Cupid's Invisible Hand:

Social Surplus and Identification in Matching Models

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

May 17, 2020[‡]

Abstract

We investigate a model of one-to-one matching with transferable utility when some of the characteristics of the agents are unobservable to the analyst. We allow for a wide class of distributions of unobserved heterogeneity, subject only to a separability assumption that generalizes Choo and Siow (2006). We first show that the stable 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 surplus in every possible match and the equilibrium utilities of all participants, given any known distribution of unobserved heterogeneity. We formulate a parametric version of the model and show several estimation techniques. We discuss computational and inference issues and provide efficient algorithms. Finally, we revisit Choo and Siow's empirical application to illustrate the potential of our more general approach.

Keywords: matching, marriage, assignment, hedonic prices.
JEL codes: C78, D61, C13.

[*] Economics Department, FAS and Mathematics Department, Courant Institute, New York University; 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, 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 grant FP7-313699.

Introduction

Since the seminal contribution of Becker (1973), many economists have modeled the marriage market as a matching problem in which each potential match generates a marital surplus. When utility is perfectly transferable, 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 can be used in all of these contexts; for concreteness, we often refer to partners as "men" and "women" 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 socially optimal matches then exhibit *positive assortative matching*: higher types pair up with higher types. Moreover, the resulting configuration is stable, it is in the core of the corresponding matching game, and it can be efficiently implemented by a simple sorting procedure. 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 distributed as standard type I extreme value terms. They used their model to evaluate the

effect of the legalization of abortion on gains to marriage; and they applied it 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, 2017) 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 (2019) extend the theory to same-sex marriage.

We revisit here the theory of matching with transferable utilities in the light of Choo and Siow's insights. Three assumptions underlie their contribution: latent variables do not mutually interact in producing matching surplus, they are distributed as iid type I extreme values, and populations are large. We maintain the first assumption, which we call "separability", 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. We prove that the optimal matching in our generalized setting maximizes the sum of a term that describes matching on the observables and a generalized entropic term that describes matching on the unobservables. While the first term tends to match partners with complementary observed characteristics, the second one pulls towards "randomly" assigning partners to each other—in a sense that we will make clear. The social gain from any matching pattern trades off between these two terms. In particular, when unobserved heterogeneity is distributed as in Choo and Siow (2006), the generalized entropy is simply the usual entropy measure.

The maximization of this social surplus function has very straightforward consequences in terms of identification, both when equilibrium transfers are observed and when they are not. In fact, joint surplus and expected utilities can be obtained from derivatives of the terms that constitute generalized entropy; and that in turn is a function of observed

matching probabilities. Moreover, if equilibrium transfers are observed, then we also identify the pre-transfer utilities on both sides of the market. In independent work, Decker, Lieb, McCann, and Stephens (2012) proved the uniqueness of the equilibrium and analyzed its properties in the Choo and Siow multinomial logit framework; and Graham (2013) also analyzed the comparative statics of the multinomial logit model. We prove that several of their results in fact hold in *all* separable models; and we show how to derive additional testable predictions for more restricted specifications.

Our first conclusion therefore is that the Choo-Siow framework can be extended to encompass much less restrictive assumptions on the unobserved heterogeneity. Our second contribution is to delineate an empirical approach to parametric estimation in this class of models. Our identification results rely on the strong assumption that the distribution of the unobservables is known. This is unavoidable: the matching surplus cannot be estimated nonparametrically along with the distribution of the unobservables, as there would be many more parameters than cells in the data matrix. In practice, the analyst will want to estimate the parameters of this distribution. This suggests using a restricted parameterization for match surpluses.

Maximum likelihood estimation is a natural way to estimate parametric separable models, which we investigate in section 4.1. We also discuss in section 4.2 an alternative to the maximum likelihood in models with a linear parameterization of surplus and known distributions of heterogeneity, This consists of a simple moment-matching estimator that minimizes a generalized entropy among those matching distributions that fit a set of moments. In addition to yielding a globally convex problem, moment-matching suggests a very simple semi-parametric specification test. In addition, it is particularly simple to implement with the Choo and Siow (2006) specification.

In practice, computational considerations loom large in matching models; and evaluating the likelihood requires solving for the optimal matching repeatedly. We provide two alternative algorithms that maximize the social surplus and compute the optimal matching, as well as the expected utilities in equilibrium. The first one, which we call the "min-Emax"

method, uses a minimization of the sum of the inclusive values. The second one adapts the Iterative Projection Fitting Procedure (IPFP—known to some economists as RAS) to the structure of this problem. We show that that both procedures are stable and efficient in practice. We also describe (in appendix D.2.1) how to discretize distributions so that the min-Emax becomes a linear programming problem; and we show how this extends to moment-matching estimation of the parameters in the model of section 4.2. Since linear programming solvers have become very fast, this provides a promising approach to estimation.

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. To do this, we start with a semilinear Choo and Siow model. We use the Bayesian Information Criterion to select a specification. This has 30 basis functions, and it fits the data very well. 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. Still, allowing for gender- and age-dependent heteroskedasticity yields a notable improvement in the fit; it also generates a finding that the husband's share of the surplus in same-age couples increases rather steeply. We also find promise in the class of flexible multinomial logit models introduced by Davis and Schiraldi (2014), which allow for local correlation patterns. They are easy to implement and seem to us to have much potential in matching models. These results illustrate that our approach is both practical and fruitful.

There are other approaches to estimating matching models with unobserved heterogeneity; see the handbook chapter by Graham (2011, 2014) and the survey by Chiappori and Salanié (2016) and Chiappori (2020). For markets with transferable utility, Fox (2010) has proposed pooling data across many similar markets and relying on a "rank-order property". This assumes that given the characteristics of the populations of men and women, a given matching is more likely than another when it produces a higher expected surplus. 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 focal points—in particular with Menzel (2015)'s investigation of large NTU markets. 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.

Section 1 sets up the model and the notation and characterizes equilibrium. We prove our main identification results in section 2, where we also discuss testable predictions of both the general class of separable models and some of its instances. We present some leading examples in section 3, including the Choo-Siow framework for which we spell out some counterintuitive predictions, and then moving beyond the logit case. Our results open the way to new and richer specifications; section 4 explains how to estimate them using maximum likelihood estimation, and how to use various restrictions to identify the underlying parameter. We also show there that a moment-based estimator is an excellent low-cost alternative in a restricted but useful class of models, which includes Choo and Siow. We present in section 5 our computational methods, which make it easy to solve for the equilibrium. Finally, section 6 applies several instances of separable models to Choo and Siow's dataset. We conclude with a brief discussion of our assumptions and of extensions to our results.

