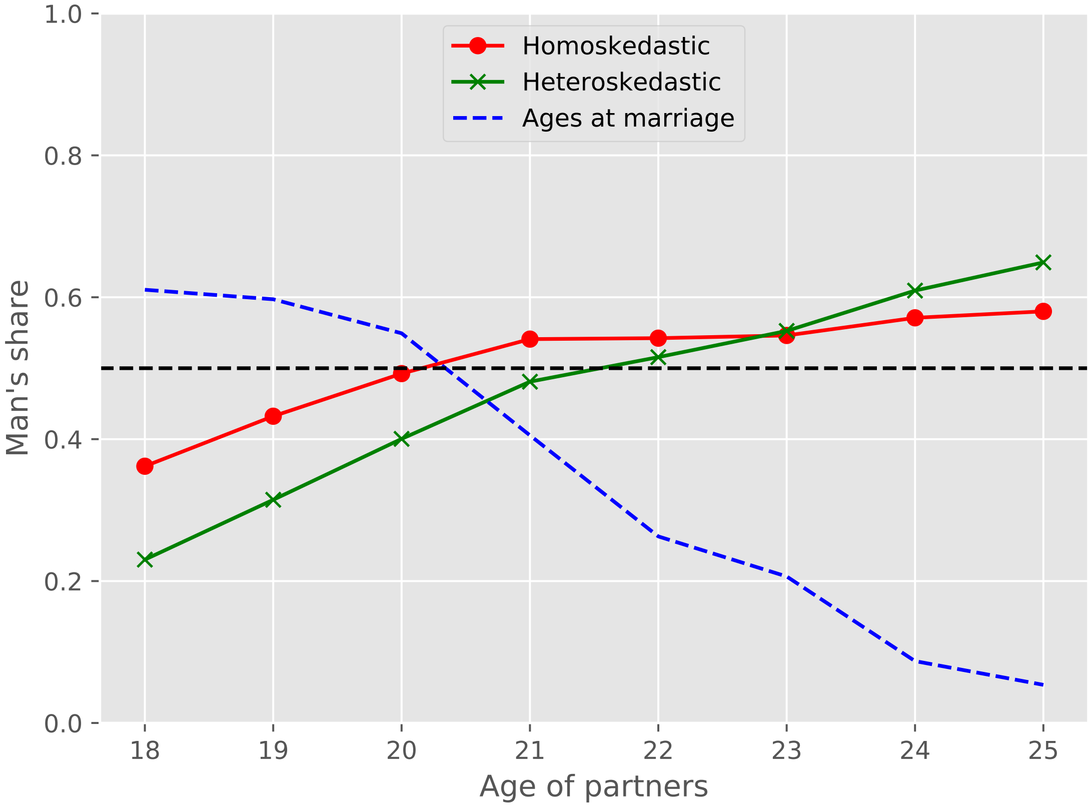
Figure 1 plots this ratio in the homoskedastic and in the heteroskedastic models for same-age couples ($x = y$). The surplus share of men clearly increases much more with age at marriage in the heteroskedastic version. Since the heteroskedastic model fits the data better, this suggests caution in interpreting the results of Choo and Siow (2006) on the effect of Roe vs Wade on the average utilities of men and women in marriage.

[12] The one difficulty we faced is in inverting the information matrix to compute the standard errors: the matrix has one or two very small eigenvalues that corresponds to two coefficients of the interactions of y and y^2 with $D = \mathbb{1}(x \geq y)$. We held them fixed when computing the standard errors.

[13] We normalize the standard error of ε to be 1 for a man of age 28—the midpoint in our sample.

[14] Recall that "age" is as recorded in 1970, while marriage occurs in 1971 or 1972.

Figure 1: Men's Share of the Marriage Surplus in the Logit Models



The dashed blue line indicate the number of same-age marriages. The dashed black line corresponds to equal sharing of the surplus.

6.2 Flexible Multinomial Logit Models

Nested logit models assign equal correlation between all the alternatives in a given nest. This is not well-suited to the kind of correlations we would like to capture[15]. What we need is a specification in which the preference shock for a partner of say age 22 is more positively correlated with the preference shock for a partner of age 23 than it is with the preference shock for a partner of age 29. In order to capture "age-local" correlations, we turned to the Flexible Coefficient Multinomial Logit (FC-MNL) model of Davis and Schiraldi (2014)[16]. This specification belongs to the class of Generalized Extreme Values models that we discussed in Appendix B.1. It allows for much more general substitution patterns between the different choices of partners, and in particular for "age-local" substitution patterns that we expect to find on the marriage market.

We estimated a few models of this family, along the lines suggested by Davis and Schiraldi (2014). All specifications we tried gave similar results; we present here the results we obtained where the matrix $\boldsymbol{b}$ that drives substitution patterns is given by

$$b^x_{y,y'} = \begin{cases} \frac{b_m(x)}{|y-y'|} & \text{if } y \neq y' \\ 1 & \text{if } y = y'; \end{cases}$$

where $b_m(x)$ is an affine function of the man's age. We used a similar specification on women's side, with an affine function $b_w(y)$ divided by $|x - x'|$.

The maximum likelihood estimator of this model achieves a meager gain of 0.5 point of the total loglikelihood over the basic Choo and Siow model. The affine functions are zero for the older men and women. Their estimated values for young men and women are positive but small[17]. Still, they do suggest more subtle patterns of substitution between partners than the Choo and Siow model allows for. We illustrate this on Figures 2 and 3. Figure 2 for instance plots the "demand semi-elasticities": $\partial \log \mu_{t|x} / \partial V_y$ for men whose age x goes

[15] We did estimate a simple two-level nested logit, and we found that the likelihood barely improves—see Appendix G.3.

[16] We thank Gautam Gowrisankaran for suggesting that we use this model.

[17] See Appendix G.3.

from 16 (in 1970) to 21. The horizontal and vertical axes represents partner's ages y and t (five on each side of x, with the obvious truncation.)

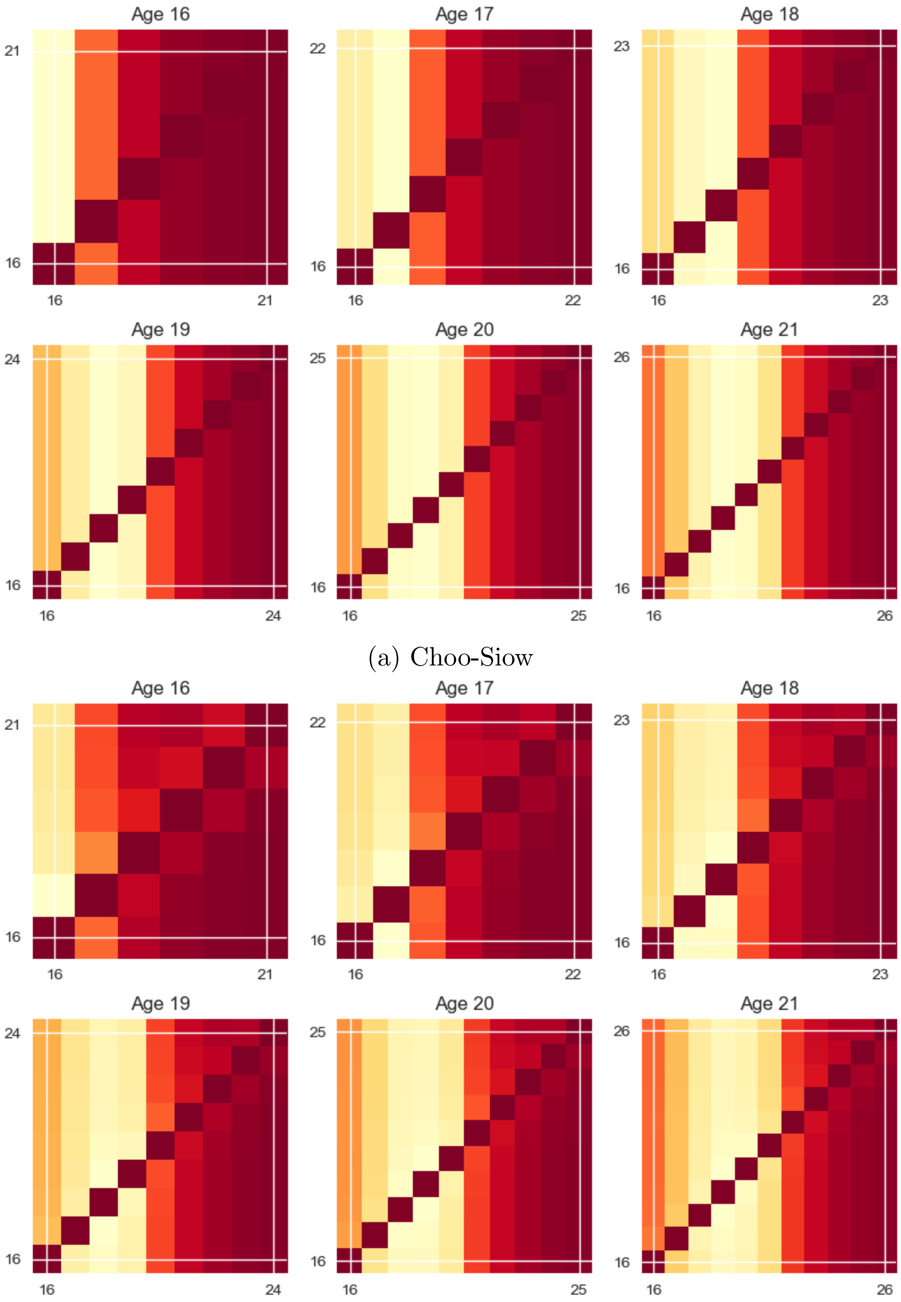
In the Choo and Siow model, the semi-elasticities are given by the usual formula:

$$\frac{\partial \log \mu_{t|x}}{\partial V_y} = \mathbb{1}(y = t) - \mu_{y|x}.$$

Aside from the diagonal $y = t$, the semi-elasticities do not depend on t. This appears as the vertical bands in the upper panel of Figure 2. The lower panel shows the same semi-elasticities for the FC-MNL model. Even with the small values of the b coefficients we estimate, richer substitution patterns appear. Figure 3 tells a similar story for women.

Concluding Remarks

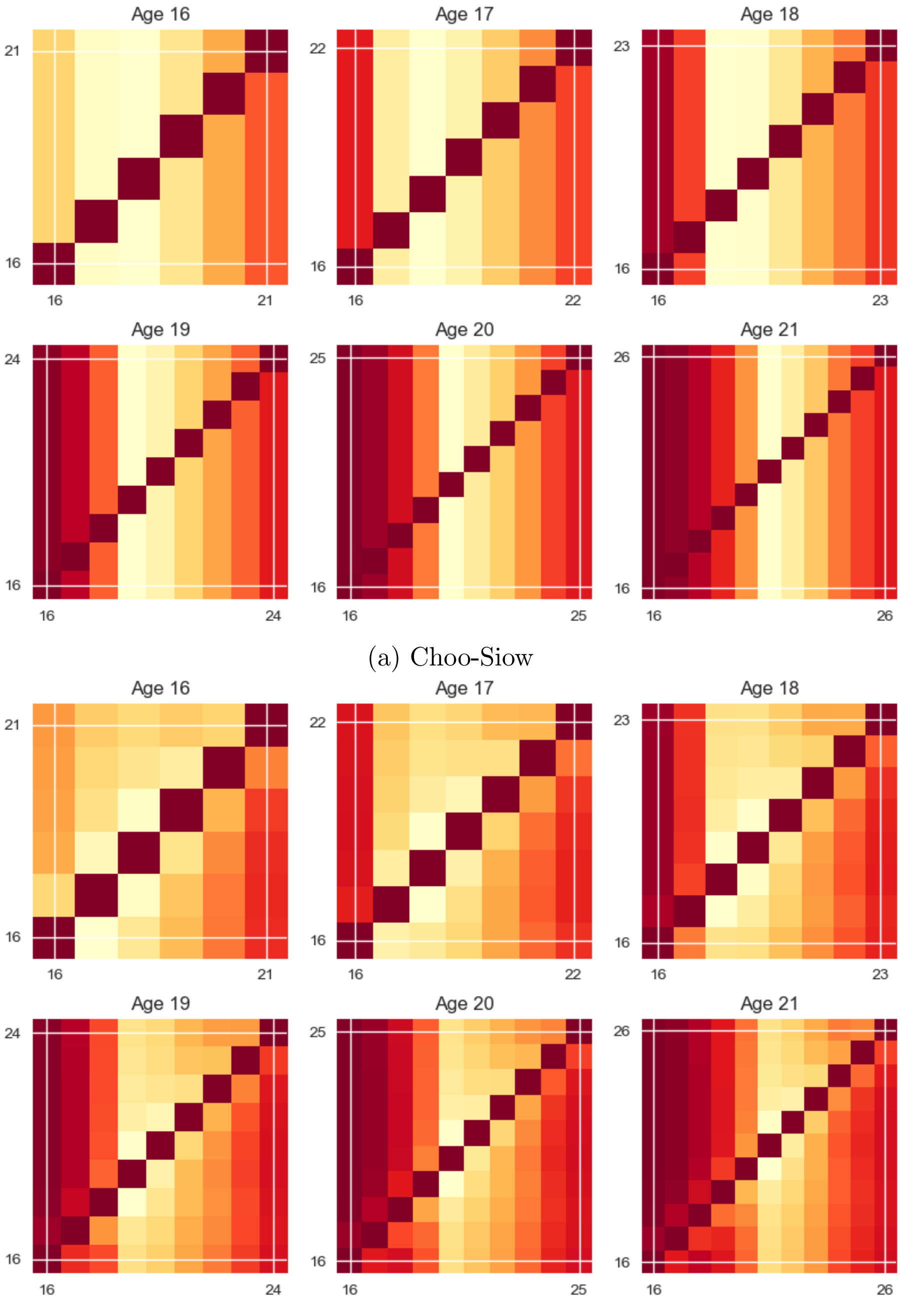
Several assumptions made in our paper, in particular the separability assumption and the large market assumption are tested on simulations by Chiappori, Nguyen, and Salanié (2019). We find these simulation results reassuring about the assumptions we have maintained in the present paper. Other assumptions we made in the present paper can also be dispensed with. In particular, one challenge is to extend our analysis to the case where the observable characteristics of the partners may be continuous. This issue is addressed by Dupuy and Galichon (2014) for the Choo and Siow model, using the theory of extreme value processes; they also propose a test of the number of relevant dimensions for the matching problem. Our results also open the way to applications beyond the bipartite, one-to-one matching framework of this paper. Chiappori, Galichon, and Salanié (2019) for instance describe a formal analogy between the "roommate" (non-bipartite) problem and the bipartite one-to-one model. We expect that this framework should also prove useful in the study of trading on networks, when transfers are allowed (thus providing an empirical counterpart to Hatfield and Kominers (2012) and Hatfield, Kominers, Nichifor, Ostrovsky, and Westkamp (2013)). Finally, our assumption that utility is fully transferable without frictions can be relaxed. Galichon, Kominers, and Weber (2019) study models with imperfectly transferable



(a) Choo-Siow

(b) FC-MNL

Figure 2: Semi-elasticities of substitution across partners: men



(a) Choo-Siow

(b) FC-MNL

Figure 3: Semi-elasticities of substitution across partners: women

utility and separable logit heterogeneity, while Galichon and Hsieh (2019) look at models with nontransferable utility and a similar form of heterogeneity.

References

Agarwal, N. (2015): "An Empirical Model of the Medical Match," *American Economic Review*, 105, 1939–1978.

Agarwal, N., and P. Somaini (2020): "Revealed Preference Analysis of School Choice Models," *Annual Review of Economics*, 12, 471–501.

Anderson, S., A. de Palma, and J.-F. Thisse (1988): "A Representative Consumer Theory of the Logit Model," *International Economic Review*, 29, 461–466.

Arcidiacono, P., and R. Miller (2011): "Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity," *Econometrica*, 79, 1823–1867.

Bajari, P., and J. Fox (2013): "Measuring the Efficiency of an FCC Spectrum Auction," *American Economic Journal: Microeconomics*, 5, 100–146.