Our arguments use tools from convex analysis as well as optimal transportation. We have tried to keep the exposition intuitive in the body of the paper; all proofs can be found in Appendix A. We offer several complements in Appendix B, including a geometric interpretation of some of our results. Appendix C specializes our results to several two-sided versions of commonly used discrete-choice models. Appendix D details our IPFP algorithm in pseudo-code, and it gives simulation results for this and other algorithms. Appendix E provides a detailed description of the data we use in Section 6 and Appendix F gives detailed estimation results.

1 The Assignment Problem with Unobserved Heterogeneity

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 but our results are clearly not restricted to a marriage context.

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—or, more precisely, of 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 feasibility constraints states that the mass of married individuals in each group cannot be greater

than the mass of individuals in that group, which is denoted $\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{r})$, where $\mathcal{M}(\boldsymbol{n}, \boldsymbol{m})$ (or $\mathcal{M}$ in the absence of ambiguity) is defined by:

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

Each element of $\mathcal{M}$ is called 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:

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

We shall now discuss the "joint surplus" Φ_{ij}, which is the sum of the utilities of man i and woman j if they match. As shown in Chiappori, Salanié, and Weiss (2017), an important assumption made implicitly in Choo and Siow is that the joint surplus (which they call the *total systematic net gains to marriage*) 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 exists a matrix* $\boldsymbol{\Phi}$ *such that*

(i) the joint surplus from a match between a man i *in group* x *and a woman* j *in group* j *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* $|\mathcal{Y}_0|$*-dimensional random vector* $\boldsymbol{\varepsilon_i} = (\varepsilon_{iy})_y$ *has probability distribution* $\boldsymbol{P}_x$*, and, conditional on* $y_j = y$*, the* $|\mathcal{X}_0|$*-dimensional distribution of random vector* $\boldsymbol{\eta_j} = (\eta_{xj})_x$ *has probability distribution* $\boldsymbol{Q}_y$*. The variables*

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

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

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 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 explain in section 3.2 why these extensions are useful for applications.

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 $\boldsymbol{Q}_y$.

1.2 Discrete choice and generalized entropy

Before we solve the two-sided matching problem, it is useful to consider an associated discrete choice problem. We consider men who choose between various marital options according to a random utility model: man i in group x chooses a partner in the group y that maximizes $(U_{xz} + \varepsilon_z)$ over z in $\mathcal{Y}_0$.

We will show that the ex-ante indirect surplus can be expressed as a sum of two terms: the weighted sum of the systematic utilities, and a "generalized entropy of choice" which comes 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 surplus, and the other one as the solution to an optimal transport problem[1] To the best of our knowledge, these results, which are summarized in Theorem 1 below, are new.

Consider a man i in group x, and a vector $\boldsymbol{U} = (U_{x1}, \ldots, U_{x|\mathcal{Y}|})$. This man could either stay single and attain utility ε_{i0}, or match with a partner in some group y and attain utility $(U_{xy} + \varepsilon_{iy})$. The expected utility of men of group x therefore is

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \boldsymbol{E}_{\boldsymbol{P}_x} \max_{y \in \mathcal{Y}_0} (U_{xy} + \varepsilon_y) = \boldsymbol{E}_{\boldsymbol{P}_x} \max \left(\max_{y \in \mathcal{Y}} (U_{xy} + \varepsilon_y), \varepsilon_0 \right) \tag{1.3}$$

[1] See Galichon (2016) for an introduction to optimal transport with a focus on economics.

where the expectation is taken over the random vector $(\varepsilon_0, \ldots, \varepsilon_{|\mathcal{Y}|}) \sim \boldsymbol{P}_x$. The function G_x is known as the *Emax operator* for group x in the discrete choice literature. It will be convenient to define the *aggregate welfare of men* by

$$G(\boldsymbol{U}, \boldsymbol{n}) = \sum_{x \in \mathcal{X}} n_x G_x(\boldsymbol{U}_{x\cdot}) \tag{1.4}$$

for $\boldsymbol{U} = (U_{xy})_{x \in \mathcal{X}, y \in \mathcal{Y}}$.

Note that as the expectation of the maximum of linear functions of the (U_{xy}), G_x is a convex function of $\boldsymbol{U}_{\boldsymbol{x}\cdot}$; and G is a convex function of $\boldsymbol{U}$. Let $Y_i^* \in \mathcal{Y}_0$ denote the optimal choice of marital option by a man i of group x; then with $U_{x0} = 0$, we have

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \boldsymbol{E}_{\boldsymbol{P}_x}\left(U_{xY_i^*} + \varepsilon_{iY_i^*}\right) = \sum_{y \in \mathcal{Y}} \mu_{y|x} U_{xy} + \boldsymbol{E}_{\boldsymbol{P}_x}\left(\varepsilon_{iY_i^*}\right), \tag{1.5}$$

where $\mu_{y|x}$ is the conditional choice probability of option y by an individual of group x.

Our analysis gives a prominent role to the *Legendre-Fenchel transform* of G_x, which is defined in convex analysis by

$$G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = \sup_{\tilde{\boldsymbol{U}}_{\boldsymbol{x}\cdot} = (\tilde{U}_{x1}, \ldots, \tilde{U}_{x|\mathcal{Y}|})} \left(\sum_{y \in \mathcal{Y}} \mu_{y|x} \tilde{U}_{xy} - G_x(\tilde{\boldsymbol{U}}_{\boldsymbol{x}\cdot}) \right) \tag{1.6}$$

whenever $\sum_{y \in \mathcal{Y}} \mu_{y|x} \leq 1$, and $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = +\infty$ otherwise. Note that the domain of G_x^* is the set of $\mu_{\cdot|x}$ that can be interpreted as vectors of choice probabilities of alternatives in $\mathcal{Y}$. As the supremum of a set of linear functions of the $\mu_{y|x}$, G_x^* is a convex function. The theory of convex duality implies that since G_x is convex, it coincides with its bitransform:

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \sup_{\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}\cdot} = (\tilde{\mu}_{1|x}, \ldots, \tilde{\mu}_{|\mathcal{Y}||x})} \left(\sum_{y \in \mathcal{Y}} \tilde{\mu}_{y|x} U_{xy} - G_x^*(\tilde{\boldsymbol{\mu}}_{\cdot|\boldsymbol{x}}) \right). \tag{1.7}$$

Now assume that $\boldsymbol{\mu}_{y|x}$ attains the supremum in (1.7). Then

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) + G_x^*(\boldsymbol{\mu}_{\cdot|x}) = \sum_{y \in \mathcal{Y}} \mu_{y|x} U_{xy};$$

and comparing with (1.5) shows that

$$G_x^*\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}\right) = -\boldsymbol{E}_{\boldsymbol{P}_x}\left(\varepsilon_{iY_i^*}\right). \tag{1.8}$$

Hence $-G_x^*\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}\right)$ is just the expected heterogeneity that is required to rationalize the conditional choice probability vector $\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}$. We will see in section 3.1 that in the logit setting, $-G_x^*$ is the usual entropy function. This motivates the following definition:

Definition 1. *We call the function $-G_x^*$ the* generalized entropy of choice *of men of group x. For $\boldsymbol{\mu} = (\mu_{xy})_{x\in\mathcal{X},y\in\mathcal{Y}}$, we denote*

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

the generalized entropy of choice of all men.

The following result goes beyond formula (1.8) 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 $\mathcal{M}\left(\boldsymbol{\mu}_{\cdot|x},\boldsymbol{P}_x\right)$ denote the set of probability distributions π of the random joint vector $(\boldsymbol{Y},\boldsymbol{\varepsilon})$, where $\boldsymbol{Y}\sim\boldsymbol{\mu}_{\cdot|x}$ is a random element of $\mathcal{Y}_0$, and $\boldsymbol{\varepsilon}\sim\boldsymbol{P}_x$ is a random vector of $\mathbb{R}^{\mathcal{Y}_0}$. Then $-G_x^*(\boldsymbol{\mu}_{\cdot|x})$ is the value of the optimal matching problem between the distribution $\boldsymbol{\mu}_{\cdot|x}$ of $\boldsymbol{Y}$ and the distribution $\boldsymbol{P}_x$ of $\boldsymbol{\varepsilon}$, when the surplus is given by $\varepsilon_{\boldsymbol{Y}}$. That is,*

$$-G_x^*(\boldsymbol{\mu}_{\cdot|x}) = \sup_{\pi\in\mathcal{M}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}},\boldsymbol{P}_x\right)}\boldsymbol{E}_\pi\left(\varepsilon_{\boldsymbol{Y}}\right). \tag{1.9}$$

if $\sum_{y\in\mathcal{Y}_0}\mu_{y|x}=1$, and $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}})=+\infty$ otherwise.

Since problem (1.9) 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.

1.3 The social surplus of matching

The results of the previous subsection will help us provide an expression for the social surplus in the matching problem. We now show that the total welfare is the sum of the systematic joint surpluses of all realized matches and an entropic term which is the sum of the generalized entropy of choice of all men and women.

We start with an intuitive derivation of our main result, Theorem 2 below. First, we argue that separability (Assumption 1) reduces the choice of partner to a one-sided discrete choice problem over the groups of partners. To see this, note that by standard results in the literature (Shapley and Shubik, 1972), the equilibrium utilities $(\tilde{u}_i)$ and $(\tilde{v}_j)$ of men and women solve the system of functional equations[2]

$$\tilde{u}_i = \max_{j,0} \left(\tilde{\Phi}_{ij} - \tilde{v}_j \right) \text{ and } \tilde{v}_j = \max_{i,0} \left(\tilde{\Phi}_{ij} - \tilde{u}_i \right),$$

where the maximization includes the option of singlehood. Focus on the first one, and recall that $\tilde{\Phi}_{ij} = \Phi_{xy_j} + \varepsilon_{iy_j} + \eta_{xj}$, so that $\tilde{u}_i = \max_j \left(\tilde{\Phi}_{ij} - \tilde{v}_j \right) = \max_y \max_{j:y_j=y} \left(\tilde{\Phi}_{ij} - \tilde{v}_j \right)$ can be rewritten as $\tilde{u}_i = \max_y \left(\Phi_{xy} + \varepsilon_{iy} - \min_{j:y_j=y} \left(\tilde{v}_j - \eta_{xj} \right) \right)$. Denoting $V_{xy} = \min_{j:y_j=y} \left(\tilde{v}_j - \eta_{xj} \right)$ and $U_{xy} = \Phi_{xy} - V_{xy}$, it follows that:

$$\tilde{u}_i = \max \left(\max_{y \in \mathcal{Y}} \left(U_{xy} + \varepsilon_{iy} \right), \varepsilon_{i0} \right) \text{ and similarly } \tilde{v}_j = \max \left(\max_{x \in \mathcal{X}} \left(V_{xy} + \eta_{xj} \right), \eta_{0j} \right).$$

Recall that G_x is the expected value of $\tilde{u}_i$ conditional on $x_i = x$, and $G(\boldsymbol{U}) = \sum_{x \in \mathcal{X}} n_x G_x \left(\boldsymbol{U}_{x\cdot} \right)$ is the aggregate social welfare. We define H_y and H on women's side similarly: a randomly drawn woman of group y expects to get utility

$$H_y(\boldsymbol{V}_{\cdot \boldsymbol{y}}) = \boldsymbol{E}_{\boldsymbol{Q}_y} \left(\max \left(\max_{x \in \mathcal{X}} \left(V_{xy} + \eta_x \right), \eta_0 \right) \right),$$

and the welfare of all women is

$$H(\boldsymbol{V}, \boldsymbol{m}) = \sum_{y \in \mathcal{Y}} m_y H_y(\boldsymbol{V}_{\cdot \boldsymbol{y}}).$$

[2] Alert readers will note that we are taking some liberties here: we are using maxima for continuous sets, we are neglecting issues that arise when $\mu_{xy} = 0$, etc. The proof in Appendix A deals with all of these points.

As we did for the men's side, we define H_y^* and H^*, respectively the generalized entropy of choice of women of group y and of all women, as the respective Legendre-Fenchel transforms of H_y and H. The social surplus $\mathcal{W}$ is simply the sum of the expected utilities of all groups of men and women:

$$\mathcal{W} = \sum_{x \in \mathcal{X}} n_x G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) + \sum_{y \in \mathcal{Y}} m_y H_y(\boldsymbol{V}_{\cdot \boldsymbol{y}}) = G(\boldsymbol{U}, \boldsymbol{n}) + H(\boldsymbol{V}, \boldsymbol{m}). \tag{1.10}$$

Expression (1.10) explains how the total surplus $\mathcal{W}$ is broken down between men and women: men get surplus $G(\boldsymbol{U}, \boldsymbol{n})$, while women get $H(\boldsymbol{V}, \boldsymbol{m})$. Note that the function G (resp. H) is homogeneous of degree 1 in $\boldsymbol{n}$ (resp. $\boldsymbol{m}$).