Bauschke, H., and J. Borwein (1997): "Legendre Functions and the Method of Random Bregman Projections," *Journal of Convex Analysis*, 4, 27–67.

Becker, G. (1973): "A theory of marriage, part I," *Journal of Political Economy*, 81, 813–846.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890.

Berry, S., and A. Pakes (2007): "The Pure Characteristics Demand Model," *International Economic Review*, 48, 1193–1225.

Botticini, M., and A. Siow (2011): "Are there Increasing Returns in Marriage Markets?," IGIER Working Paper 395.

Boyd, S., and L. Vandenberghe (2004): *Convex Oprtimization.* Cambridge Universitry Press.

Byrd, R., J. Nocedal, and R. Waltz (2006): "KNITRO: An Integrated Package for Nonlinear Optimization," in *Large-Scale Nonlinear Optimization*, p. 3559. Springer Verlag.

Chernozhukov, V., A. Galichon, M. Hallin, and M. Henry (2017): "Monge-Kantorovich Depth, Quantiles, Ranks and Signs," *Annals of Statistics*, 45, 223–256.

Chiappori, P.-A. (2017): *Matching with Transfers: The Economics of Love and Marriage.* Princeton University Press.

——— (2020): "The Theory and Empirics of the Marriage Market," *Annual Review of Economics*, 12(1), 547–578.

Chiappori, P.-A., A. Galichon, and B. Salanié (2019): "On Human Capital and Team Stability," *Journal of Human Capital*, 13, 236–259.

Chiappori, P.-A., R. McCann, and L. Nesheim (2010): "Hedonic Price Equilibria, Stable Matching, and Optimal Transport: Equivalence, Topology, and Uniqueness," *Economic Theory*, 42, 317–354.

Chiappori, P.-A., D. L. Nguyen, and B. Salanié (2019): "Matching with Random Components: Simulations," Columbia University mimeo.

Chiappori, P.-A., and B. Salanié (2016): "The Econometrics of Matching Models," *Journal of Economic Literature*, 54, 832–861.

Chiappori, P.-A., B. Salanié, and Y. Weiss (2017): "Partner Choice, Investment in Children, and the Marital College Premium," *American Economic Review*, 107, 2109–67.

Chiong, K.-X., A. Galichon, and M. Shum (2016): "Duality in dynamic discrete-choice models," *Quantitative Economics*, 7, 83–115.

Choo, E., and A. Siow (2006): "Who Marries Whom and Why," *Journal of Political Economy*, 114, 175–201.

Ciscato, E., A. Galichon, and M. Goussé (2020): "Like Attract Like: A Structural Comparison of Homogamy Across Same-Sex and Different-Sex Households," *Journal of Political Economy*, 128, 740–781.

Costinot, A., and J. Vogel (2015): "Beyond Ricardo: Assignment Models in International Trade," *Annual Review of Economics*, 7, 31–62.

Csiszár, I. (1975): "*I*-divergence Geometry of Probability Distributions and Minimization Problems," *Annals of Probability*, 3, 146–158.

Dagsvik, J. (2000): "Aggregation in Matching Markets," *International Economic Review*, 41, 27–58.

Daly, A., and S. Zachary (1978): "Improved Multiple Choice Models," in *Identifying and Measuring the Determinants of Mode Choice*, ed. by D. Henscher, and Q. Dalvi. Teakfields, London.

Davis, P., and P. Schiraldi (2014): "The Flexible Coefficient Multinomial Logit (FC-MNL) Model of Demand for Differentiated Products," *Rand Journal of Economics*, 45, 32–63.

Debreu, G. (1960): "Review of R. D. Luce, *Individual choice behavior: A theoretical analysis*," *American Economic Review*, 50, 186–188.

Decker, C., E. Lieb, R. McCann, and B. Stephens (2012): "Unique Equilibria and Substitution Effects in a Stochastic Model of the Marriage Market," *Journal of Economic Theory*, 148, 778–792.

Dupuy, A., and A. Galichon (2014): "Personality traits and the marriage market," *Journal of Political Economy*, 122, 1271–1319.

Ekeland, I., J. Heckman, and L. Nesheim (2004): "Identification and Estimation of Hedonic Models," *Journal of Political Economy*, 112, S60–S109.

Fox, J. (2010): "Identification in Matching Games," *Quantitative Economics*, 1, 203–254.

——— (2018): "Estimating Matching Games with Transfers," *Quantitative Economics*, 8, 1–38.

Fox, J., C. Yang, and D. Hsu (2018): "Unobserved Heterogeneity in Matching Games with an Appplication to Venture Capital," *Journal of Political Economy*, 126, 1339–1373.

Gabaix, X., and A. Landier (2008): "Why Has CEO Pay Increased So Much?," *Quarterly Journal of Economics*, 123, 49–100.

Galichon, A. (2016): *Optimal Transport Methods in Economics*. Princeton University Press.

Galichon, A., and Y.-W. Hsieh (2019): "A model of decentralized matching markets without transfers," Unpublished manuscript.

Galichon, A., S. Kominers, and S. Weber (2019): "Costly Concessions: An Empirical Framework for Matching with Imperfectly Transferable Utility," *Journal of Political Economy*, 127, 2875–2925.

Galichon, A., and B. Salanié (2017): "The Econometrics and Some Properties of Separable Matching Models," *American Economic Review Papers and Proceedings*, 107, 251–255.

Galichon, A., and B. Salanié (2019): "Labeling Dependence in Separable Matching Markets," Columbia University mimeo.

Galichon, A., and B. Salani (2021): "Structural Estimation of Matching Markets with Transferable Utility," *Handbook of Market Design*, forthcoming.

Graham, B. (2011): "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers," in *Handbook of Social Economics*, ed. by J. Benhabib, A. Bisin, and M. Jackson. Elsevier.

Graham, B. (2013): "Uniqueness, Comparative Static, And Computational Methods for an Empirical One-to-one Transferable Utility Matching Model," *Structural Econometric Models*, 31, 153–181.

Graham, B. (2014): "Errata on "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers"," mimeo Berkeley.

Gretsky, N., J. Ostroy, and W. Zame (1992): "The Nonatomic Assignment Model," *Economic Theory*, 2, 103–127.

Gualdani, C., and S. Sinha (2019): "Partial Identification in Nonparametric One-to-One Matching Models," TSE Working Paper n. 19-993.

Hatfield, J., S. Kominers, A. Nichifor, M. Ostrovsky, and A. Westkamp (2013): "Stability and Competitive Equilibrium in Trading Networks," *Journal of Political Economy*, 121, 966–1005.

Hatfield, J. W., and S. D. Kominers (2012): "Matching in Networks with Bilateral Contracts," *American Economic Journal: Microeconomics*, 4, 176–208.

Hiriart-Urruty, J.-B., and C. Lemaréchal (2001): *Fundamentals of Convex Analysis.* Springer.

Hotz, J., and R. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60, 497–529.

Luce, R. D. (1959): *Games and Decisions.* New York: Wiley.

McFadden, D. (1978): "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Residential Location*, ed. by A. K. et al., pp. 75–96. North Holland.

Menzel, K. (2015): "Large Matching Markets as Two-Sided Demand Systems," *Econometrica*, 83, 897–941.

Mourifié, I. (2019): "A Marriage Matching Function with Flexible Spillover and Substitution Patterns," *Economic Theory*, 67, 421–461.

Mourifié, I., and A. Siow (2021): "The Cobb Douglas Marriage Matching function: Marriage Matching with Peer and Scale Effects," *Journal of Labor Economics*, 39, 239–274.

Ruggles, S., K. Genadek, R. Goeken, J. Grover, and M. Sobek (2015): "Integrated Public Use Microdata Series: Version 6.0," Discussion paper, Minneapolis: University of Minnesota.

Shapley, L., and M. Shubik (1972): "The Assignment Game I: The Core," *International Journal of Game Theory*, 1, 111–130.

Siow, A., and E. Choo (2006): "Estimating a Marriage Matching Model with Spillover Effects," *Demography*, 43, 463–490.

Tervio, M. (2008): "The difference that CEO make: An Assignment Model Approach," *American Economic Review*, 98, 642–668.

Train, K. E. (2009): *Discrete choice methods with simulation.* Cambridge University Press.

Williams, H. (1977): "On the Formulation of Travel Demand Models and Economic Measures of User Benefit," *Environment and Planning A*, 9, 285–344.

Appendix

A Proofs

A.1 Proof of Proposition 1

Denote by $(\tilde{u}_i)$ and $(\tilde{v}_j)$ the equilibrium utilities of men and women. Stability requires that for all (i, j),

- $\tilde{u}_i \geq \tilde{\Phi}_{i0}$, with equality if i is single
- $\tilde{v}_j \geq \tilde{\Phi}_{0j}$, with equality if j is single
- $\tilde{u}_i + \tilde{v}_j \geq \tilde{\Phi}_{ij}$, with equality if i and j are matched.

Let us focus on man i in group x. This man must be single or matched. If he is matched, then $\tilde{u}_i = \max_j \left(\tilde{\Phi}_{ij} - \tilde{v}_j\right)$; and by Assumption 1, we have $\tilde{\Phi}_{ij} = \Phi_{xy_j} + \varepsilon_{iy_j} + \eta_{xj}$ so that

$$\tilde{u}_i = \max_y \left(\Phi_{xy} + \varepsilon_{iy} + \max_{j:y_j=y} \left(\eta_{xj} - \tilde{v}_j\right)\right).$$

If he is single, then $\tilde{u}_i = \tilde{\Phi}_{i0} = \varepsilon_{i0}$.

Let $V_{xy} = \inf_{j:y_j=y} \left(\tilde{v}_j - \eta_{xj}\right)$ and $V_{x0} = 0$. Then

$$\tilde{u}_i = \max\left(\max_{y\in\mathcal{Y}} \left(\Phi_{xy} - V_{xy} + \varepsilon_{iy}\right), \varepsilon_{i0}\right) = \max_{y\in\mathcal{Y}_0} \left(\Phi_{xy} - V_{xy} + \varepsilon_{iy}\right).$$

Considering women would lead us to define $U_{xy} = \inf_{i:x_i=x}(\tilde{u}_i - \varepsilon_{iy})$ and $U_{0y} = 0$. Since $\tilde{\Phi}_{ij} = \Phi_{x_iy_j} + \varepsilon_{iy_j} + \eta_{x_ij}$ cannot be larger than $\tilde{u}_i + \tilde{v}_j$, we obtain

$$\Phi_{x_iy_j} \leq (\tilde{u}_i - \varepsilon_{iy_j}) + (\tilde{v}_j - \eta_{xj}); \tag{A.1}$$

taking lower bounds gives $\Phi_{xy} \leq U_{xy} + V_{xy}$. Finally, if $\mu_{xy} > 0$ then there is a couple (i, j) with $x_i = x, y_j = y$ for which (A.1) is an equality, so that $\Phi_{xy} = U_{xy} + V_{xy}$. ∎

A.2 Proof of Theorem 1

Replacing the expression of G given by (2.1) in formula (2.3) for G^* gives

$$-G^*(\boldsymbol{\mu}) = \inf_{\tilde{\boldsymbol{U}}} \left(-\sum_{y \in \mathcal{Y}_0} \mu_y \tilde{U}_y + \mathbb{E}_{\boldsymbol{P}} \max_{y \in \mathcal{Y}_0} \left(\varepsilon_y + \tilde{U}_y \right) \right)$$

where the minimization is over $\tilde{\boldsymbol{U}}$ such that $\tilde{U}_0 = 0$. The first term in the minimand can be seen as the expectation of the random variable $-\tilde{U}_Y$ under the distribution $Y \sim \mu_Y$. The term $\max_{y \in \mathcal{Y}_0} \left(\varepsilon_y + \tilde{U}_y \right)$ is the maximized utility of a man with mean utilities $\tilde{\boldsymbol{U}}$ and random taste shocks $\boldsymbol{\varepsilon}$. Alternatively, it is the value of the problem

$$\min \tilde{u} \text{ s.t. } \tilde{u} \geq \tilde{U}_y + \varepsilon_y \text{ for all } y \in \mathcal{Y}_0, \text{ with one equality.}$$

Therefore

$$\begin{aligned} -G^*(\boldsymbol{\mu}) = \inf_{\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{u}}} & \left(-\sum_{y \in \mathcal{Y}_0} \mu_y \tilde{U}_y + \mathbb{E}_{\boldsymbol{P}} \tilde{u}(\boldsymbol{\varepsilon}) \right) \\ \text{s.t. } & \tilde{u}(\boldsymbol{\varepsilon}) \geq \tilde{U}_y + \varepsilon_y \text{ for all } y \in \mathcal{Y}_0, \boldsymbol{\varepsilon}. \end{aligned}$$

Setting $V_y = -U_y$, we finally have

$$\begin{aligned} -G^*(\boldsymbol{\mu}) &= \inf_{\boldsymbol{V}, \tilde{u}} \left(\mathbb{E}_{\mu_Y} V_Y + \mathbb{E}_{\boldsymbol{P}} \tilde{u}(\boldsymbol{\varepsilon}) \right) \\ \text{s.t. } V_0 &= 0 \text{ and } V_y + \tilde{u}(\boldsymbol{\varepsilon}) \geq \varepsilon_y \quad \forall y \in \mathcal{Y}_0, \boldsymbol{\varepsilon} \in \operatorname{supp}(\boldsymbol{P}). \end{aligned}$$

This is exactly the value of the dual of an optimal transport problem in which the margins are μ_Y and $\boldsymbol{P}$ and the surplus ε_y is split into V_y and $\tilde{u}(\boldsymbol{\varepsilon})$. By the equivalence of the primal and the dual, this yields expression (2.6). ∎

A.3 Proof of Theorem 2

Since $\boldsymbol{P}$ has full support and is absolutely continuous, each y achieves the maximum with positive probability; the function G is strictly convex and by the envelope theorem, it is continuous differentiable and $\frac{\partial G}{\partial U_y}(\boldsymbol{U})$ is the probability that y achieves the maximum. This is just the classical Daly-Zachary-Williams theorem. By the same token, G^* is also strictly convex and continuously differentiable. The general theory of convex duality—or a straightforward application of the envelope theorem—tells us that $\mu_y = \left(\partial G / \partial \mu_y\right)(\boldsymbol{U})$ if and only if $U_y = \left(\partial G^* / \partial \mu_y\right)(\boldsymbol{\mu})$, which proves Part 2.

Now consider the strictly convex function $\tilde{\boldsymbol{U}} \longmapsto G\left(\tilde{\boldsymbol{U}}\right) - \sum_{y \in \mathcal{Y}} \mu_y \tilde{U}_y$. Part 3 follows from the fact that by the envelope theorem, $\boldsymbol{U}$ minimizes the value of this function if and only if $U_y = \left(\partial G^* / \partial \mu_y\right)(\boldsymbol{\mu})$. Since $G(\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{P}} \max_{y \in \mathcal{Y}_0}(U_y + \varepsilon_y)$, defining $\tilde{u}(\boldsymbol{\varepsilon})$ as in our proof of Theorem 1 yields (2.9). ∎

A.4 Proof of Theorem 3

In this proof we denote $\tilde{n}$ the distribution of $(x, \boldsymbol{\varepsilon})$ when the distribution of x is $\boldsymbol{n}$ and the distribution of $\boldsymbol{\varepsilon}$ conditional on x is $\boldsymbol{P}_x$. Formally, for $S \subseteq \mathcal{X} \times \mathbb{R}^{\mathcal{Y}_0}$, we get

$$\tilde{n}(S) = \sum_x n_x \int_{\mathbb{R}^{\mathcal{Y}_0}} \mathbb{1}(x, \boldsymbol{\varepsilon} \in S)\, d\boldsymbol{P}_x(\boldsymbol{\varepsilon}).$$

We define $\tilde{m}$ in the same way.