Letting (μ_{xy}) be the optimal matching, summing the expressions (1.7) over x and doing similarly on the other side, we get that at the optimum,

$$G(\boldsymbol{U}, \boldsymbol{n}) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mu_{xy} U_{xy} - G^*(\boldsymbol{\mu}, \boldsymbol{n}) \text{ and } H(\boldsymbol{V}, \boldsymbol{m}) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mu_{xy} V_{xy} - H^*(\boldsymbol{\mu}, \boldsymbol{m}).$$

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

$$\mathcal{W} = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mu_{xy} \Phi_{xy} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{m}) \tag{1.11}$$

where we have defined

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{m}) := -G^*(\boldsymbol{\mu}, \boldsymbol{n}) - H^*(\boldsymbol{\mu}, \boldsymbol{m}). \tag{1.12}$$

In contrast with expression (1.10) which explains the *destination* of the surplus shared at equilibrium between men and women, expression (1.11) explains the *origin* of the surplus: $\sum \mu_{xy} \Phi_{xy}$ originates from the part of the surplus that comes from the interaction between observable characteristics, while $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{m})$ originates from unobservable heterogeneities in tastes. Note that like its components G^* and H^*, the function $\mathcal{E}$ takes infinite values for infeasible matchings $\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{r})$.

The following result makes this intuition precise by showing that the values of all of these variables emerge from the solution to an optimization problem:

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

$$\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{r}) = \max_{\boldsymbol{\mu} \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}} \left(\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \mu_{xy} \Phi_{xy} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{r}) \right). \tag{1.13}$$

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

$$\begin{aligned} \mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{r}) = \quad & \min_{\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}} \left(G(\boldsymbol{U}, \boldsymbol{n}) + H(\boldsymbol{V}, \boldsymbol{m}) \right) \\ s.t. \quad & U_{xy} + V_{xy} \geq \Phi_{xy} \ \forall x \in \mathcal{X}, y \in \mathcal{Y}. \end{aligned} \tag{1.14}$$

The proof of this result, which is given in appendix A, consists of writing the infinite-dimensional linear programming duality of Gretsky, Ostroy, and Zame (1992) and making use of the separability assumption to transform it into a finite-dimensional nonlinear (convex) programming duality. While (1.14) only gives us the equilibrium $\boldsymbol{U}$ and $\boldsymbol{V} = \boldsymbol{\Phi} - \boldsymbol{U}$, the matching patterns obtain easily by using either

$$\mu_{xy} = \frac{\partial G}{\partial U_{xy}}(\boldsymbol{U}, \boldsymbol{n}) \ \text{ or } \ \mu_{xy} = \frac{\partial H}{\partial V_{xy}}(\boldsymbol{V}, \boldsymbol{m}). \tag{1.15}$$

Note that the first-order conditions of (1.14) 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}.

The right-hand side of equation (1.13) gives the value of the social surplus when the matching patterns are $\boldsymbol{\mu}$. Its first term $\sum_{xy} \mu_{xy} \Phi_{xy}$ reflects "group preferences": if groups x and y generate more surplus when matched, then in the absence of unobserved heterogeneity they should be matched with higher probability. In this case Theorem 2 boils down to the classical linear programming duality for the assignment problem described in Shapley and Shubik (1972). On the other hand, the second term $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{r})$ reflects the effect of the dispersion of individual affinities, conditional on observed characteristics: those men i in a group x that have more affinity to women of group y should be matched to this

group with higher probability. If available data were so poor that unobserved heterogeneity dominates ($\mathbf{\Phi} \simeq 0$), then the analyst should observe something that, to her, looks like completely random matching. Information theory tells us that entropy is a natural measure of statistical disorder; and as we will see in Example 1, in the simple case analyzed by Choo and Siow the function $\mathcal{E}$ is just the usual notion of entropy. This is why we call it the *generalized entropy of matching*. When some of the variation in marital surplus is driven by group characteristics (through the Φ_{xy}) and some is carried by the unobserved heterogeneity terms ε_{iy} and η_{xj}, the market equilibrium trades off matching on group characteristics and matching on unobserved characteristics, as measured by the generalized entropy terms in $\mathcal{E}$.

Theorem 2 has two corollaries which will prove useful later. The first one rewrites the dual formulation as a lower-dimensional program, with expected utilities $\boldsymbol{u}$ and $\boldsymbol{v}$ as its arguments.

Corollary 1. *Let $E(\boldsymbol{\mu};\boldsymbol{r})$ be any strictly concave function of $\boldsymbol{\mu}$ that coincides with the generalized entropy $\mathcal{E}(\boldsymbol{\mu};\boldsymbol{r})$ over the set of feasible matchings $\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{r})$. For $\boldsymbol{a} \in \mathbb{R}^{\mathcal{X}}$ and $\boldsymbol{b} \in \mathbb{R}^{\mathcal{Y}}$, define $S(\boldsymbol{a},\boldsymbol{b})$ as the value of*

$$\max_{\boldsymbol{\mu}} \left(E(\boldsymbol{\mu};\boldsymbol{r}) + \sum_{x,y \in \mathcal{X}\times\mathcal{Y}} \mu_{xy}(\Phi_{xy} - a_x - b_y) + \sum_{x\in\mathcal{X}} (n_x - \mu_{x0}) a_x + \sum_{y\in\mathcal{Y}} (m_y - \mu_{0y}) b_y \right). \tag{1.16}$$

Then S is a convex function of $(\boldsymbol{a},\boldsymbol{b})$. The social welfare is the matching problem $\mathcal{W}$ which is the value of (1.14) is obtained by minimizing $S(\boldsymbol{u},\boldsymbol{v})$ over u and v, and the solutions $\boldsymbol{u}$ and $\boldsymbol{v}$ to (1.16) obtained this way are the expected utilities of the different types of men and women in equilibrium, and the solutions to (1.14) for $\boldsymbol{a} = \boldsymbol{u}$ and $\boldsymbol{b} = \boldsymbol{v}$ are the equilibrium matching patterns.

Corollary 1 will be instrumental in the derivation of an efficient algorithm in sections 5.2. We will also use it in our Proposition 5 to show a simple way of estimating variants of the Choo and Siow (2006) model.

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

implication of Theorem 2.

Corollary 2. *The function $\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{n}, \boldsymbol{m})$ is convex in $\boldsymbol{\Phi}$. It is homogeneous of degree 1 and concave in $\boldsymbol{r} = (\boldsymbol{n}, \boldsymbol{m})$.*

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

Proposition 1 (Individual and group surpluses)**.** *Under Assumption 1:*

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

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

where $U_{x0} = 0$, and $\boldsymbol{U}$ solves (1.14).

(ii) The average expected utility of the men of group x is

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

(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.