By the dual formulation of the matching problem (see Gretsky, Ostroy, and Zame (1992)), the value of total welfare in equilibrium is obtained by solving

$$\begin{aligned} \mathcal{W} = \inf_{\tilde{u}, \tilde{v}} \quad & \left(\int \tilde{u}(x, \boldsymbol{\varepsilon})\, d\tilde{n}(x, \boldsymbol{\varepsilon}) + \int \tilde{v}(y, \boldsymbol{\eta})\, d\tilde{m}(y, \boldsymbol{\eta}) \right) \\ \text{s.t.} \quad & \tilde{u}(x, \boldsymbol{\varepsilon}) + \tilde{v}(y, \boldsymbol{\eta}) \geq \Phi_{xy} + \varepsilon_y + \eta_x \;\; \forall (x, y, \boldsymbol{\varepsilon}, \boldsymbol{\eta}) \\ & \tilde{u}(x, \boldsymbol{\varepsilon}) \geq \varepsilon_0 \;\; \forall (x, \boldsymbol{\varepsilon}) \\ & \tilde{v}(y, \boldsymbol{\eta}) \geq \eta_0 \;\; \forall (y, \boldsymbol{\eta}). \end{aligned} \tag{A.2}$$

Fix any $\tilde{u}, \tilde{v}$ that satisfies all constraints in this program. As in the proof of Proposition 1, for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ we define

$$U_{xy} = \inf_{\boldsymbol{\varepsilon}} \left\{ \tilde{u}\left(x, \boldsymbol{\varepsilon}\right) - \varepsilon_y \right\} \text{ and } V_{xy} = \inf_{\boldsymbol{\eta}} \left\{ \tilde{v}\left(y, \boldsymbol{\eta}\right) - \eta_x \right\};$$

and we let $U_{x0} = V_{0y} = 0$. Then $\tilde{u}\left(x, \boldsymbol{\varepsilon}\right) \geq \max_{y \in \mathcal{Y}_0} \left\{ U_{xy} + \varepsilon_y \right\}$ and $\tilde{v}\left(y, \boldsymbol{\eta}\right) \geq \max_{x \in \mathcal{X}_0} \left\{ V_{xy} + \eta_x \right\}$; and the first constraint in (A.2) is simply $U_{xy} + V_{xy} \geq \Phi_{xy}$. Reciprocally, assume that $U_{x0} = V_{0y} = 0$ and $U_{xy} + V_{xy} \geq \Phi_{xy}$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and define

$$\tilde{u}\left(x, \boldsymbol{\varepsilon}\right) = \max_{y \in \mathcal{Y}_0} \left\{ U_{xy} + \varepsilon_y \right\} \quad \text{and} \quad \tilde{v}\left(y, \boldsymbol{\eta}\right) \geq \max_{x \in \mathcal{X}_0} \left\{ V_{xy} + \eta_x \right\};$$

Then $(\tilde{u}, \tilde{v})$ satisfies all constraints. Therefore we can rewrite the whole program as:

$$\begin{array}{rcl} \mathcal{W} & = & \displaystyle\min_{U,V} \left(\int \max_{y \in \mathcal{Y}_0} \left\{ U_{xy} + \varepsilon_y \right\} d\tilde{n}\left(x, \boldsymbol{\varepsilon}\right) + \int \max_{x \in \mathcal{X}_0} \left\{ V_{xy} + \eta_x \right\} d\tilde{m}\left(y, \boldsymbol{\eta}\right) \right) \\ \text{s.t.} & & U_{xy} + V_{xy} \geq \Phi_{xy} \ \forall x \in \mathcal{X}, y \in \mathcal{Y} \\ \text{and} & & U_{x0} = V_{0y} = 0 \ \forall x \in \mathcal{X}, y \in \mathcal{Y}. \end{array}$$

Now remember that we defined $G_x(\boldsymbol{U}_{x\cdot}) = \int \max_{y \in \mathcal{Y}_0} (U_{xy} + \varepsilon_y) dP_x(\boldsymbol{\varepsilon})$ and $G(\boldsymbol{U}, \boldsymbol{n}) = \sum_x n_x G_x(\boldsymbol{U}_{x\cdot})$. Under Assumption 1,

$$\left| \max_{y \in \mathcal{Y}_0} \left(U_{xy} + \varepsilon_y \right) \right| \leq \max_{y \in \mathcal{Y}_0} |U_{xy}| + \max_{y \in \mathcal{Y}_0} |\varepsilon_y|$$

is integrable, so that G_x is well-defined. It follows that

$$\begin{array}{rcl} \mathcal{W} & = & \displaystyle\min_{U,V} \left(G\left(\boldsymbol{U}, \boldsymbol{n}\right) + H\left(\boldsymbol{V}, \boldsymbol{m}\right) \right) \\ \text{s.t.} & & U_{xy} + V_{xy} \geq \Phi_{xy} \ \forall x \in \mathcal{X}, y \in \mathcal{Y} \end{array}$$

which is expression (3.6). Introducing multipliers (μ_{xy}), this convex minimization problem can be written in a minimax form as

$$\begin{aligned}\mathcal{W} &= \min_{\boldsymbol{U},\boldsymbol{V}} \max_{\mu\geq 0} \left(G\left(\boldsymbol{U},\boldsymbol{n}\right) + H\left(\boldsymbol{V},\boldsymbol{m}\right) + \sum_{xy} \mu_{xy}\Phi_{xy} - \sum_{xy} \mu_{xy}U_{xy} - \sum_{xy} \mu_{xy}V_{xy} \right) \\ &= \max_{\boldsymbol{\mu}\geq 0} \left(\sum_{xy} \mu_{xy}\Phi_{xy} - \max_{\boldsymbol{U},\boldsymbol{V}} \left(\sum_{xy} \mu_{xy}U_{xy} + \sum_{xy} \mu_{xy}V_{xy} - G\left(\boldsymbol{U},\boldsymbol{n}\right) - H\left(\boldsymbol{V},\boldsymbol{m}\right) \right) \right)\end{aligned}$$

which is (3.5); and (3.7) are its first-order conditions. ∎

A.5 Proof of Proposition 2

Part (i) and $U_{xy} + V_{xy} = \Phi_{xy}$ restate Proposition 1 (since Assumption 2 guarantees that $\mu_{xy} > 0$ for all $(x,y) \in \mathcal{A}$). For part (ii), note that applying the envelope theorem twice,

$$\frac{\partial \mathcal{W}}{\partial n_x} = -\frac{\partial G^*}{\partial n_x} = \frac{\partial G}{\partial n_x}$$

which equals G_x by the definition (3.1). Part (iii) is similar. ∎

A.6 Proof of Theorem 4

Part (i) follows from Theorem 2 (ii). Moreover, the $\boldsymbol{\mu}$'s are the multipliers in (3.6); since they are all positive, the constraints must be saturated, proving (ii). ∎

A.7 Extending the Entropy

Lemma 1 below is instrumental in the derivation of an efficient algorithm in Section 4.2.

While the generalized entropy $\mathcal{E}$ defined in (3.4) is concave in the matching patterns $\boldsymbol{\mu}$, it is only strictly concave when $\boldsymbol{\mu}$ has the margins $\boldsymbol{r}$ (otherwise $\mathcal{E}$ is infinite). We will need to extend it to a function that is strictly concave everywhere.

Definition 2 (Extended Entropy)**.** *Let $\mathcal{E}(\boldsymbol{\mu};\boldsymbol{r})$ be the generalized entropy of matching. We say that a function E* extends *$\mathcal{E}$ if it is a strictly concave function of $\boldsymbol{\mu}$ that coincides with over the set of feasible matchings $\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{r})$.*

There are many ways of extending a given generalized entropy function $\mathcal{E}$. Any choice of

$$E\left(\boldsymbol{\mu};\boldsymbol{r}\right)=\mathcal{E}\left(\boldsymbol{\mu},\sum_{y}\mu_{xy}+\mu_{x0},\sum_{x}\mu_{xy}+\mu_{0y}\right)+K\left(\boldsymbol{\mu};\boldsymbol{r}\right)$$

will work, where

$$K\left(\boldsymbol{\mu};\boldsymbol{r}\right)=\sum_{x}\left\{A_x\left(\sum_{y}\mu_{xy}+\mu_{x0}\right)-A_x\left(n_x\right)\right\}+\sum_{y}\left\{B_y\left(\sum_{x}\mu_{xy}+\mu_{0y}\right)-B_y\left(m_y\right)\right\}, \tag{A.3}$$

and A_x and B_y are concave functions from $\mathbb{R}$ to $\mathbb{R}$. Defining E in this way ensures that it coincides with $\mathcal{E}(\boldsymbol{\mu},\boldsymbol{r})$ for any feasible matching; and adding the term K makes E strictly concave in $\boldsymbol{\mu}$.

Lemma 1. *Let E extend $\mathcal{E}$. For $\boldsymbol{u}\in\mathbb{R}^{\mathcal{X}}$ and $\boldsymbol{v}\in\mathbb{R}^{\mathcal{Y}}$, define $S(\boldsymbol{u},\boldsymbol{v})$ as the value of*

$$\max_{\boldsymbol{\mu}}\left(E(\boldsymbol{\mu};\boldsymbol{r})+\sum_{x,y\in\mathcal{X}\times\mathcal{Y}}\mu_{xy}(\Phi_{xy}-u_x-v_y)+\sum_{x\in\mathcal{X}}(n_x-\mu_{x0})u_x+\sum_{y\in\mathcal{Y}}(m_y-\mu_{0y})v_y\right). \tag{A.4}$$

Then S is a convex function of $(\boldsymbol{u},\boldsymbol{v})$. The social welfare $\mathcal{W}$ is its minimum value; the minimizers $\boldsymbol{u}$ and $\boldsymbol{v}$ are the average utilities of the different types of men and women in equilibrium; and the solutions $\boldsymbol{\mu}$ to (A.4) at $(\boldsymbol{u},\boldsymbol{v})$ are the equilibrium matching patterns.

Proof. Recall from equation (3.5) that the equilibrium matching $\boldsymbol{\mu}$ maximizes $\sum_{x,y}\mu_{xy}\Phi_{xy}+\mathcal{E}(\boldsymbol{\mu},\boldsymbol{r})$ over μ in $\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$. This can be rewritten as

$$\begin{aligned}
\max_{\boldsymbol{\mu}}\quad & \sum_{x,y\in\mathcal{X}\times\mathcal{Y}}\mu_{xy}\Phi_{xy}+E(\boldsymbol{\mu};\boldsymbol{r}) \\
s.t.\quad & \mu_{x0}+\sum_{y\in\mathcal{Y}}\mu_{xy}=n_x \\
& \mu_{0y}+\sum_{x\in\mathcal{X}}\mu_{xy}=m_y.
\end{aligned} \tag{A.5}$$

Denote a_x and b_y the multipliers of the constraints. The Lagrangian of (A.5) can be written as

$$\mathcal{L} = \max_{\boldsymbol{\mu}} \min_{\boldsymbol{a},\boldsymbol{b}} \left(\begin{array}{c} \sum_{x,y \in \mathcal{X}\times\mathcal{Y}} \mu_{xy}\Phi_{xy} + E(\boldsymbol{\mu};\boldsymbol{r}) \\ -\sum_{x\in\mathcal{X}} a_x \left(\mu_{x0} + \sum_{y\in\mathcal{Y}} \mu_{xy} - n_x\right) - \sum_{y\in\mathcal{Y}} b_y \left(\mu_{0y} + \sum_{x\in\mathcal{X}} \mu_{xy} - m_y\right) \end{array} \right)$$

$$= \max_{\boldsymbol{\mu}} \min_{\boldsymbol{a},\boldsymbol{b}} \left(\begin{array}{c} \sum_{x,y \in \mathcal{X}\times\mathcal{Y}} \mu_{xy} \left(\Phi_{xy} - a_x - b_y\right) + E(\boldsymbol{\mu};\boldsymbol{r}) \\ +\sum_{x\in\mathcal{X}} a_x \left(n_x - \mu_{x0}\right) + \sum_{y\in\mathcal{Y}} b_y \left(m_y - \mu_{0y}\right) \end{array} \right).$$

Interchanging min and max gives $\mathcal{L} = \min_{\boldsymbol{a},\boldsymbol{b}} S(\boldsymbol{a},\boldsymbol{b};\boldsymbol{\Phi},\boldsymbol{r})$, where S is defined in the corollary. It is a maximum of linear functions of $\boldsymbol{a},\boldsymbol{b}$ and therefore convex. Since the constraints are binding at the optimum, $\mathcal{W} = \mathcal{L}$. Moreover, by the envelope theorem $\frac{\partial \mathcal{W}}{\partial n_x} = \frac{\partial S}{\partial n_x} = a_x$. By Proposition 2, this gives $a_x = u_x$; and the μ's are the corresponding matching patterns.

A.8 Proof of Theorem 5

We start by extending the generalized entropy $\mathcal{E}$ to a strictly concave function E as explained in A.7. For notational simplicity, we now drop the arguments $\boldsymbol{r}$ and $\boldsymbol{\Phi}$. Proposition 1 shows that the value of the matching problem is $\min_{\boldsymbol{a},\boldsymbol{b}} S(\boldsymbol{a},\boldsymbol{b})$. We solve for the minimum iteratively by coordinate descent. At step $2k$, we first fix $\boldsymbol{b} = \boldsymbol{b}^{(2k)}$ and we solve the convex minimization problem over $\boldsymbol{a}$ only:

$$\boldsymbol{a}^{(2k+1)} \equiv \arg\min_{\boldsymbol{a}} S(\boldsymbol{a},\boldsymbol{b}^{(2k)}).$$

Then we keep $\boldsymbol{a} = \boldsymbol{a}^{(2k+1)}$ fixed at this new value and we solve the minimization problem over $\boldsymbol{b}$:

$$\boldsymbol{b}^{(2k+2)} \equiv \arg\min_{\boldsymbol{b}} S(\boldsymbol{a}^{(2k+1)},\boldsymbol{b}).$$

We stop the iterations when $\boldsymbol{b}^{(2k+2)}$ and $\boldsymbol{b}^{(2k)}$ are close enough. We take $\boldsymbol{u}^{(2k+1)}$ and $\boldsymbol{v}^{(2k+2)}$ to be the average utilities, and the associated $\boldsymbol{\mu}$ to be the equilibrium matching patterns.

Let us now prove that the algorithm converges to the global minimum $(\boldsymbol{u},\boldsymbol{v})$ of S. We rely on results in Bauschke and Borwein (1997), which builds on Csiszár (1975). The map $\boldsymbol{\mu} \to -E(\boldsymbol{\mu})$ is smooth and strictly convex; hence it is a "Legendre function" in their terminology. Introduce the associated "Bregman divergence" D as

$$D\left(\boldsymbol{\mu},\bar{\boldsymbol{\nu}}\right)=E\left(\bar{\boldsymbol{\nu}}\right)-E\left(\boldsymbol{\mu}\right)+\left\langle \nabla E\left(\bar{\boldsymbol{\nu}}\right),\boldsymbol{\mu}-\bar{\boldsymbol{\nu}}\right\rangle ,$$

where ∇ denotes the gradient wrt $\bar{\boldsymbol{\nu}}$; and define the linear subspaces $\mathcal{L}\left(\boldsymbol{n}\right)$ and $\mathcal{L}\left(\boldsymbol{m}\right)$ by

$$\mathcal{L}\left(\boldsymbol{n}\right)=\{\boldsymbol{\mu}\geq 0:\forall x\in\mathcal{X},\ \sum_{y\in\mathcal{Y}_0}\mu_{xy}=n_x\}\text{ and }\mathcal{L}\left(\boldsymbol{m}\right)=\{\boldsymbol{\mu}\geq 0:\forall y\in\mathcal{Y},\ \sum_{x\in\mathcal{X}_0}\mu_{xy}=m_y\}$$

so that $\mathcal{M}(\boldsymbol{r})=\mathcal{L}\left(\boldsymbol{n}\right)\cap\mathcal{L}\left(\boldsymbol{m}\right)$. It is easy to see that $\boldsymbol{\mu}^{(k)}$ results from iterative projections with respect to D on the linear subspaces $\mathcal{L}(\boldsymbol{n})$ and $\mathcal{L}(\boldsymbol{m})$:

$$\boldsymbol{\mu}^{(2k+1)}=\arg\min_{\boldsymbol{\mu}\in\mathcal{L}(\boldsymbol{n})}D\left(\boldsymbol{\mu},\boldsymbol{\mu}^{(2k)}\right)\text{ and }\boldsymbol{\mu}^{(2k+2)}=\arg\min_{\boldsymbol{\mu}\in\mathcal{L}(\boldsymbol{m})}D\left(\boldsymbol{\mu},\boldsymbol{\mu}^{(2k+1)}\right).\tag{A.6}$$

By Theorem 8.4 of Bauschke and Borwein, the iterated projection algorithm converges to the projection $\boldsymbol{\mu}$ of $\boldsymbol{\mu}^{(0)}$ on $\mathcal{M}(\boldsymbol{r})$, which is also the maximizer $\boldsymbol{\mu}$ of (3.5).

As mentioned earlier, there are many possible ways of extending $\mathcal{E}$ to E, depending on the choice of the functions A_x and B_y in (A.3). In practice, good judgement should be exercised, as the choice of an extension E that makes it easy to solve the systems in A.6 is crucial for the performance of the algorithm.

B Examples of random utility models

B.1 The Generalized Extreme Value Framework

Consider a function $g:\mathbb{R}^{\mathcal{Y}_0}\to\mathbb{R}$ that (i) is positive homogeneous of degree one; (ii) goes to $+\infty$ whenever any of its arguments goes to $+\infty$; (iii) has partial derivatives (outside of $\mathbf{0}$) at any order k of sign $(-1)^k$; (iv) is such that the function defined by $F\left(w_0,\ldots,w_{|\mathcal{Y}|}\right)=\exp\left(-g\left(e^{-w_0},\ldots,e^{-w_{|\mathcal{Y}|}}\right)\right)$ is a multivariate cumulative distribution function associated to some distribution, which we denote $\boldsymbol{P}$. Then introducing utility shocks $\boldsymbol{\varepsilon} \sim \boldsymbol{P}$, we have by a theorem of McFadden (1978):

$$G(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{P}}\left[\max_{y\in\mathcal{Y}_0}\{w_y + \varepsilon_y\}\right] = \log g\left(e^{\boldsymbol{w}}\right) + \gamma \tag{B.1}$$

where γ is the Euler constant $\gamma \simeq 0.577$.

For any vector $\boldsymbol{p} \in \mathbb{R}^{\mathcal{Y}}$ such that $\sum_{y\in\mathcal{Y}} p_y = 1$, we denote $\bar{\boldsymbol{p}} = (p_1, \ldots, p_{|\mathcal{Y}|})$. Then

$$G^*\left(\bar{\boldsymbol{p}}\right) = \log g\left(e^{\boldsymbol{w}(\boldsymbol{p})}\right) + \gamma - \sum_{y\in\mathcal{Y}_0} p_y w_y\left(\boldsymbol{p}\right),$$

where the vector $\boldsymbol{w}\left(p\right)$ solves the system of equations

$$p_y = \frac{\partial \log g}{\partial w_y}\left(e^{\boldsymbol{w}}\right) \quad \text{for all} \quad y \in \mathcal{Y}_0. \tag{B.2}$$

Now take a vector $\boldsymbol{\mu} = (\mu_y)_{y\in\mathcal{Y}}$ such that $\sum_{y\in\mathcal{Y}} \mu_y \leq 1$. The generalized entropy of choice arising from this heterogeneity is

$$G^*(\boldsymbol{\mu}) = \log g\left(e^{\boldsymbol{w}(\boldsymbol{\mu})}\right) - \sum_{y\in\mathcal{Y}_0} \mu_y w_y\left(\boldsymbol{\mu}\right) + \gamma. \tag{B.3}$$

Applying the envelope theorem, the derivative of this expression with respect to μ_y is $-w_y\left(\boldsymbol{\mu}\right)$. Therefore the $\boldsymbol{U}$ vector is identified by

$$U_y = w_y\left(\boldsymbol{\mu}\right). \tag{B.4}$$

B.2 The nested logit model

We consider the two-layer nested logit model of Example 2.1: alternative 0 is alone in a nest and each other nest $n \in \mathcal{N}$ contains alternatives $y \in \mathcal{Y}\left(n\right)$. The correlation of alternatives whithin nest n is proxied by $(1 - \lambda_n^2)$.

B.2.1 The entropy of choice of the one-sided nested logit model

It is well-known that[18]

$$G(\boldsymbol{U}) = \log\left(1 + \sum_{n\in\mathcal{N}} \exp(I_n(\boldsymbol{U}))\right)$$

where $I_n(\boldsymbol{U}) \equiv \lambda_n \log\left(\sum_{y\in\mathcal{Y}(n)} \exp(U_y/\lambda_n)\right)$ is the *inclusive value* of nest n. For $y \in \mathcal{Y}_n$, this gives

$$\mu_y = \frac{\partial G}{\partial U_y}(\boldsymbol{U}) = \mu_n \times \frac{\exp(U_y/\lambda_n)}{\exp\left(I_n(\boldsymbol{U})/\lambda_n\right)},$$

where

$$\mu_n := \sum_{y\in\mathcal{Y}(n)} \mu_y = \frac{\exp\left(I_n(\boldsymbol{U})\right)}{1 + \sum_{m\in\mathcal{N}} \exp\left(I_m(\boldsymbol{U})\right)}.$$

As a result, $\log\mu_n = I_n(\boldsymbol{U}) - G(\boldsymbol{U})$ and $\log\mu_y = \log\mu_n + (U_y - I_n(\boldsymbol{U}))/\lambda_n$. Moreover,

$$\mu_0 = 1 - \sum_{n\in\mathcal{N}} \mu_n = \exp(-G(\boldsymbol{U})),$$

so that we can solve for

$$\begin{aligned} G(\boldsymbol{U}) &= -\log\mu_0 \\ I_n(\boldsymbol{U}) &= \log(\mu_n/\mu_0) \\ U_y &= \lambda_n \log\frac{\mu_y}{\mu_0} + (1-\lambda_n)\log\frac{\mu_n}{\mu_0}. \end{aligned} \tag{B.5}$$

Since $G^*(\boldsymbol{\mu}) = \sum_{y\neq 0} \mu_y U_y - G(\boldsymbol{U})$ at the optimum, this gives

$$\begin{aligned} G^*(\boldsymbol{\mu}) = \sum_{n\in\mathcal{N}} \lambda_n \sum_{y\in\mathcal{Y}(n)} \mu_y \log\mu_y - \left(\sum_{y\neq 0}\mu_y\right)\log\mu_0 \\ + \sum_{n\in\mathcal{N}} (1-\lambda_n)\left(\sum_{y\in\mathcal{Y}(n)} \mu_y\right)\log\mu_n + \log\mu_0; \end{aligned}$$

[18] We omit the Euler constant γ from now on, as it plays no role in any of our calculations.

using $\sum_{y\neq 0}\mu_y = 1-\mu_0$ and $\sum_{y\in\mathcal{Y}(n)}\mu_y = \mu_n$, we get the generalized entropy of choice

$$G^*(\boldsymbol{\mu}) = \sum_{n\in\mathcal{N}}\left(\lambda_n \sum_{y\in\mathcal{Y}(n)} \mu_y \log\mu_y + (1-\lambda_n)\mu_n\log\mu_n\right) + \mu_0\log\mu_0.$$

B.2.2 The two-sided nested logit model

Now suppose that the above (indexed by x as $\lambda_n^x, \mathcal{N}^x, \mathcal{Y}^x(n)$) describes the structure of errors for men of group x, and that women of group y have a similar error structure with parameters $\nu_n^y, \mathcal{N}^y, \mathcal{X}^y(n)$. We denote $n(y;x)$ the nest of partner group y for men of group x, and $n(x;y)$ the nest of partner group x for women of group y. Then the matrix $\boldsymbol{U}$ is identified as

$$U_{xy} = \lambda^x_{n(y;x)}\log\frac{\mu_{xy}}{\mu_{x0}} + \left(1-\lambda^x_{n(y;x)}\right)\log\frac{\mu_{x,n(y;x)}}{\mu_{x0}}.$$

Along with the corresponding formula for $\boldsymbol{V}$, this identifies the joint surplus as

$$\begin{aligned}\Phi_{xy} &= (\lambda^x_{n(y;x)} + \nu^y_{n(x;y)})\log\mu_{xy} - \log\mu_{x0} - \log\mu_{0y}\\ &+ \left(1-\lambda^x_{n(y;x)}\right)\log\mu_{x,n(y;x)} + \left(1-\nu^y_{n(x;y)}\right)\log\mu_{n(x;y),y}\end{aligned}$$

for any given values of the parameters of the nested logit errors.

B.3 The random coefficients logit model

Recall that Example 2.2 had $\boldsymbol{\varepsilon} = \boldsymbol{Ze} + T\boldsymbol{\eta}$, where $\boldsymbol{e}$ is a random vector on $\mathbb{R}^d$ with distribution $\mathbf{P}_\epsilon$; $\boldsymbol{Z}$ is a $|\mathcal{Y}_0|\times d$ matrix; $T>0$; and $\boldsymbol{\eta}$ is an extreme value type-I (Gumbel) random variable i.i.d. on $\mathbb{R}^{\mathcal{Y}_0}$ and independent from $\boldsymbol{e}$.

By the law of iterated expectations, making use of the independence of $\boldsymbol{e}$ and $\boldsymbol{\eta}$, we get

$$\begin{aligned} G(\boldsymbol{U}) &= \mathbb{E}\left[\mathbb{E}\left[\max_{y\in\mathcal{Y}_0}\left\{U_y + (\boldsymbol{Ze})_y + T\eta_y\right\}|\boldsymbol{e}\right]\right] && \text{(B.6)}\\ &= \int G_0\left(U+\boldsymbol{Ze}\right)f\left(e\right)de && \text{(B.7)}\end{aligned}$$

where

$$G_0\left(\boldsymbol{U}\right)=T\log\sum_{y\in\mathcal{Y}_0}\exp\left(\frac{U_y}{T}\right)$$

is the Emax operator associated with the plain multinomial logit model. It is easy to compute its convex conjugate: $G_0^*\left(\boldsymbol{\pi}\right)=T\sum_y\pi_y\log\pi_y$ if $\sum_y\pi_y=1$, and $+\infty$ otherwise.

We will use two well-known properties of convex conjugates (see e.g. Hiriart-Urruty and Lemaréchal, 2001, part E):

- the convex conjugate of a translated function $\boldsymbol{x}\rightarrow g_t(\boldsymbol{x})\equiv g(\boldsymbol{x}+\boldsymbol{t})$ is $g_t^*(\boldsymbol{y})=g^*(\boldsymbol{y})+\boldsymbol{y}\cdot\boldsymbol{t}$
- the convex conjugate of a sum of convex functions is the infimum-convolution of their convex conjugates:

$$(f_1+f_2)^*(\boldsymbol{y})=\inf_{\boldsymbol{y}_1+\boldsymbol{y}_2=\boldsymbol{y}}\left(f_1^*(\boldsymbol{y}_1)+f_2^*(\boldsymbol{y}_2)\right).$$

Together, they imply that

$$G^*\left(\boldsymbol{\mu}\right)=\inf_{\boldsymbol{\pi}(\cdot)\geq 0}\left\{\int G_0^*\left(\boldsymbol{\pi}(\boldsymbol{e})\right)d\boldsymbol{P}_{\boldsymbol{e}}(\boldsymbol{e})-\sum_y\int\left(Ze\right)_y\pi_y\left(\boldsymbol{e}\right)d\boldsymbol{P}_{\boldsymbol{e}}(\boldsymbol{e}):\ \int_{\boldsymbol{e}}\pi_y\left(\boldsymbol{e}\right)d\boldsymbol{P}_{\boldsymbol{e}}(\boldsymbol{e})=\mu_y\ \forall y\right\}. \tag{B.8}$$

It follows that

$$\begin{aligned}-G^*\left(\boldsymbol{\mu}\right) &= \max_{\boldsymbol{\pi}(\cdot)\geq 0}\left\{\sum_y\int\left(Ze\right)_y\pi_y\left(\boldsymbol{e}\right)-T\sum_y\pi_y\left(\boldsymbol{e}\right)\log\pi_y\left(\boldsymbol{e}\right)\right\}d\boldsymbol{P}_{\boldsymbol{e}}(\boldsymbol{e})\\ s.t. & \quad \int\pi_y\left(\boldsymbol{e}\right)d\boldsymbol{P}_{\boldsymbol{e}}(\boldsymbol{e})=\mu_y\ \forall y\\ & \quad \sum_y\pi_y\left(\boldsymbol{e}\right)=1\ \forall\boldsymbol{e}.\end{aligned}$$

This is an optimal transport problem with entropic regularization, (see Galichon, 2016, Chapter 7). In the absence of the second term in the objective function, it would be an optimal transport problem between the discrete random variable $Y\sim\boldsymbol{\mu}$ and the continuous random vector $\boldsymbol{e} \sim \boldsymbol{P}_{\boldsymbol{e}}$, with transport surplus $(y, \boldsymbol{e}) \to -(\boldsymbol{Z}\boldsymbol{e})_y$. The second term is an entropic regularization.

B.4 The pure characteristics model

The second part of Example 2.2 is obtained by setting $T = 0$ in (B.7). The regularization term in (B.8) disappears, and

$$G^*(\boldsymbol{\mu}) = \max_{\boldsymbol{\pi} \in \mathcal{M}} \sum_{y \in \mathcal{Y}_0} \mu_y \int_{\boldsymbol{e} \in \mathbb{R}^d} -(Z\boldsymbol{e})\, d\boldsymbol{P}_{\boldsymbol{e}}(\boldsymbol{e}) \tag{B.9}$$

which is a standard optimal transport problem (this time without the entropic regularization) between a discrete random variable on $\mathbb{R}^d$ $\tilde{z}$ such that $\tilde{z}_i = Z_{\tilde{y}i}$ where $\tilde{y} \sim \mu$, and the continuous random variable $\boldsymbol{e} \sim \boldsymbol{P}$, where the transport surplus is now the scalar product $(z, \boldsymbol{e}) \to z^\top \epsilon$. This is exactly the power diagram situation described in Chapter 5 of Galichon (2016).

B.5 The FC-MNL Model

Davis and Schiraldi (2014) introduced a flexible GEV specification which they called the Flexible Coefficients-Multinomial Choice Model.

Example B.1 (FC-MNL)**.** *The function g that appears in (B.1) takes the following form:*

$$g(\boldsymbol{t}) = \sum_{(y,y') \in \mathcal{Y}_0^2} b_{y,y'} \left(\frac{t_y^{1/\sigma} + t_{y'}^{1/\sigma}}{2} \right)^{\tau\sigma}$$

where $(b_{y,y'})$ is a non-negative symmetric matrix, and the parameters satisfy the inequalities $0 < \sigma < 1$, $\tau > 1$, $\tau\sigma \leq 1$. We can set $b_{yy} = 1$ for every y. Note that we recover the standard multinomial logit model when $\boldsymbol{b}$ is the identity matrix.

We followed Davis and Schiraldi (2014) in making g a τ-homogeneous function, rather than 1-homogeneous. This is a harmless normalization. It gives

$$G(\boldsymbol{U}) = \frac{1}{\tau}\left(\log \sum_{(y,y')\in\mathcal{Y}_0^2} b_{y,y'}\left(\frac{\exp(U_y/\sigma)+\exp(U_{y'}/\sigma)}{2}\right)^{\tau\sigma}\right) + \gamma.$$

While this may look forbidding, it is easy to evaluate and it yields simple demands:

$$\mu_y = \frac{1}{g}\exp(U_y/\sigma)\sum_{y'\in\mathcal{Y}_0} b_{y,y'}\left(\frac{\exp(U_y/\sigma)+\exp(U_{y'}/\sigma)}{2}\right)^{\tau\sigma-1}.$$

It is apparent from the formulæ that the "cross-price elasticities" (the dependence of $\boldsymbol{\mu}$ on $\boldsymbol{U}$ are largely driven by the matrix $\boldsymbol{b}$.) In fact Davis and Schiraldi (2014) show that for any fixed σ and τ, $\boldsymbol{b}$ can be chosen to replicate any given set of own- and cross-price elasticities.

Suggested online appendices

C More on the assumptions [online]

In this online appendix, we discuss the separability assumption (which we maintain throughout), and the type I extreme value assumption of Choo and Siow (2006) (which we relax).