Point (i) also appears in Chiappori, Salanié, and Weiss (2017), with a different proof. It reduces the two-sided matching problem to a series of one-sided discrete choice problems that are only linked through the adding-up formula $U_{xy} + V_{xy} = \Phi_{xy}$. 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, then he would be equally well-off with any other woman of this group[3].

2 Identification and comparative statics

In this section, we focus on identifying $\boldsymbol{\Phi}$ (an array of $|\mathcal{X}| \times |\mathcal{Y}|$ unknown numbers) given the observation of $\boldsymbol{\mu}$ (an array of the same size.) In order to study the relation between

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

these two objects, we need to make assumptions on the distribution of the unobserved heterogeneity terms.

2.1 Identification of discrete choice problems

As in section 1.2, we start by considering one-sided discrete choice problems before moving on to matching problems. In this paragraph, we show that the generalized entropy of choice functions $-G_x^*$ allow for straightforward identification of systematic utilities based on the conditional choice probabilities.

Again, we first give the intuition of our result. By the Daly-Zachary-Williams theorem[4], we know that the derivative of the ex-ante indirect surplus of a man of group x with respect to U_{xy}, the systematic utility associated to marital option y, is equal to the conditional choice probability that this man chooses a partner in group y, that is

$$\frac{\partial G_x}{\partial U_{xy}}(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \frac{\mu_{xy}}{n_x} =: \mu_{y|x}. \tag{2.1}$$

The Fenchel duality theorem[5] implies that if G_x and G_x^* are continuously differentiable, this is equivalent to

$$\frac{\partial G_x^*}{\partial \mu_{y|x}}(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = U_{xy}. \tag{2.2}$$

As a consequence, for any $y \in \mathcal{Y}$, U_{xy} is identified from $\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}$, the observed matching patterns of men of group x. This is only true at points of differentiability. Since the functions G_x^* and H_y^* are convex, they are differentiable almost everywhere. For simplicity, we impose a full support assumption which ensures that G, G^*, H and H^* are continuously differentiable. It is not essential to our approach[6], but it makes for simpler formulæ and the functions G_x^* and $\boldsymbol{U}_{x\cdot}$ can be evaluated by solving an optimal transport problem (see Galichon, 2016).

Assumption 2 (Full support)**.** *The distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ all have full support and are absolutely continuous with respect to the Lebesgue measure.*

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

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

[6] It holds in most popular specifications, including the logit model of Choo and Siow of course.

Assumption 2 leads to the following identification result:

Proposition 2 (General identification of the systematic surpluses)**.** *Under Assumptions 1 and 2, the following statements are equivalent:*

(i) for every $y \in \mathcal{Y}_0$, $\mu_{y|x} = \dfrac{\partial G_x}{\partial U_{xy}}(\boldsymbol{U}_{\boldsymbol{x}\cdot})$

(ii) for every $y \in \mathcal{Y}_0$, $U_{xy} = \dfrac{\partial G_x^*}{\partial \mu_{y|x}}(\boldsymbol{\mu}_{\cdot|x})$

(iii) there exists a function $u_x(\boldsymbol{\varepsilon})$*, integrable with respect to* $\boldsymbol{P}_x$*, such that* $(u_x(\cdot), U_{x\cdot})$ *are the unique minimizers of the dual problem to* (1.9)*, that is of:*

$$\begin{aligned} \min_{\bar{U}_{x\cdot}, \bar{u}_x(\cdot)} \quad & \int \bar{u}_x(\boldsymbol{\varepsilon})\, d\boldsymbol{P}_x(\boldsymbol{\varepsilon}) - \sum_{y \in \mathcal{Y}} \mu_{y|x} \bar{U}_{xy} \\ s.t. \quad & \bar{u}_x(\boldsymbol{\varepsilon}) - \bar{U}_{xy} \geq \varepsilon_y \quad \forall y \in \mathcal{Y}, \forall \boldsymbol{\varepsilon} \in \mathbb{R}^{\mathcal{Y}_0} \\ & \bar{U}_{x0} = 0. \end{aligned} \tag{2.3}$$

The primal part, point (i), is well-known in the discrete choice literature; the dual parts, points (ii) and (iii), do not seem to be. The only related 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 (ii) and (iii) provides a constructive method to identify U_{xy} based on the conditional choice probabilities $\boldsymbol{\mu}_{\cdot|x}$, as the solution to a convex optimization problem (part (ii)) or more particularly, to an optimal transport problem (part (iii)). The intuition behind part (iii) is simply that each observed choice probability $\mu_{y|x}$ must be matched to the values of idiosyncratic preference shocks $\boldsymbol{\varepsilon}_i \sim \boldsymbol{P}_x$ for which y is the most preferred choice. The $\boldsymbol{U}_{\cdot|x}$ are the shadow prices that support this matching. Chiong, Galichon, and Shum (2016) apply our approach to dynamic discrete-choice models.

2.2 Identification of the Matching Surplus

We move on to the two-sided matching problem which is the focus of the paper. Given Proposition 1, we can decompose the matching problem into two series of discrete choice problems; and we can use Proposition 2 in order to identify the equilibrium utilities $\boldsymbol{U}$ and $\boldsymbol{V}$ as functions of $\boldsymbol{\mu}$:

Proposition 3. *Under Assumptions 1 and 2:*

(i) $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *are identified from* $\boldsymbol{\mu}$ *by*

$$\boldsymbol{U} = \frac{\partial G^*}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}) \text{ and } \boldsymbol{V} = \frac{\partial H^*}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}) \tag{2.4}$$

(ii) The constraint in (1.14) is always saturated: $U_{xy} + V_{xy} = \Phi_{xy}$ *for every* $x \in \mathcal{X}, y \in \mathcal{Y}$. *As a result, the matching surplus* $\boldsymbol{\Phi}$ *is identified by*

$$\Phi_{xy} = \frac{\partial G_x^*}{\partial \boldsymbol{\mu}_{y|x}}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}\right) + \frac{\partial H_y^*}{\partial \mu_{x|y}}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{y}}\right), \tag{2.5}$$

that is

$$\Phi_{xy} = -\frac{\partial \mathcal{E}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{r}). \tag{2.6}$$

Combining Proposition 2 with Proposition 3, all of the quantities in Theorem 2 can be computed by solving simple convex optimization problems.

2.3 Comparative statics and testable predictions

Recall from Proposition 1 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

$$\frac{\partial u_x}{\partial n_{x'}} = \frac{\partial u_{x'}}{\partial n_x}. \tag{2.7}$$

Hence the "unexpected symmetry" result proven by Decker, Lieb, McCann, and Stephens (2012) for the multinomial logit Choo and Siow model is a direct consequence of the symmetry of the Hessian of $\mathcal{W}$; in fact, it holds for *all* separable models. We show in Appendix B.3

that most of the comparative statics results of Decker et al extend to the present framework, and can be used as testable predictions for the general class of separable models.