C.1 The separability assumption

Assumption 1 imposes that the matching surplus $\tilde{\Phi}$ be separable in the sense that

$$\tilde{\Phi}_{ij} = \Phi_{xy} + \varepsilon_{iy} + \eta_{xj}.$$

It is easy to see that Assumption 1 is equivalent to the follwing:

Assumption 3 (Separability restated)**.** *If two men i and i' belong to the same group x, and their respective partners j and j' belong to the same group y, then the total surplus generated by these two matches is unchanged if partners are shuffled:*

$$\tilde{\Phi}_{ij} + \tilde{\Phi}_{i'j'} = \tilde{\Phi}_{ij'} + \tilde{\Phi}_{i'j}.$$

It should be clear from this equivalent definition that we need not adopt Choo and Siow's original interpretation, in which $\boldsymbol{\varepsilon}$ was a vector of preference shocks of the husband and $\boldsymbol{\eta}$ was a vector of preference shocks of the wife. More precisely, they assumed that the utility of a man i of group x who marries a woman j of group y was given by

$$\alpha_{xy} + \tau + \varepsilon_{iy}, \tag{C.1}$$

where α_{xy} was the "systematic" part of the surplus; τ represented the utility transfer (possibly negative) that the husband gets from his partner in equilibrium; and ε_{iy} was a standard type I extreme value random term[19]. The utility of this man's wife would be

[19] For a single, $\alpha_{x0} = \tau = 0$.

written as

$$\gamma_{xy} - \tau + \eta_{xj}. \tag{C.2}$$

This formulation clearly implies separability, but it is much stronger than we need. To take an extreme example, assume that men are indifferent over partners and are only interested in the transfer they receive; while women also care about some attractiveness characteristic of men, in a way that may depend on the woman's group. In a marriage between man i of group x and woman j of group y, if the wife transfers τ to the husband his net utility would be τ, and hers would be $(\varepsilon_{iy} - \tau)$. Since the joint surplus is ε_{iy}, it clearly satisfies Assumption 1. All of our results would apply in this case. Since there is a continuum of women in each group y, but only one man i, he must capture all joint surplus if he marries a woman of group y: his net utility must be ε_{iy}, and hers zero. In other words, this man will receive a transfer $\tau_i = \max_{y \in \mathcal{Y}} \varepsilon_{iy}$, which depends on his unobservable characteristic. In contrast, in Choo and Siow's preferred interpretation equilibrium transfers only depend on characteristics that are observed by the analyst. Once again, this is a matter of modelling choice and not a logical necessity since the $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ terms are observed by all agents.

C.2 The logit assumption

A second major assumption in the Choo and Siow model states that the distribution of the unobserved heterogeneity terms ε_{iy} and η_{xj} are distributed as type I extreme value iid random vectors. This brings in familiar but restrictive features of the logit model, and in particular, the Independence of Irrelevant Alternatives (IIA) property.

The literature on single-agent discrete choice models has long stressed the links between the type I-EV specification and IIA. In his famous discussion of Luce (1959), Debreu (1960) showed that given IIA, introducing irrelevant attributes would change choice probabilities. Matching markets are two-sided by their very nature, and defining IIA is less straightforward than in single-agent models—we propose two definitions and draw out their implications in Galichon and Salanié (2019). Still, it is not hard to construct illustrations similar to Debreu's example within the Choo and Siow model.

Let x and y consist of education, with two levels C (college) and N (no college). Now suppose that the analyst distinguishes two types of college graduates: those whose Commencement fell on an even-numbered day C_e and those for whom it was on an odd-numbered day C_o. Assume that this difference in fact is payoff-irrelevant: the joint surplus of any match does not depend on whether the college graduates in it (if any) had Commencement on an even day. We show in Galichon and Salanié (2019) that adding the Commencement distinction to the model changes equilibrium marriage patterns: it reduces the number of singles, and it increases the number of matches between college graduates while reducing the number of matches between non-graduates. These are clearly unappealing properties: since the Commencement date is irrelevant to all market participants, a more reasonable model would imply none of these changes.

The Choo and Siow model has other stark comparative statics predictions. Since $u_x = -\log(\mu_{x0}/n_x)$ in this framework, average utilities are in a one-to-one relationship with the probabilities of singlehood. Property (D.1) becomes a statement on semi-elasticities of these probabilities. Moreover, the equilibrium equation (3.13) implies that for any 4-tuple of characteristics (x, y, x', y'),

$$\frac{\mu_{y|x}\mu_{y'|x'}}{\mu_{y|x'}\mu_{y'|x}} = \exp((\Phi_{xy} + \Phi_{x'y'} - \Phi_{x'y} - \Phi_{xy'})/2).$$

Therefore the log-odds ratio $(\mu_{y|x}\mu_{y'|x'})/(\mu_{y|x'}\mu_{y'|x})$ should only depend on the joint surplus matrix $\mathbf{\Phi}$, and not on the availability of different types $\boldsymbol{n}, \boldsymbol{m}$. It is easy to see that none of the other specifications we study in this section has this invariance property. It is in principle testable, given data for several markets which can be assumed to have the same surplus function. This property was first pointed out by Graham (2013), who also describes other predictions of the Choo and Siow framework[20].

[20] Mourifié and Siow (2021) and Mourifié (2019) extend this and other results of Graham (2013) to models with peer effects.

D Some properties of the stable matching [online]

We now state additional results which took too much space to fit into the main text.

D.1 Symmetry

Recall from Proposition 2 that the partial derivative of the social surplus $\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{r})$ with respect to n_x is u_x. It follows immediately that

$$\frac{\partial u_x}{\partial n_{x'}} = \frac{\partial u_{x'}}{\partial n_x}. \tag{D.1}$$

Hence the "unexpected symmetry" result proven by Decker, Lieb, McCann, and Stephens (2012) for Choo and Siow model is a direct consequence of the symmetry of the Hessian of $\mathcal{W}$; and it holds for *all* separable models.

Our second corollary states some properties of the objective function $\mathcal{W}$, as a direct implication of Theorem 3.

Corollary 1. *The function $\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{n}, \boldsymbol{m})$ is convex in $\boldsymbol{\Phi}$. It is homogeneous of degree 1 and concave in $\boldsymbol{r} = (\boldsymbol{n}, \boldsymbol{m})$.*

Proof. The convexity of $\mathcal{W}$ w.r.t. Φ follows immediately from (3.5); the concavity of $\mathcal{W}$ w.r.t. $(\boldsymbol{r})$ similarly follows from (3.6). Since $G(\boldsymbol{U}, \boldsymbol{n})$ is 1-homogeneous in $\boldsymbol{n}$ and $H(\boldsymbol{V}, \boldsymbol{m})$ is 1-homogeneous in $\boldsymbol{m}$, the dual program shows that $\mathcal{W}$ is 1-homogeneous in $\boldsymbol{r} = (\boldsymbol{n}, \boldsymbol{m})$.

Corollary 1 entails further consequences. Since the function $\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{r})$ is concave in $\boldsymbol{r}$, the matrix $\partial^2 \mathcal{W}/\partial \boldsymbol{r} \partial \boldsymbol{r'}$ must be semidefinite negative. This implies the symmetry result above, and much more—including sign constraints on the minors[21]. Similarly, since $\mathcal{W}$ is convex in $\boldsymbol{\Phi}$ the matrix of general term $\partial^2 \mathcal{W}/\partial \Phi_{xy} \partial \Phi_{zt}$ must be semi-definite positive, which implies

[21] The most obvious one implies that the expected utility of a type must decrease with the mass of its members:

$$\frac{\partial u_x}{\partial n_x} = \frac{\partial^2 \mathcal{W}}{\partial n_x^2} \leq 0.$$

certain symmetry and determinant sign constraints. Galichon and Salanié (2017) studies the comparative statics of separable models in more detail.

Finally, the homogeneity of $\mathcal{W}$ in $\boldsymbol{r}$ implies that all utilities (e.g. U_{xy} and v_t) and all conditional matching probabilities $\mu_{y|x}$ must be homogeneous of degree 0 in $\boldsymbol{r}$. In that sense, all separable models exhibit constant returns to scale. This property distinguishes separable models from those in Dagsvik (2000) or Menzel (2015). It can be viewed either as a feature or as a bug. Mourifié and Siow (2021) and Mourifié (2019) argue for a class of "Cobb-Douglas marriage matching functions" that extends the multinomial logit specification of Choo and Siow (2006) beyond separable models and allows for scale and peer effects.

D.2 Other comparative statics results

Theorem 3 can be used to show that other comparative statics results of Decker, Lieb, McCann, and Stephens (2012) extend beyond the logit model to our generalized framework, beyond those stated in Subsection D.1. Many of these results are collected in Galichon and Salanié (2017), but we recall some here for completeness. From the results of Section 3.1, recall that $\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{r})$ is given by the dual expressions

$$\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{r}) = \max_{\mu \in \mathcal{M}(\boldsymbol{r})} \left(\sum_{xy} \mu_{xy} \Phi_{xy} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{r}) \right), \text{ and} \tag{D.2}$$

$$\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{r}) = \min_{U_{xy}+V_{xy}=\Phi_{xy}} \left(\sum n_x G_x(U_{xy}) + \sum m_y H_y(V_{xy}) \right); \tag{D.3}$$

and that

$$\frac{\partial \mathcal{W}}{\partial \Phi_{xy}} = \mu_{xy}, \; \frac{\partial \mathcal{W}}{\partial n_x} = G_x(U_{xy}) = u_x, \text{ and } \frac{\partial \mathcal{W}}{\partial m_y} = H_y(V_{xy}) = v_y.$$

By the same logic as the one that obtained (D.1), the cross-derivative of $\mathcal{W}$ with respect to $n_{x'}$ and Φ_{xy} yields

$$\frac{\partial \mu_{xy}}{\partial n_{x'}} = \frac{\partial^2 \mathcal{W}}{\partial n_{x'} \partial \Phi_{xy}} = \frac{\partial u_{x'}}{\partial \Phi_{xy}} \tag{D.4}$$

which is proven (again in the case of the multinomial logit Choo and Siow model) in Decker, Lieb, McCann, and Stephens (2012, section 3). The effect of an increase in the matching surplus between groups x and y on the surplus of individual of group x' equals the effect of the mass of individuals of group x' on the mass of matches between groups x and y. Let us provide an interpretation for this result. Assume that groups x and y are men and women with a PhD, and that x' are men with a college degree. Suppose that $\partial \mu_{xy}/\partial n_{x'} < 0$, so that an increase in the mass of men with a college degree causes the mass of matches between men and women with a PhD to decrease. This suggests that men with a college degree or with a PhD are substitutes for women with a PhD. Hence, if there is an increase in the matching surplus between men and women with a PhD, men with a college degree will become less of a substitute for men with a PhD. Therefore their share of surplus will decrease, and $\partial u_{x'}/\partial \Phi_{xy} < 0$.

Finally, differentiating $\mathcal{W}$ twice with respect to Φ_{xy} and $\Phi_{x'y'}$ yields

$$\frac{\partial \mu_{xy}}{\partial \Phi_{x'y'}} = \frac{\partial^2 \mathcal{W}}{\partial \Phi_{xy} \partial \Phi_{x'y'}} = \frac{\partial \mu_{x'y'}}{\partial \Phi_{xy}}. \tag{D.5}$$

The interpretation is the following: if increasing the matching surplus between groups x and y has a positive effect on marriages between groups x' and y', then increasing the matching surplus between groups x' and y' has a positive (and equal) effect on marriages between groups x and y. Again, all comparative statics results derived in this section hold in *any* model that satisfies our assumptions.

E Additional results on estimation [online]

E.1 Moment matching

Assume that the specification of the joint surplus $\boldsymbol{\Phi}^{\boldsymbol{\lambda}}$ is linear in $\boldsymbol{\lambda}$ and that the distributions of the unobserved heterogeneity terms $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ are known. Let (ϕ^k_{xy}) be the basis functions, and define the *comoments* $C^k(\mu) = \sum_{xy} \mu_{xy} \phi^k_{xy}$ for any matching μ. This appendix proves the following result:

Theorem 6 (Comoments and a specification test). *Denote* $\hat{\boldsymbol{\lambda}}^{MM}$ *the moment-matching* *estimator defined by* (5.4).

1. *It makes predicted comoments equal to observed comoments:* $C^k(\hat{\boldsymbol{\mu}}) = C^k(\boldsymbol{\mu}^{\boldsymbol{\lambda}})$ *for all* k *when* $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^{MM}$.

2. *It is also the vector of Lagrange multipliers of the comoment equality constraints in the program*

$$\mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right) = \max_{\boldsymbol{\mu} \in \mathcal{M}(\hat{\boldsymbol{r}})} \left(\mathcal{E}\left(\boldsymbol{\mu}, \hat{\boldsymbol{r}}\right) : C^k(\boldsymbol{\mu}) = C^k(\hat{\boldsymbol{\mu}})\ \forall k\right). \tag{E.1}$$

3. *The value of* $\mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right)$ *is* $\mathcal{E}\left(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}^{MM}}, \hat{\boldsymbol{r}}\right)$. *Moreover,* $\mathcal{E}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right) \leq \mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right)$, *with equality if and only if there is a value* $\boldsymbol{\lambda}$ *of the parameter such that* $\boldsymbol{\Phi}^{\boldsymbol{\lambda}} = \boldsymbol{\Phi}$.

Proof. We denote $\hat{\boldsymbol{\lambda}} := \hat{\boldsymbol{\lambda}}^{MM}$ to simplify the notation.

1. By definition, $\sum_{x,y} \hat{\mu}_{xy} \phi^k_{xy} = \left(\partial \mathcal{W} / \partial \lambda_k\right)\left(\boldsymbol{\Phi}^{\hat{\boldsymbol{\lambda}}}, \hat{\boldsymbol{r}}\right)$. Applying the envelope theorem to (3.5) shows that
$$\frac{\partial \mathcal{W}}{\partial \lambda_k}(\boldsymbol{\Phi}^{\hat{\boldsymbol{\lambda}}}, \hat{\boldsymbol{r}}) = \sum_{x,y} \mu^{\hat{\boldsymbol{\lambda}}}_{xy} \phi^k_{xy}.$$
Therefore $\sum_{xy} \mu^{\hat{\boldsymbol{\lambda}}}_{xy} \phi^k_{xy} = \sum_{xy} \hat{\mu}_{xy} \phi^k_{xy}$.

2. Given (3.5), the program (5.4) can be rewritten as
$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^K} \min_{\boldsymbol{\mu} \in \mathcal{M}(\hat{\boldsymbol{r}})} \left(\sum_k \lambda_k \sum_{x,y} \left(\hat{\mu}_{xy} - \mu_{xy}\right) \phi^k_{xy} - \mathcal{E}\left(\boldsymbol{\mu}, \hat{\boldsymbol{r}}\right)\right).$$
Since the objective function is convex in $\boldsymbol{\mu}$ and linear in $\boldsymbol{\lambda}$, we can exchange the max and the min. Consider a value of $\boldsymbol{\mu}$ such that $\sum_{x,y} \left(\hat{\mu}_{xy} - \mu_{xy}\right) \phi^k_{xy} \neq 0$ for some k; then minimizing over $\boldsymbol{\lambda}$ gives $-\infty$. Therefore these K equalities must hold at the optimum, and $\boldsymbol{\mu}$ minimizes $\mathcal{E}$ over the set of $\boldsymbol{\mu} \in \mathcal{M}(\hat{\boldsymbol{r}})$ such that $\sum_{x,y} \left(\hat{\mu}_{xy} - \mu_{xy}\right) \phi^k_{xy} = 0$ for all k.

3. Since $\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}}$ maximizes $\sum_{x,y}\mu_{xy}\Phi^{\hat{\boldsymbol{\lambda}}}_{xy}+\mathcal{E}(\boldsymbol{\mu};\hat{\boldsymbol{r}})$ over $\boldsymbol{\mu}$,

$$\mathcal{E}\left(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}},\hat{\boldsymbol{r}}\right)-\mathcal{E}\left(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{r}}\right)\geq\sum_{x,y}\left(\hat{\mu}_{xy}-\mu^{\hat{\boldsymbol{\lambda}}}_{xy}\right)\Phi^{\hat{\boldsymbol{\lambda}}}_{xy}$$

and the inequality is strict unless $\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}}=\hat{\boldsymbol{\mu}}$, since $\mathcal{E}$ is strictly concave in $\boldsymbol{\mu}$. By part 1, the RHS is zero. Therefore $\mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{r}}\right)=\mathcal{E}(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}},\hat{\boldsymbol{r}})\geq\mathcal{E}\left(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{r}}\right)$, with equality if and only if $\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}}=\hat{\boldsymbol{\mu}}$.

If $\boldsymbol{\Phi}=\boldsymbol{\Phi}^{\boldsymbol{\lambda}}$, then $\hat{\boldsymbol{\mu}}$ maximizes $\sum_{x,y}\mu_{xy}\Phi^{\boldsymbol{\lambda}}_{xy}+\mathcal{E}(\boldsymbol{\mu},\hat{\boldsymbol{r}})$, and $\boldsymbol{\mu}^{\boldsymbol{\lambda}}=\hat{\boldsymbol{\mu}}$. Therefore $\mathcal{E}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{r}})=\mathcal{E}_{\max}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{r}})$.

Comparing the values of $\mathcal{E}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{r}})$ and $\mathcal{E}\left(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}^{MM}},\hat{\boldsymbol{r}}\right)$ gives a simple specification test. Its critical values can be obtained by bootstrapping for instance. One could also run the test for different specifications of the distributions of heterogeneities and invert it to obtain confidence intervals for the parameters of $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$.

E.2 Geometric interpretation of the estimation procedure

The approach to inference we describe in Section 5.2 has a simple geometric interpretation. In this appendix, we fix the distributions $\boldsymbol{P}_x$ and a specification $(\phi^k_{xy})_{k=1,\dots,K}$ of the linear model of surplus $\boldsymbol{Q}_y$; and we vary the parameter vector $\boldsymbol{\lambda}$. Now consider the set of moments associated to all feasible matchings:

$$\mathcal{F}=\left\{\left(C^1,\dots,C^K\right):C^k=\sum_{xy}\mu_{xy}\phi^k_{xy},\ \boldsymbol{\mu}\in\mathcal{M}\left(\hat{\boldsymbol{r}}\right)\right\}$$

This is a convex polyhedron, which we call the *covariogram*. It includes the observed commoments $\hat{\boldsymbol{C}}$, as well as the vector of moments $C^{\boldsymbol{\lambda}}$ generated by the optimal matching $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ for any value of the parameter vector $\boldsymbol{\lambda}$. Each feasible matching $\boldsymbol{\mu}$ also has a generalized entropy $\mathcal{E}(\boldsymbol{\mu},\hat{\boldsymbol{r}})$; we denote $\mathcal{E}^{\boldsymbol{\lambda}}\equiv\mathcal{E}(\boldsymbol{\mu}^{\boldsymbol{\lambda}},\hat{\boldsymbol{r}})$ the generalized entropy associated with parameter vector $\boldsymbol{\lambda}$. Since the vectors ϕ are linearly independent, the mapping $\boldsymbol{\lambda}\longrightarrow C^{\boldsymbol{\lambda}}$ is invertible on the covariogram. Denote $\boldsymbol{\lambda}(C)$ its inverse. The corresponding optimal matching has

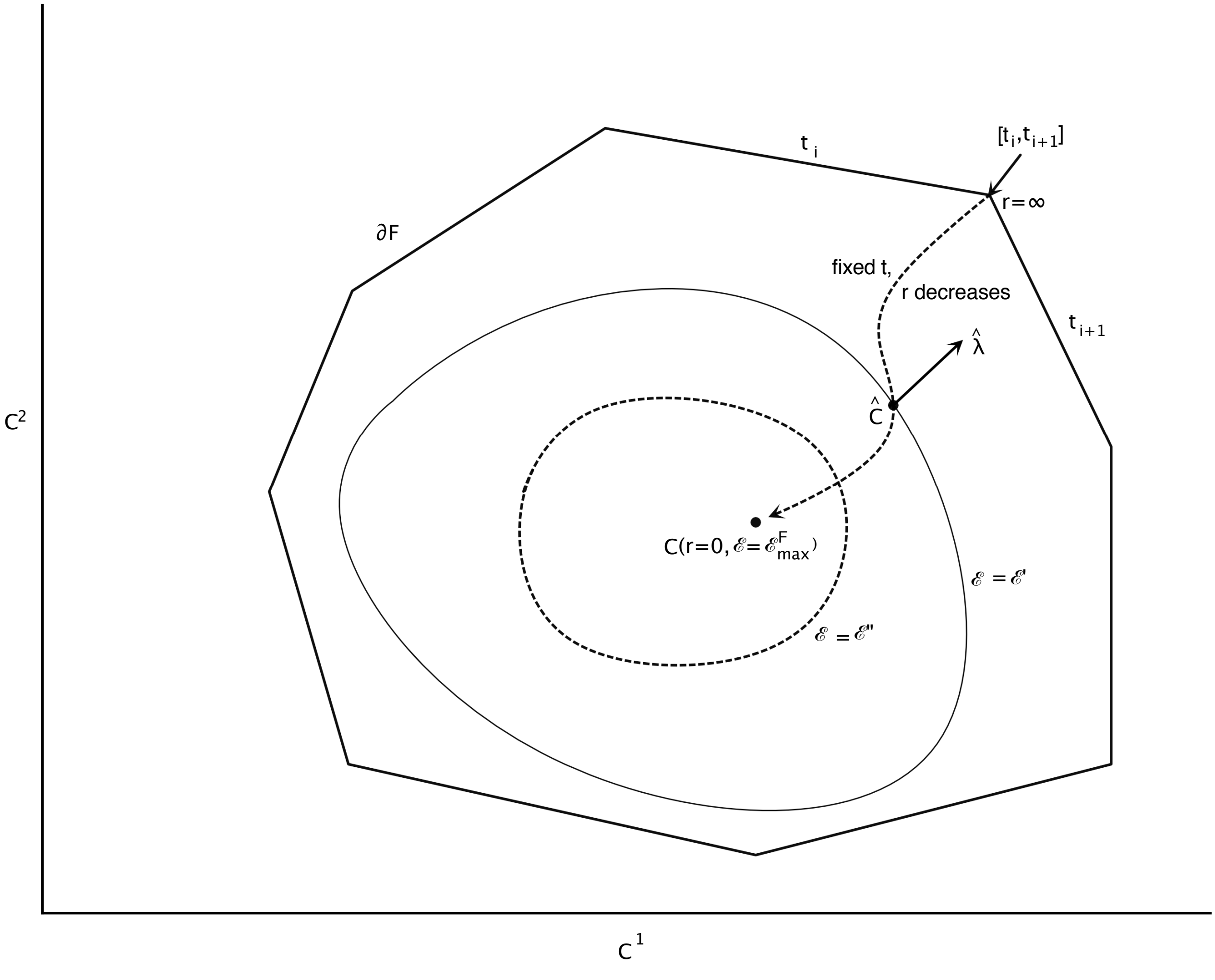


Figure 4: The covariogram and related objects

generalized entropy $\mathcal{E}_r\left(C\right) = \mathcal{E}^{\boldsymbol{\lambda}(C)}$. The level sets of the function $\mathcal{E}_r$ are *isoentropy surfaces* in the covariogram.

Figure 4 illustrates these concepts. It assumes $K = 2$ basis functions, so that the covariogram is a convex polyhedron in (C^1, C^2) plane. Since $\boldsymbol{\lambda}$ also is two-dimensional, it can be represented in polar coordinates. Let the data be generated by $\boldsymbol{\lambda} = r\exp(it)$. For $r = 0$, the model is uninformative: matching is random and generalized entropy takes its maximum possible value $\mathcal{E}^F_{\max}$ among all possible matchings. We denote C_0 the corresponding moments. At the other extreme, the boundary ∂F of the covariogram corresponds to $r = \infty$. There is no unobserved heterogeneity; generically over t, the moments generated by $\boldsymbol{\lambda}$ must belong to a finite set of vertices, so that $\boldsymbol{\lambda}$ is only set-identified.

As r decreases for a given t, the corresponding moments follow a trajectory indicated by the dashed line on Figure 4, from the boundary ∂F to the point C_0. The entropy $\mathcal{E}^{\boldsymbol{\lambda}}$ increases as this trajectory crosses contours of higher entropy ($\mathcal{E}'$ then $\mathcal{E}''$ on the figure.)

We know from Theorem 6.2 that the moment-matching estimator $\hat{\boldsymbol{\lambda}}^{MM}$ is the vector of multipliers of the program that maximizes entropy over the matchings that generate the observed values of the moments. Therefore $\partial\mathcal{E}_r(\hat{C})/\partial C^k = \hat{\lambda}^{MM}_k$; and the moment-matching estimator lies on the normal to the isoentropy contour that goes through the observed moments $\hat{\boldsymbol{C}}$. This is shown as $\hat{\lambda}$ on Figure 4.

E.3 Parameterization, testing, and multimarket data

Proposition 4 shows that, given a specification of the distribution of the unobserved heterogeneities $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$, there is a one-to-one correspondence between $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$. To put it differently: any matching on a single market can be rationalized by exactly one model that satisfies Assumptions 1 and 2, for any such vector of distributions. This has several consequences for analysts using data on a single market. Without further restrictions, it is impossible to test separability, even assuming perfect knowledge of the distributions of unobserved heterogeneity. It is also impossible to discriminate between separable models based on different distributions. One way out of this conundrum is to incorporate credible restrictions (inspired by theoretical restrictions, or by institutional features of the market) into both the surplus matrix $\mathbf{\Phi}$ and the distributions of unobservable heterogeneity $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$. To take a simple example, suppose that we know that there is no interaction between partner characteristics x^k and y^l in the production of joint surplus: for fixed values of the other characteristics, Φ_{xy} is additive in x^k and y^l. Given our identification formula (3.9) and observed matching patterns, this translates into a set of constraints on the derivatives of the generalized entropy, and therefore on the distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$. Adding constraints on the distributions would make the model testable[22]. As another example, consider a semiparametric specification in the spirit of Ekeland, Heckman, and Nesheim (2004): $\Phi_{xy} = \boldsymbol{b}_y' \boldsymbol{\phi}_x$, with known d-dimensional vectors $\boldsymbol{\phi}_x$ and unknown vectors $\boldsymbol{b}_y$. If $d < |Y|$, this would restrict the number of degrees of freedom in $\mathbf{\Phi}$, freeing parameters to specify the distributions of heterogeneity and/or to test the model. An alternative empirical strategy is to use multiple markets with restricted parametric variation in the joint surplus $\mathbf{\Phi}$ and the distributions of unobserved heterogeneity $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$. The variations in the group sizes $\boldsymbol{n}$ and $\boldsymbol{m}$ across markets then generate variation in optimal matchings that can be used to overidentify the model and generate testable restrictions. Chiappori, Salanié, and Weiss (2017) relied on a variant of this approach.

F Computational Methods and Benchmarks [online]

Section 4 described two classes of methods to compute the equilibrium matching patterns: min-Emax, and IPFP. Min-Emax is more generally applicable than IPFP; on the other hand, IPFP is much faster. To document these claims, we present in this appendix a small simulation of the Choo and Siow model that explores the computational performance of four different methods: a general-purpose equation solver, the min-Emax method, the minimization of the function F expressed in (3.14), and IPFP.

In the second part of this appendix, we show how linear programming techniques can

[22] As a trivial illustration, finding that $\log \hat{\mu}_{xy}$ is not additive in x^k and y^l would reject the Choo and Siow model in this example.

be used to solve and estimate a model with discretized error distributions.

F.1 Benchmarks

We simulated ten cases, with a number of categories $|\mathcal{X}| = |\mathcal{Y}|$ that goes from 100 to 5,000. For each of these ten cases, we draw the n_x and m_y uniformly in $\{1, \dots, 100\}$; and for each (x, y) match we draw $\Phi_{xy}/2$ from $\mathcal{N}(0, 1)$.

F.1.1 Minpack

We applied the Levenberg-Marquardt solver Minpack to the system of $(\mathcal{X}| + |\mathcal{Y})$ equations that characterizes the optimal matching (see Section 4). Minpack is probably the most-used solver in scientific applications; it underlies many statistical and numerical packages.

F.1.2 Min-Emax

The min-Emax method we described in Section 4 minimizes $(G(\boldsymbol{U}, \boldsymbol{n}) + H(\boldsymbol{\Phi} - \boldsymbol{U}, \boldsymbol{m}))$ over the $|\mathcal{X}| \times |\mathcal{Y}|$ object $\boldsymbol{U} = (U_{xy})$. In the particular case of the Choo and Siow model, the function G is given by

$$G(\boldsymbol{U}, \boldsymbol{n}) = \sum_{x \in \mathcal{X}} n_x \log \left(1 + \sum_{y \in \mathcal{Y}} \exp(U_{xy}) \right)$$

and H has the same form.

We used the Knitro optimizer[23] to obtain the solution.

F.1.3 Minimizing F

Formula (3.14) provides us with an alternative method that works on the smaller object (u_x, v_y) of group average utilities. Here again we used the Knitro optimizer.

[23] See Byrd, Nocedal, and Waltz (2006).

F.1.4 IPFP

Consider the logit model of Choo and Siow.

Fix a value of $\boldsymbol{\lambda}$ and drop it from the notation: let the joint surplus function be $\boldsymbol{\Phi}$, with optimal matching $\boldsymbol{\mu}$. Formula (3.12) can be rewritten as

$$\mu_{xy} = \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}\mu_{0y}}. \tag{F.1}$$

As noted by Decker, Lieb, McCann, and Stephens (2012) we could just plug this into the feasibility constraints $\sum_y \mu_{xy} + \mu_{x0} = n_x$ and $\sum_x \mu_{xy} + \mu_{0y} = m_y$ and solve for the masses of singles μ_{x0} and μ_{0y}. This results in a system of $|\mathcal{X}| + |\mathcal{Y}|$ equations:

$$\mu_{x0} + \left(\sum_{y\in\mathcal{Y}} \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{0y}}\right)\sqrt{\mu_{x0}} = n_x \tag{F.2}$$

$$\mu_{0y} + \left(\sum_{x\in\mathcal{X}} \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}}\right)\sqrt{\mu_{0y}} = m_y. \tag{F.3}$$

Taking the unknowns to be $\sqrt{\mu_{x0}}$ and $\sqrt{\mu_{0y}}$, each of these equations is quadratic in the unknowns. IPFP simply consists of solving the system (F.2) iteratively. Starting from an arbitrary guess $\mu_{0y}^{(0)}$, at step $(2k+1)$ we find the following updating equation (4.2). These equations are easily solved in closed form. The pseudo-code in Algorithm 1 gives a detailed implementation. Note that since in the Choo and Siow model the shadow prices u_x and v_y are simply minus the logarithms of the corresponding μ_{x0} and μ_{0y}, this algorithm in fact operates on u_x and v_y.