Corollary 2 entails further consequences. Since the function $\mathcal{W}(\mathbf{\Phi}, \boldsymbol{r})$ 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[7]. Similarly, since $\mathcal{W}$is convex in $\mathbf{\Phi}$ the matrix of general term $\partial^2\mathcal{W}/\partial\Phi_{xy}\partial\Phi_{zt}$ must be semi-definite positive, which implies 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 can be viewed as a feature, or as a bug. Mourifié and Siow (2017) 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.

3 Examples of separable models

While Proposition 2 and Theorem 2 provide a general way of computing surplus and utilities, they can be derived in closed form in important special cases. In all formulæ below, the proportions and masses of single men in feasible matchings are computed as

$$\mu_{0|x} = 1 - \sum_{y\in Y} \mu_{y|x} \quad \text{and} \quad \mu_{x0} = n_x - \sum_{y\in Y} \mu_{xy}, \tag{3.1}$$

and similarly for women. We maintain Assumption 1 in this section.

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

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

3.1 Choo and Siow and its predictions

Our first example is the multinomial logit model used by Choo and Siow, which is a particular case of the results in Section 2 when the $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ distributions are iid standard type I extreme value:[8]

Example 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*

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \log\left(1+\sum_{y\in\mathcal{Y}}\exp(U_{xy})\right) \text{ and } G_x^*(\boldsymbol{\mu}_{\cdot|x}) = \mu_{0|x}\log(\mu_{0|x})+\sum_{y\in\mathcal{Y}}\mu_{y|x}\log\mu_{y|x}. \quad (3.2)$$

where the term $\mu_{0|x}$ is a function of $\boldsymbol{\mu}_{\cdot|x}$ defined in (3.1). Expected utilities are $u_x = -\log\mu_{0|x}$ and $v_y = -\log\mu_{0|y}$. The generalized entropy is

$$\mathcal{E} = -\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}_0}}\mu_{xy}\log\mu_{y|x} - \sum_{\substack{y\in\mathcal{Y}\\ x\in\mathcal{X}_0}}\mu_{xy}\log\mu_{x|y}, \quad (3.3)$$

which is a standard entropy[9]*.*

Surplus and matching patterns are linked by

$$\Phi_{xy} = 2\log\mu_{xy} - \log\mu_{x0} - \log\mu_{0y}, \quad (3.4)$$

which is Choo and Siow (2006)'s identification result, more familiar under the form

$$\mu_{xy} = \sqrt{\mu_{x0}\mu_{0y}}\exp(\Phi_{xy}/2). \quad (3.5)$$

A direct application of Corollary 1 shows that the expected utilities u_x and v_y minimize the globally convex function $F(\boldsymbol{u},\boldsymbol{v};\boldsymbol{\Phi},\boldsymbol{r})$ defined as

$$\sum_{x\in\mathcal{X}} n_x\left(u_x+e^{-u_x}-1\right)+\sum_{y\in\mathcal{Y}} m_y\left(v_y+e^{-v_y}-1\right)+2\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}}\sqrt{n_x m_y}e^{\frac{\Phi_{xy}-u_x-v_y}{2}} \quad (3.6)$$

[8] From now on we deviate from Choo and Siow by centering all type I extreme values distributions we use; the only effect of this normalization is that it eliminates the Euler constant $\gamma = 0.577\ldots$ from expected utilities.

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

and that $\mu_{x0} = n_x \exp(-u_x); \mu_{0y} = \exp(-v_y)$*; and* $\mu_{xy} = \sqrt{n_x m_y} \exp\left((\Phi_{xy} - u_x - v_y)/2\right)$*. See Appendix C.1.1 for details.*

Choo and Siow's assumption that errors are iid draws from a standard type-I extreme value generates a very convenient multinomial logit form; but it has severe drawbacks, as it has both counterintuitive aspects and counterfactual implications. We show in Galichon and Salanié (2019) how classical paradoxes of multinomial logit extend to matching markets[10]. The Choo and Siow model has other stark comparative statics predictions. Since $u_x = -\log(\mu_{x0}/n_x)$ in the Choo and Siow framework, expected utilities are in a one-to-one relationship with the probabilities of singlehood. Property (2.7) becomes a statement on semi-elasticities of these probabilities. Moreover, the equilibrium equation (3.5) 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[11].

3.2 Beyond Logit

In separable models, substitution patterns are driven both by the unobserved and by the observed heterogeneity; either one may dominate. To see this, fix the variances of the unobserved heterogeneity terms and multiply $\mathbf{\Phi}$ by a large positive constant. Then unobserved

[10] See B.2 for a summary of the argument.

[11] Mourifié and Siow (2017) and Mourifié (2019) extend this and other results of Graham (2013) to models with peer effects.

heterogeneity becomes almost irrelevant, and substitution patterns will reflect the properties of the function $\mathbf{\Phi}$. If on the other hand we multiply $\mathbf{\Phi}$ by a small positive constant, then matching will be dominated by unobserved heterogeneity. In the multinomial logit model, the resulting substitution patterns will be just as constrained as in the one-sided case. As is well-known, the price cross-elasticities in single-agent multinomial logit models are highly restricted—to such an extent that much of the applied literature now uses mixed multinomial logit instead. This should likewise inspire caution with respect to the multilogit matching model, as well as a quest for alternative specifications.

We illustrate this in the next subsections with two families of separable models for which some formulæ can be obtained in closed-form. The class of separable models is much larger, however; it contains models whose error terms have discrete support, as well as models whose G and H functions can only be evaluated via simulations. We elaborate on these two issues in section 5.

3.2.1 Generalized extreme value

The need to go beyond the logit framework has long been recognized in the literature on industrial organization and on consumer demand. This has led to a huge literature on random utility models, initiated by McFadden's seminal work on generalized extreme value (GEV) theory (McFadden (1978); see also Anderson, de Palma, and Thisse (1992) for an exposition and applications.) We give general formulæ for GEV models in Appendix C.1.

The multinomial logit Choo and Siow model is the simplest example of the GEV framework. A direct extension is the heteroskedastic model considered by Chiappori, Salanié, and Weiss (2017); it allows the scale parameters of the type I extreme value distributions to vary across genders or groups. We estimate it in section 6.

The well-known *nested logit model* is another member of the GEV family; we study it as Example 3 in Appendix C. We found the flexible coefficients multinomial logit (FC-MNL) models of Davis and Schiraldi (2014) more useful for our purposes, as they easily incorporate local correlation patterns. We study them as Example 5 in Appendix C and we present

estimates on the Choo and Siow (2006) dataset in section 6.