Algorithm 1. ***Solving for the optimal matching by IPFP***

Require: *two non-negative vectors* $\boldsymbol{n}$ *and* $\boldsymbol{m}$ *(sizes* M *and* N*); a matrix* $\boldsymbol{\Phi}$ *of size* (M, N)

Require: *a tolerance* τ *and a maximum number of iterations* I

Ensure: *the matrix* $\boldsymbol{\mu}$ *of size* (M, N) *holds the marriage patterns at the optimal matching*

```
  $X \leftarrow size(\boldsymbol{n})$
  $Y \leftarrow size(\boldsymbol{m})$
```

```
$\boldsymbol{K} \leftarrow \exp(\boldsymbol{\Phi}/2)$
$\delta \leftarrow \infty, i \leftarrow 0$
$\boldsymbol{T} \leftarrow 0_Y$
**while** $\delta > \tau$ *and* $i < I$ **do**
    $\boldsymbol{S} \leftarrow \boldsymbol{K}\boldsymbol{T}$                                  ▷ Project on $\boldsymbol{n}$ margins
    $\boldsymbol{t} \leftarrow \left(\sqrt{\boldsymbol{S}^2 + 4\boldsymbol{n}} - \boldsymbol{S}\right)/2$
    $S \leftarrow \boldsymbol{K}'\boldsymbol{t}$                                  ▷ Project on $\boldsymbol{m}$ margins
    $\boldsymbol{T} \leftarrow \left(\sqrt{\boldsymbol{S}^2 + 4\boldsymbol{m}} - \boldsymbol{S}\right)/2$
    $\delta_1 \leftarrow \max|\boldsymbol{t}^2 + \boldsymbol{t} \odot \boldsymbol{K}\boldsymbol{T} - \boldsymbol{n}|$          ▷ Error on $\boldsymbol{n}$ margins
    $\delta_2 \leftarrow \max|\boldsymbol{T}^2 + \boldsymbol{T} \odot \boldsymbol{K}'\boldsymbol{t} - \boldsymbol{m}|$          ▷ Error on $\boldsymbol{m}$ margins
    $\delta \leftarrow \max(\delta_1, \delta_2)$
    $i \Leftarrow i + 1$
**end while**
**if** $i \geq I$ **then**
    *Failed to achieve requested precision*
**else**
    $\boldsymbol{\mu} \leftarrow \boldsymbol{K} \odot (\boldsymbol{t} \otimes \boldsymbol{T})$          ▷ $\otimes$ denotes outer product and $\odot$ element-wise product
**end if**
```

F.1.5 Results

For all four methods, we used C/C++ programs run on a single processor of a Mac desktop. We set the convergence criterion for all methods as a relative estimated error of 10^{-6}. This is not as straightforward as one would like: both Knitro and Minpack rescale the problem before solving it, while we did not attempt to do it for IPFP. Still, varying the tolerance within reasonable bounds hardly changes the results, which we present in Figure 5. Each panel gives the distribution of CPU times for one of the four methods, in the form of a Tukey box-and-whiskers graph[24].

[24] The box goes from the first to the third quartile; the horizontal bar is at the median; the lower (resp. upper) whisker is at the first (resp. third) quartile minus (resp. plus) 1.5 times the interquartile range, and

There are three things to note about these graphs. First, distances on the x-axis are not drawn to scale, except for the smaller number of categories; second the y-axis is logarithmic; third, for some methods we only report results on the lower range of categories. The reasons are obvious from the graphs. Minpack solving not scale up well. The min-Emax method that minimizes $(G(\boldsymbol{U})+H(\boldsymbol{\Phi}-\boldsymbol{U}))$ is even worse: in this "logit" case, it is not competitive beyond 100 categories as it minimizes in a high-dimensional space. On the other hand, the min-Emax method that optimizes over $\boldsymbol{u}$ and $\boldsymbol{v}$ and the IPFP algorithm both perform remarkably well, even with several thousands of categories.

Choo and Siow only used 60 categories in their application (ages from 16 to 75). For such numbers, all four methods work well, but IPFP and min-Emax on (u, v) again clearly dominate. We should emphasize that only the special structure of the Choo and Siow model allowed us to reduce the dimensionality by minimizing over $\boldsymbol{u}$ and $\boldsymbol{v}$. IPFP, on the other hand, can be used in a broader class of models. While IPFP has more variability than the other methods (perhaps because we did not rescale the problem beforehand), even the slowest convergence times for each problem size are at least three times smaller than those of Minpack. This is all the more remarkable that IPFP does not require any calculation of derivatives; by comparison, we fed the code for the Jacobian of the system of equations into Minpack. IPFP also compares very well with the min-Emax method on (u, v), even though we fed the Jacobian and the Hessian into Knitro.

Finally, while we have run these experiments on a single processor, it is clear that IPFP is much more amenable to parallel implementation than the optimization methods, since each iteration solves $|\mathcal{X}|$ or $|\mathcal{Y}|$ equations that are independent of each other.

F.2 An additional method: linear programming

Min-Emax and IPFP both exploit the structure of the separable matching problem. A more "brute-force" alternative is to simply solve the underlying linear programming problem. This requires discretizing the distribution of the error terms. We now explain how it can

the circles plot all points beyond that.

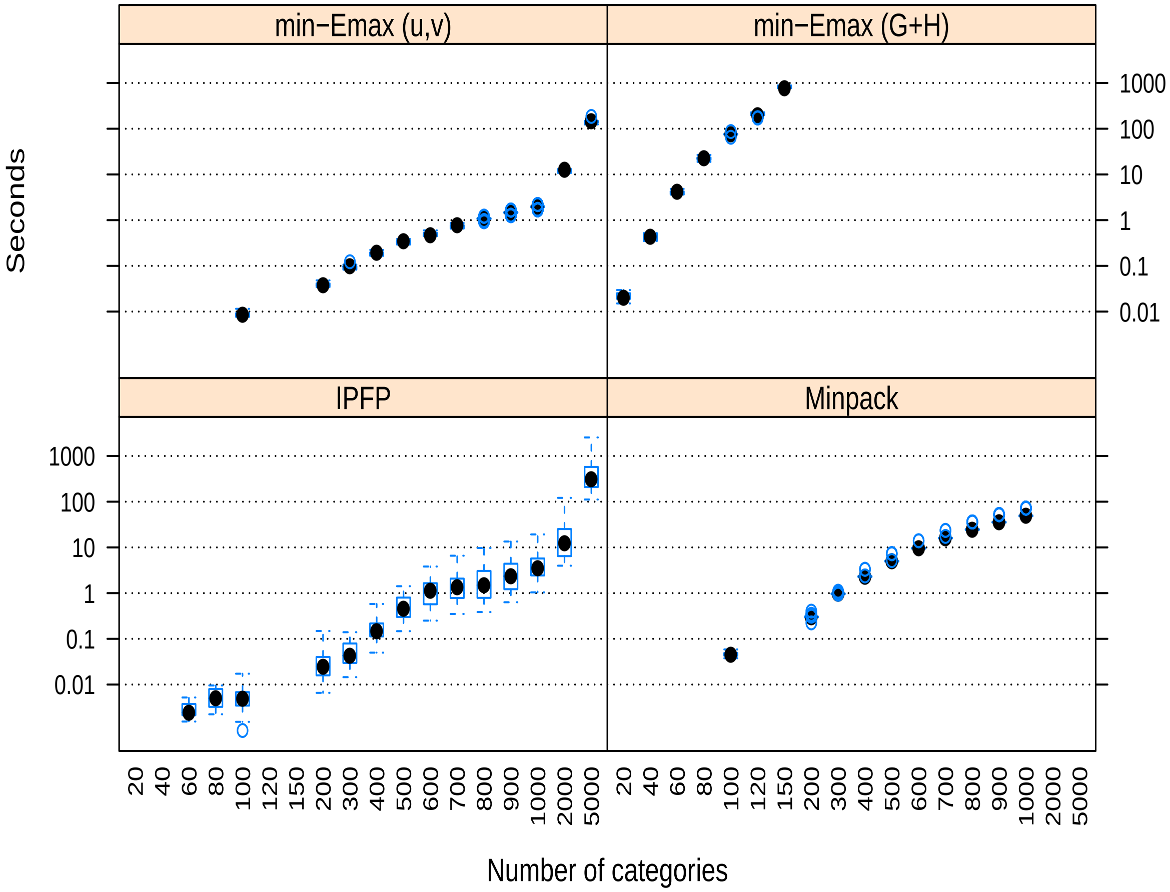


Figure 5: Solving for the optimal matching

be done, and we extend it to obtain the moment-based estimator in a semilinear model.

F.2.1 Equilibrium via linear programming

Now suppose that the vectors ε and η, instead of having full support, only take a finite number of values: these are analogous to the unobserved "types" of many structural econometric models. We define $(\varepsilon_y^{xk})_{y\in\mathcal{Y}_0,k=1,\ldots,K_x}$ to be the points of support of $\mathbf{P}_x$, and (r_x^k) their probabilities; and we define (η_x^{yl}) and (s_y^l) similarly. In this case, $G(\boldsymbol{U},\boldsymbol{n})$ is $\sum_x n_x E_{\mathbf{P}_x} \max_y(U_{xy}+\varepsilon_y)$, that is

$$G(\boldsymbol{U},\boldsymbol{n}) = \sum_{x\in\mathcal{X}} n_x \sum_{k=1}^{K_x} r_x^k \max\left(\varepsilon_0^{xk}, \max_{y\in\mathcal{Y}}(U_{xy}+\varepsilon_y^{xk})\right).$$

Define $u_x^k = \max\left(\varepsilon_0^{xk}, \max_{y\in\mathcal{Y}}(U_{xy}+\varepsilon_y^{xk})\right)$, and $v_y^l = \max\left(\eta_0^{yl}, \max_{x\in\mathcal{X}}(V_{xy}+\eta_x^{yl})\right)$. By construction,

$$u_x^k \geq U_{xy}+\varepsilon_y^{xk} \;\forall y \quad\text{and}\quad u_x^k \geq \varepsilon_0^{xk} \tag{F.4}$$

$$v_y^l \geq V_{xy}+\eta_x^{yl} \;\forall x \quad\text{and}\quad v_y^l \geq \eta_0^{yl}. \tag{F.5}$$

It follows from (3.6) that we minimize the objective function and given the constraint $U_{xy}+V_{xy}\geq\Phi_{xy}$, it is easy to see that the optimal matching solves

$$\mathcal{W}(\boldsymbol{\Phi},\boldsymbol{n},\boldsymbol{m}) = \min_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{U}} \left(\sum_{x\in\mathcal{X}} n_x \sum_{k=1}^{K_x} r_x^k u_x^k + \sum_{y\in\mathcal{Y}} m_y \sum_{l=1}^{L_y} s_y^l v_y^l\right)$$

subject to the constraints (F.4) and (F.5) with $V_{xy}=\Phi_{xy}-U_{xy}$. Note that the objective function and the constraints are linear in the variables. Therefore solving for equilibrium with finite types boils down to a linear programming problem, for which very fast algorithms are available (even with many variables). The multipliers of the constraints at the optimum give the matching patterns for each type in each group, and can be averaged over types to yield the μ_{xy}. This idea can be taken further: any distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$ can be discretized. Solving the program above for a given finite-support approximation of the distributions gives an approximation that can be shown to converge to the optimum for the limit of the discrete distributions, by adapting an argument of Chernozhukov, Galichon, Hallin, and Henry (2017, Theorem 3.1). Hence the approach described in this subsection is applicable to any separable model.

F.2.2 Computing the moment-matching estimator

The linear programming approach of Subsection F.2.1 can be extended in order to compute the moment-matching estimator in the semilinear models of Section 5.2. Equation (5.4) shows that the moment-matching estimator minimizes $\min_{\boldsymbol{\lambda}} \left(\mathcal{W}(\boldsymbol{\lambda}'\tilde{\boldsymbol{\phi}}, \hat{\boldsymbol{r}}) - \boldsymbol{\lambda}'\hat{C} \right)$ over λ. This suggests a general approach to the estimation of separable models with known distributions of heterogeneity. First, specify a linear surplus function and distributions of unobservable heterogeneity $\mathcal{P}_x$ and $\mathcal{Q}_y$. Second, discretize the latter distributions. Third, solve the following linear program:

$$\min_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{U},\boldsymbol{\lambda}} \left(\sum_{x\in\mathcal{X}} \hat{n}_x \sum_{k=1}^{K_x} r_x^k u_x^k + \sum_{y\in\mathcal{Y}} \hat{m}_y \sum_{l=1}^{L_y} s_y^l v_y^l - \boldsymbol{\lambda}'\hat{C} \right)$$

under the constraints (F.4) and (F.5), replacing V_{xy} with $\boldsymbol{\lambda}'\tilde{\boldsymbol{\phi}}_{\boldsymbol{xy}} - U_{xy}$. The objective and the constraints are still linear with respect to all variables, which now also include $\boldsymbol{\lambda}$. Once again, this program can be solved efficiently by linear programming algorithms, yielding both the moment-matching estimator and the corresponding average utilities and matching patterns.

A summary

Each computational method has pros and cons. The min-Emax method can be applied quite generally. It requires many evaluations of G and H however, which may be costly for large $|\mathcal{X}|, |\mathcal{Y}|$. Linear programming is attractive in semilinear models, at the price of discretization. IPFP requires no discretization, provides easy estimation of linear model, and is highly scalable. It does require evaluating the extended entropy E of Section A.7, which is straightforward in logit-type models.

G Application to Choo and Siow's data [online]

Our empirical application uses the data Choo and Siow (2006) put together, with some minor changes. We also put more emphasis on the treatment of those (x, y) cells that have zero observations.

G.1 The data

Choo and Siow used data from the Census to evaluate the numbers $\boldsymbol{n}$ and $\boldsymbol{m}$ of men and women of every age in every state; and they relied on National Center for Health Statistics data to estimate the number of marriages by state and by age cell. They were kind enough to share with us their samples and programs; the description that follows is very similar to that in their paper, and in fact quotes freely from it.

G.1.1 The populations

Data on the populations of men and women of every age and state were extracted from the Integrated Public-Use Microdata Sample files of the U.S. Census (see Ruggles, Genadek, Goeken, Grover, and Sobek, 2015). Choo and Siow used data from the 1970 and 1980 U.S. Census to construct population vectors:

> The samples used were the 5 percent state samples for 1980 and the 1 percent Form 1 and Form 2 samples for 1970. The 1970 data sets were appropriately scaled to be comparable with the 1980 files[25].

[25] State of residence in the 1970 census files can be identified only in the state samples (Form 1 and Form 2 samples, both of which are 1 percent samples). This is the reason that the other samples were not used for 1970 calculations. Further, the age of marriage variable is available only in Form 1 samples in 1970, which meant that only one sample, the Form 1 state sample, was used for calculations involving married couples in the 1970 Census.

> [. . .]
>
> We use the `marst` variable in the census to identify a person as either never married, currently married, or previously married (divorced or widowed). To calculate the number of available individuals, we simply add the never marrieds and previously married.

Choo and Siow kept all individuals aged 16 to 75. Since the number of first marriages in which either partner is older than 40 is rather small in the 70s and 80s, we decided to focus on the populations aged 16 to 40 instead. The "state" of an individal is defined as his/her place of residence.

G.1.2 The marriages

Choo and Siow obtained data on marriages from the Vital Statistics reports that many states send to the National Center for Health Statistics (NCHS):

> Marriage records from the 1971/72 and 1981/82 Vital Statistics were used to construct the bivariate distributions of marriages. A state has to report the number of marriages to the National Center for Health Statistics to be in the sample.

We deviated from their paper in two respects.

- To be consistent with our age window for populations in the basis year we only keep marriages in which either partner is at most 41 (in the Census year+1) or 42 (in the Census year+2). We corrected a small mistake in the construction of the data—Choo and Siow (2006) did not update the ages of the subjects between Census year+1 and Census year+2. This does not affect their main conclusions.
- The list of states we include in our application is slightly different. They excluded Iowa, Minnesota, and South Carolina which we do use since they reporteed to the

NCHS in both waves. Colorado only reported to the NCHS after 1980. Choo and Siow excluded it from their study; we keep it in the 1980s wave. Choo and Siow also excluded New York City from New York State. We eventually decided to exclude both.

A "reform" state is one in which the Roe v. Wade Supreme Court decision affected the legal status of abortion. Our list of reform states comprises Alaska, California, Delaware, Florida, Georgia, Hawaii, Kansas, Maryland, and (in the 1980s only) Colorado. Our non-reform states are Alabama, Connecticut, Idaho, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Massachusetts, Michigan, Mississippi, Missouri, Montana, Nebraska, New Hampshire, New Jersey, Ohio, Pennsylvania, Rhode Island, South Dakota, Tennessee, Utah, Vermont, West Virginia, Wyoming, and the District of Columbia. We exclude from our study Arizona, Arkansas, Nevada, New Mexico, New York, North Dakota, Oklahoma, Texas, Washington, and (in the 1970s) Colorado.

G.1.3 Merging availables and marriages

Table 1 describes our data on the populations of men and women. The numbers between parenthesis refer to the population, the other numbers to the sample. With a total of 2.19m observations representing 58.67m individuals, our universe of men and women is about 40% smaller than Choo and Siow's. This is a direct consequence of our focus on younger ages. The reform states have 34.6% of the population in 1970 and 37.9% in 1980. The sample is much larger in 1980, as the ACS dataset we use had a better sampling rate then.

Census		1970	1980	Population increase
Reform states	Men	81,260 (4.32m)	351,231 (7.20m)	66.7%
	Women	66,920 (3.63m)	308,808 (6.37m)	76.2%
Non-reform states	Men	150,887 (7.82m)	566,460 (11.51m)	47.2%
	Women	137,839 (7.16m)	524,741 (10.68m)	49.2%
Total	Men	232,147 (12.14m)	917,691 (18.71m)	54.2%
	Women	204,759 (10.77m)	833,549 (17.05m)	58.3%

Table 1: **Numbers of men and women**

Table 2 describes our subsample from the NCHS dataset. In this table (rt, N) for instance refers to marriages in which the husband lists a reform state as his residence, and the wife lists a non-reform state. In more than 95% of marriages, husband and wife list a state with the same "reform status". This is not surprising since a large majority of marriages in fact unite partners from the same state. As in Choo and Siow, the number of marriages increased much more in reform states than in non-reform states; but also less than the general population.

Wave	1971–72	1981–82	Population increase
(r,R)	138,483 (838,140)	424,416 (1.00m)	19.4%
(r,N)	5,866 (38,518)	10,383 (32,952)	−14.5%
(n,R)	6,108 (33,440)	10,182 (24,530)	−26.6%
(n,N)	216,428 (1.70m)	506,953 (1.79m)	4.9%
Total	366,885 (2.61m)	951,934 (2.84m)	8.9%

Table 2: **Numbers of marriages**

Finally, Table 3 shows that the average age at marriage increased by two years, quite uniformly across reform status and genders. As a consequence, the age difference also did not change, with husbands two years older than their wives.

Wave		1971–72	1981–82	Increase
Reform states	Men	23.0	25.1	2.1
	Women	20.9	23.0	2.1
Non-reform states	Men	22.7	24.7	2.0
	Women	20.6	22.6	2.0

Table 3: **Ages at marriage**

G.2 Zero cells

Like much discrete-valued economic data, the Choo and Siow data contains a small but non-negligible percentage of (x, y) cells with no observed match—up to 3%, depending on the subsample[26]. The CS specification by construction rules out zero probability cells,

[26] Trade is another area where matching methods have become popular in recent years (see Costinot and Vogel, 2015); and trade data also has typically many zero cells.

and Choo and Siow (2006, footnote 15, p. 186) used kernel smoothers to impute positive probabilities in these "zero cells". More generally, no separable model with full support can simulate zero cells (see our discussion of Assumption 2).

This is not an issue with unrestricted estimation, since we only need to assign a value of $-\infty$ to the corresponding Φ_{xy}. With parametric inference, maintaining Assumption 2 implies that the model is misspecified. This is a minor consideration in practice, as the estimated probabilities of these cells turn out to be very small. A cleaner alternative is to specify error distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ that do not have full support, either because their supports have lower dimension and/or because their supports are bounded.

G.3 Detailed Estimation Results

G.3.1 Selecting Basis Functions

We used our moment matching method to estimate 625 semilinear versions of the original Choo and Siow (2006) specification, which we will call "the homoskedastic logit model". They all include the two basis functions $\phi^1_{xy} \equiv 1$ and $\phi^2_{xy} = D_{xy} \equiv \mathbb{1}(x \geq y)$, where x is the age of the husband and y that of the wife—both linearly transformed to be in [-1,1]. The D term accounts for possible jumps or kinks in surplus when the wife is older than the husband ($D = 0$). In addition to these two basis functions, we include a varying set of functions of the form $x^i y^j$ and $x^i y^j D$. Our richest candidate specification has 98 basis functions; note that the nonparametric model has 625 (as many as marriage cells.)

Figure 6 plots the values of the Akaike Information Criterion (on the horizontal axis) and of the Bayesian Information Criterion (on the vertical axis) for the 625 models, and for the nonparametric model NP. The location of NP shows that even for our sample of a couple hundred thousand observations, it is severely overparameterized: no fewer than 490 of our 625 models have a better AIC, and all of them have a better BIC.

Our best AIC model is still large: it has 60 coefficients, of which 49 are significant at 5%. With such a large sample, we could probably have included even higher-degree terms

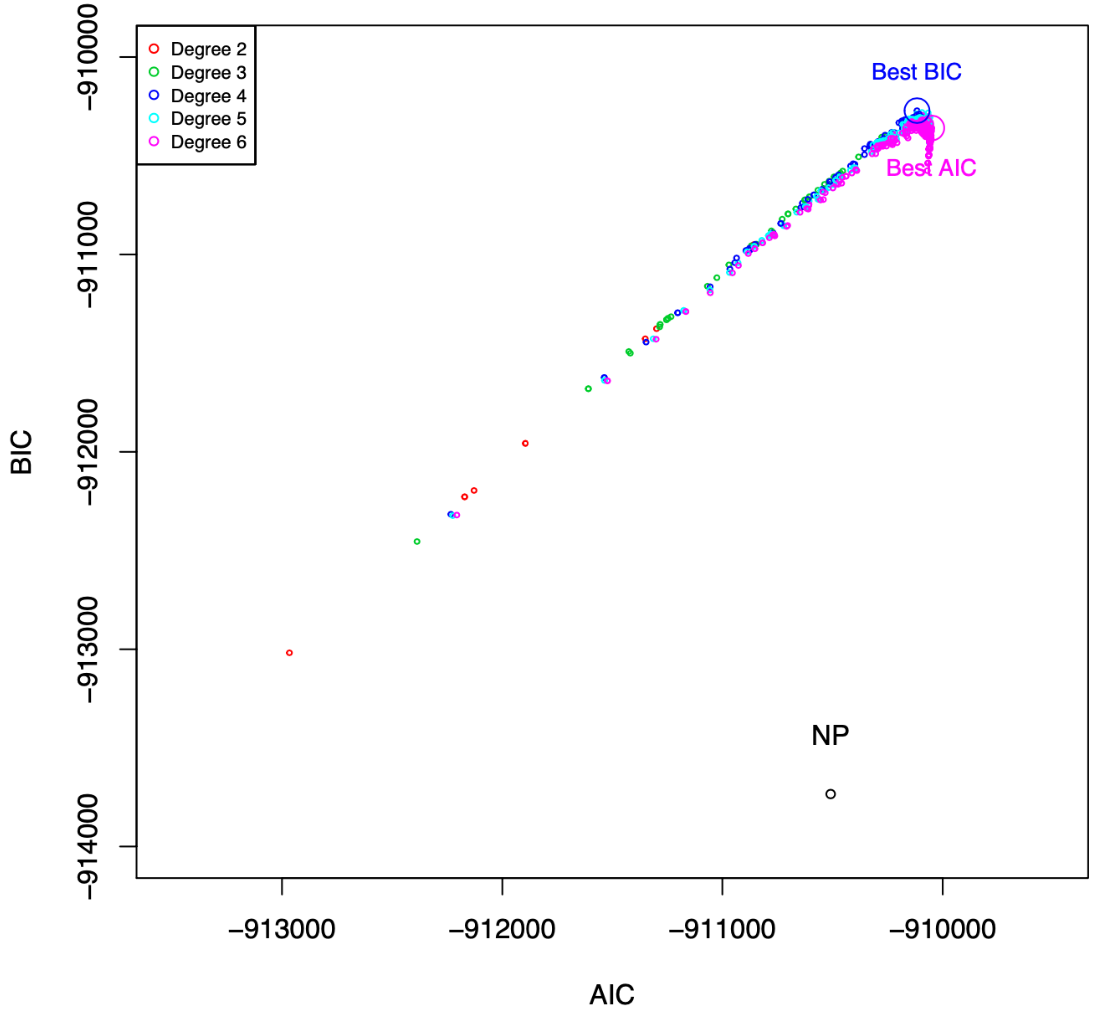


Figure 6: AIC and BIC Values for the Parametric and Nonparametric Choo and Siow Models

and improved the AIC slightly. While the AIC criterion subtracts twice the number of parameters from the log-likelihood, the BIC criterion penalizes it by half of the logarithm of the number of observations. With our 224,068 observations, this amounts to 6.2 rather than 2 times the number of parameters. As a result, the BIC-selected model only has 30 coefficients, of which 28 differ significantly from 0 at the 5% level. For model selection (as opposed to forecasting), BIC is more appropriate than AIC and we will work with its 30 selected basis functions from now on: all terms $x^m y^n$ and $x^m y^n D$ for $1 \leq m \leq 2$ and $1 \leq n \leq 4$.

G.3.2 The Homoskedastic Choo and Siow Model

Table 4 gives the estimated coefficients and their bootstrapped standard errors and Students for the BIC-preferred modelin this class.

Estimates Table 4 in Appendix G.3 collects our estimates for the coefficients of the BIC-preferred model with iid standard type I EV errors. Since the distributions $\mathbb{P}_x$ and $\mathbb{Q}_y$ are parameter-free in this model, the table shows the estimated coefficients for the 30 basis functions in its first column. We evaluated their standard errors (third column) with a bootstrap procedure based on 999 draws from the estimated variance-covariance matrix of the observed matching patterns $\hat{\boldsymbol{\mu}}$.

The bootstrap also allows us to compute a p-value for the entropy test described in Theorem 6. The value of the entropy test statistic in the sample has a bootstrapped p-value is 0.856. Recall that this tests the hypothesis that the true surplus function is a linear combination of our 30 basis functions, conditional on the distributional assumptions being true. The p-value tells us that this "spanning hypothesis" would only be rejected at the 15% level. This confirms that the 30-bases model is a very good approximation to the data-generating process. The Choo and Siow model aims at explaining marriage patterns by age, from age 16 to age 75. In the early 1970s, close to 80% of marriages occurred before either partner was 30 years old, so that the number of data points to fit is rather small.

	Estimates	Standard Errors	Students
1	-11.163	0.023	-490.9
D	1.147	0.066	17.3
X	-14.759	0.336	-44.0
XD	5.204	0.134	38.7
X^2	-13.211	0.208	-63.4
X^2D	5.656	0.104	54.2
Y	-1.220	0.066	-18.4
YD	4.757	0.127	37.5
Y^2	-2.064	0.041	-50.7
Y^2D	5.950	0.118	50.5
Y^3	1.097	0.054	20.4
Y^3D	1.659	0.029	57.4
Y^4	-0.563	0.033	-17.0
Y^4D	-0.637	0.018	-35.5
XY	26.379	0.403	65.4
XYD	-16.697	0.336	-49.7
XY^2	-16.956	0.455	-37.3
XY^2D	10.238	0.298	34.3
XY^3	6.206	0.336	18.4
XY^3D	-3.936	0.227	-17.3
XY^4	-0.997	0.144	-6.9
XY^4D	0.881	0.092	9.6
X^2Y	12.940	0.276	46.9
X^2YD	-11.549	0.226	-51.1
X^2Y^2	-5.636	0.303	-18.6
X^2Y^2D	4.938	0.229	21.5
X^2Y^3	1.131	0.196	5.8
X^2Y^3D	-1.053	0.137	-7.7
X^2Y^4	0.085	0.060	1.4
X^2Y^4D	-0.072	0.050	-1.4

Table 4: Estimates for the Homoskedastic Logit Model

Even using BIC to reward parsimony, with more than 200,000 observations we end up with a rich model and an excellent fit.

As a consequence, the distributional parameters we introduce can only improve the fit marginally. We did find, however, that allowing for gender- and age-dependent heteroskedasticity yielded a notable improvement in the fit. Interestingly, it also changes the profile of surplus-sharing within couples: the share that goes to the husband increases much more steeply than in the original (homoskedastic) Choo and Siow specification. We also fitted several Generalized Extreme Values models. The most promising ones seem to be those of the FC-MNL family (Davis and Schiraldi, 2014), which incorporate the type of local correlation patterns that are missing from the multinomial logit framework. While they do not outperform the basic Choo and Siow specification in our application, they are easy to implement and seem to us to have much potential in matching models.

Heteroskedastic Logit Models We explored several ways of adding heteroskedasticity to the benchmark model. It is clear from 1.2 that the parameters can only be identified up to a scale normalization: multiplying both $\mathbf{\Phi}$ and the error terms $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ by the same positive number has no effect on the equilibrium matching. The Choo and Siow (2006) model normalizes the scale (twice) by using standard type I EV errors for both $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$. When adding heteroskedasticity to $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$, we need to maintain one normalization.

Our simplest heteroskedastic model still uses a standard type I EV $\boldsymbol{\varepsilon}$ (our scale normalization) and adds only one parameter τ, with

$$\tau^2 = \frac{V\eta}{V\varepsilon}.$$

This model allows for heteroskedasticity across genders, but not across types. Somewhat surprisingly, the profiled loglikelihood of the model is very flat with respect to τ. While we did obtain an estimate of 0.927 that is slightly lower than one, the improvement in the loglikelihood is so small that the values of both AIC and BIC deteriorate.

Going further, we allow for type- and gender-dependent heteroskedasticity[27]. To do this, we multiply the terms $\boldsymbol{\varepsilon}_{i\cdot}$ (resp. $\boldsymbol{\eta}_{j\cdot}$) by scale factors σ_x (resp. τ_y). We experimented with specifications of the form

$$\begin{aligned}\sigma_x &= \exp(\sigma_1 x + \ldots + \sigma_p x^p) \\ \tau_y &= \exp(\tau_0 + \tau_1 y + \ldots + \tau_q y^q)\end{aligned}$$

Note that we do not allow for a constant term σ_0; this gives us the requisite scale normal-

[27] Chiappori, Salanié, and Weiss (2017) attempted to estimate a similar model, with education as the type.

ization.

Of all such specifications for $0 \leq p \leq 4$ and $1 \leq q \leq 4$, this yields the largest improvement in the fit: a sizeable +38.5 points of loglikelihood, and +25.2 points on BIC. The estimates of the parameters of σ_x and τ_y can be found in Table 5.

	Estimates	Standard Errors	Students
σ_1	0.793	0.051	15.4
τ_0	-0.751	0.161	-4.7

Table 5: Estimates for the Heteroskedastic Logit Model: Distributional Parameters

Two-level, Two-nest Nested Logit We estimated a two-level nested logit model in which we separate the singlehood option from all others. This model has two nests: one corresponding to singlehood, and one to the 25 possible ages of the partner. It introduces two additional parameters, γ_m on the men side and γ_w for women. The familiar equation from Choo and Siow (2006):

$$2 \log \mu_{xy} = \log \mu_{x0} + \log \mu_{0y} + \Phi_{xy}$$

becomes

$$\gamma_m \log \frac{\mu_{xy}}{\sum_{t \in \mathcal{Y}} \mu_{xt}} + \gamma_w \log \frac{\mu_{xy}}{\sum_{z \in \mathcal{X}} \mu_{zy}} = \log \frac{\mu_{x0}}{\sum_{t \in \mathcal{Y}} \mu_{xt}} + \log \frac{\mu_{0y}}{\sum_{z \in \mathcal{X}} \mu_{zy}} + \Phi_{xy}.$$

The values of $(1 - \gamma_m)$ and $(1 - \gamma_w)$ can be interpreted as “within-nest correlations”; they equal zero in the Choo and Siow (2006) model.

We chose this specific nested logit model because we showed in Galichon and Salanié (2019) that it satisfies a “weak IIA” property–and we conjectured that it is the only separable model that does. When we tried to estimate this two-nest specification, we consistently found a corner maximum at $\gamma_m = 1$. The other parameter γ_w has a weak maximum at 0.91, and the loglikelihood barely improves.

FC-MNL Davis and Schiraldi (2014) show that for any admissible values of σ and τ, there exist values of the b matrix that rationalize a given set of elasticities of substitution. We followed their suggestion of using $\sigma = 0.5$ and $\tau = 1.1$; and we chose the very parsimonious specification of the b matrix described in Section 6.2. The maximum likelihood estimates for the distributional parameters[28] are in Table 6.

	Estimates
$b_m(16)$	0.011
$b_m(40)$	0.000
$b_w(16)$	0.060
$b_w(40)$	0.000

Table 6: Estimates for the FC-MNL Model: Distributional Parameters

[28] Given the small gain in the loglikelihood, the standard errors are large.