3.2.2 Mixed Logit and Pure Characteristics

While the GEV framework is convenient, it is common in the applied literature to allow for random variation in preferences over observed characteristics of products. The modern approach to empirical industrial organization, for instance, allows different buyers to have idiosyncratic preferences over observed characteristics of products[12]. We describe a *mixed logit model* for matching as Example 4 in Appendix C.

Closer to our framework, hedonic models also build on idiosyncratic preferences for observed characteristics, on both sides of a match[13]. Our setup allows for such specifications. We can for instance extend the "pure characteristics" specification of Berry and Pakes Berry and Pakes (2007) in a matching context. Assume that men of group x care for a vector of observed characteristics of partners $\boldsymbol{\zeta}_x(y)$, but the intensity of the preferences of each man i in the group depends on a vector $\boldsymbol{\varepsilon}_i$ which is drawn from some given distribution. Then we could take $\boldsymbol{P}_x$ to be the joint distribution of the vector $(\boldsymbol{\zeta}_x(y) \cdot \boldsymbol{\varepsilon}_i)_{y \in \mathcal{Y}}$. We investigate a particular case of this specification in the next example: the Random Scalar Coefficient (RSC) model, where the common dimension of $\boldsymbol{\zeta}_x(y)$ and $\boldsymbol{\varepsilon}_i$ is one.

Example 2 (Random Scalar Coefficient (RSC) models)**.** *Assume that for each man i in group x, $\varepsilon_{iy} = \varepsilon_i \times \zeta_x(y)$, where $\zeta_x(y)$ is a scalar index of the observable characteristics of women which is the same for all men in the same group x, and the ε_i's are iid random variables which are assumed to be continuously distributed according to a c.d.f. F_x. We provide formulas for this model in Appendix C.2.*

The RSC model is related to the hedonic specifications investigated in Ekeland, Heckman, and Nesheim (2004) and Heckman, Matzkin, and Nesheim (2010). In these two papers, y is a continuous and scalar quality parameter, and the heterogeneity is multiplicative: $\varepsilon_{iy} = y\varepsilon_i$. Ekeland, Heckman, and Nesheim (2004) assume that Φ_{xy} is additively

[12] See the literature surveyed in Ackerberg, Benkard, Berry, and Pakes (2007) or Reiss and Wolak (2007).
[13] See Ekeland, Heckman, and Nesheim (2004) and Heckman, Matzkin, and Nesheim (2010).

separable. This allows them to identify the distribution of ε_i, which is assumed to be known in our setting. In contrast, Heckman, Matzkin, and Nesheim (2010) assume, like us, that the distribution of ε_i is known. To obtain identification they use a quantile transformation, which does not extend to the discrete case our paper investigates.

4 Parametric Inference

First assume that all observations concern a single matching market. While the formula in Theorem 2 (i) gives a straightforward estimator of the systematic surplus function $\mathbf{\Phi}$, with multiple payoff-relevant observed characteristics x and y it is bound to be very unreliable. 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 $\Phi_{xy}^{\boldsymbol{\lambda}}$ and distributions $\boldsymbol{P}_x^{\boldsymbol{\lambda}}$ and $\boldsymbol{Q}_y^{\boldsymbol{\lambda}}$ 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 $\hat{S} = \sum_x \widehat{N}_x + \sum_y \widehat{M}_y$ individuals, where $\widehat{N}_x$ (resp. $\widehat{M}_y$) 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 a particular type in 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

$$\begin{cases} \sum_{y\in\mathcal{Y}} \mu^{\boldsymbol{\lambda}}_{xy} + \mu^{\boldsymbol{\lambda}}_{x0} = \hat{n}_x \\ \sum_{x\in\mathcal{X}} \mu^{\boldsymbol{\lambda}}_{xy} + \mu^{\boldsymbol{\lambda}}_{0y} = \hat{m}_y \end{cases} \tag{4.1}$$

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 optimal matching for parameters $\boldsymbol{\lambda}$ and margins $\hat{\boldsymbol{r}}$. We will show in Section 5 how it can be computed, often very efficiently. For now we focus on statistical inference on $\boldsymbol{\lambda}$. We propose two methods: a very general Maximum Likelihood method, and a simpler moment-based method that gives consistent estimators for an important subclass of models.

4.1 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 (4.1). 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[14]. It is easy to see that the log-likelihood of this sample can be written as

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

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.

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

4.2 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 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 $\mathbf{\Phi}$ matrix must be linear in the parameter vector:

$$\Phi_{xy}^{\boldsymbol{\lambda}} = \sum_{k=1}^{K} \lambda_k \phi_{xy}^k \tag{4.2}$$

where the parameter $\boldsymbol{\lambda} \in \mathbb{R}^K$ and $\tilde{\boldsymbol{\phi}} := (\boldsymbol{\phi^1}, \ldots, \boldsymbol{\phi^K})$ are K known linearly independent *basis surplus vectors*. If the number of basis surplus vectors is rich enough, this can generate any surplus function. We compute the (joint) moments of any feasible matching $\boldsymbol{\mu}$ as the average values of the basis surplus vectors:

$$C^k(\boldsymbol{\mu}) = \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \mu_{xy} \phi_{xy}^k.$$

In particular, the empirical moments are associated with the observed matching $\hat{\boldsymbol{\mu}}$. 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

$$C^k(\hat{\boldsymbol{\mu}}) = C^k(\boldsymbol{\mu}^{\boldsymbol{\lambda}}) \text{ for all } k. \tag{4.3}$$

Then the moment-matching estimator is

$$\hat{\boldsymbol{\lambda}}^{MM} := \arg\max_{\boldsymbol{\lambda} \in \mathbb{R}^k} \left(\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \hat{\mu}_{xy} \Phi_{xy}^{\boldsymbol{\lambda}} - \mathcal{W}\left(\mathbf{\Phi}^{\boldsymbol{\lambda}}, \hat{\boldsymbol{r}}\right) \right). \tag{4.4}$$

Note that the objective function in this program is concave, as $\mathcal{W}$ is convex in $\mathbf{\Phi}$ and $\mathbf{\Phi}^{\boldsymbol{\lambda}}$ is linear in $\boldsymbol{\lambda}$. We now show that these two definitions are equivalent:

Theorem 3. *Under Assumptions 1 and 2, assume that* $\mathbf{\Phi}^{\boldsymbol{\lambda}}$ *is linear in* $\boldsymbol{\lambda}$ *and that the distributions of the unobserved heterogeneity terms* $\boldsymbol{P}_x$ *and* $\boldsymbol{Q}_y$ *are known. Then:*

(i) The moment-matching estimator defined by (4.4) makes predicted moments equal to observed moments: $C^k(\hat{\boldsymbol{\mu}}) = C^k(\mu^{\boldsymbol{\lambda}})$ *for all* k *when* $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^{MM}$.

(ii) The moment-matching estimator $\hat{\boldsymbol{\lambda}}^{MM}$ *is also the vector of Lagrange multipliers of the moment constraints in the program*

$$\mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right) = \max_{\boldsymbol{\mu} \in \mathcal{M}(\hat{\boldsymbol{r}})} \left(\mathcal{E}\left(\boldsymbol{\mu}, \hat{\boldsymbol{r}}\right) : C^k(\boldsymbol{\mu}) = C^k(\hat{\boldsymbol{\mu}}) \; \forall k \right), \tag{4.5}$$

and $\mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right) = \mathcal{E}\left(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}^{MM}}, \hat{\boldsymbol{r}}\right)$.

(iii) The program (4.5) has the dual formulation:

$$\begin{aligned} \mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right) = \quad & \min_{\substack{U,V \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}} \\ \boldsymbol{\lambda} \in \mathbb{R}^k}} \left(G(\boldsymbol{U}, \hat{\boldsymbol{n}}) + H(\boldsymbol{V}, \hat{\boldsymbol{m}}) - \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \hat{\mu}_{xy} \Phi^{\boldsymbol{\lambda}}_{xy} \right) \\ s.t. \quad & U_{xy} + V_{xy} \geq \Phi^{\boldsymbol{\lambda}}_{xy} \; \forall x \in \mathcal{X}, y \in \mathcal{Y}. \end{aligned} \tag{4.6}$$

The intuition for (i) is simple: note that all three programs (4.4), (4.5) and (4.6) are globally convex optimization problems, and hence very easy to solve numerically. we know from (1.13) that if $\boldsymbol{\mu}$ is optimal for $\mathbf{\Phi}$, then $\partial \mathcal{W} / \partial \mathbf{\Phi} = \boldsymbol{\mu}$, and the first-order condition in (4.4) is simply the equality between predicted and observed moments, that is equality (4.3).

Part (ii) of Theorem 3 generates a very simple specification test. Compare the *actual* value $\mathcal{E}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right)$ of the generalized entropy associated to the empirical distribution to the value $\mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right)$ of the program (4.6). By definition, $\mathcal{E}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right) \leq \mathcal{E}_{\max}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}\right)$; moreover, these two values coincide if and only if the model is well-specified. We state this in the following proposition:[15]

[15] The critical values of the test 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 $\boldsymbol{Q}_y$.

Proposition 4. *(**A Specification Test**) Under Assumptions 1 and 2, assume that the distributions of the unobserved heterogeneity terms $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ are known. Then $\mathcal{E}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}) \leq \mathcal{E}_{\max}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}})$, with equality if and only if there is a value $\boldsymbol{\lambda}$ of the parameter such that $\boldsymbol{\Phi}^{\boldsymbol{\lambda}} = \boldsymbol{\Phi}$.*

In the Choo and Siow case, program (4.4) has a particularly simple form.

Proposition 5 (Estimation in the Logit Case)**.** *When ε and η both have Gumbel distributions with scaling parameter one, the moment matching estimator $\hat{\boldsymbol{\lambda}}$ solves*

$$\min_{\boldsymbol{u},\boldsymbol{v},\boldsymbol{\lambda}} F(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\Phi}^{\boldsymbol{\lambda}}, \hat{\boldsymbol{r}}) - \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \hat{\mu}_{xy} \Phi^{\boldsymbol{\lambda}}_{xy}$$

where F is defined in (3.6).

Proposition 5 is a direct consequence of formula (3.6), as the objective function is smooth and globally convex in its $|\mathcal{X}| + |\mathcal{Y}| + K$ arguments. It is therefore easy to minimize. In addition to the estimator of $\boldsymbol{\lambda}$, it also yields estimators of the expected utilities $\boldsymbol{u}$ and $\boldsymbol{v}$ as a by-product.

4.3 Parameterization, testing, and multimarket data

Proposition 3 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 $\boldsymbol{\Phi}$ 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 (2.6) 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[16]. As another example, consider a semiparametric specification in the spirit of Ekeland, Heckman, and Nesheim (2004): $\Phi_{xy} = b_y'\phi_x$, with known d-dimensional vectors ϕ_x and unknown vectors b_y. If $d < |Y|$, this would restrict the number of degrees of freedom in $\boldsymbol{\Phi}$, 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 $\boldsymbol{\Phi}$ 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.

5 Computation

Estimating the model requires confronting two layers of computational challenges. The first layer consists of computing the optimal matching $\boldsymbol{\mu}$ for a given surplus function. Fortunately, the properties of the problem make it much less costly than one would naturally expect. First, they allow the use of straightforward descent methods that work very well in such globally convex problems. Second, in models when the unobserved heterogeneity has a discrete support (or when it has been discretized), computing the equilibrium matching becomes a finite-dimensional linear programming problem. Third, when the unobserved heterogeneity belongs to the logit family (of which the Choo and Siow multinomial logit model is the simplest instance), a lower-dimensional alternative based on the Iterative Projection Fitting Procedure (IPFP) proves to be particularly efficient. We present these three

[16] 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.

methods here and we test their performance in Appendix D. The second layer consists of estimating the parameters of the surplus function. Each of the three methods can be nested inside an estimation algorithm. We show that in addition, when the surplus function takes the linear form described in section 4.2, one can extend the min-Emax and linear programming methods and bypass this inner loop to estimate the parameters by moment-matching.

We present two methods to compute the equilibrium: Min-Emax (gradient descent), and IPFP (coordinate descent. In appendix D, we present a third one, linear programming based on simulated draws.

5.1 Min-Emax method

Theorem 2 gave two expressions for the social surplus. Program (1.13) solves for the equilibrium matching patterns $\boldsymbol{\mu}$. Alternatively, program (1.14) 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 3, the constraints $\boldsymbol{U} + \boldsymbol{V} \geq \boldsymbol{\Phi}$ in the second program are all saturated at the optimum. Therefore by theorem 2, we can obtain the equilibrium matching patterns by solving the globally concave unconstrained maximization problem (1.13), and we can obtain the U and V equilibrium utility components by solving its dual, the the globally convex unconstrained minimization problem

$$\min_{\boldsymbol{U} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}} \left(G(\boldsymbol{U}, \boldsymbol{n}) + H(\boldsymbol{\Phi} - \boldsymbol{U}, \boldsymbol{m}) \right). \tag{5.1}$$

Since $G = \sum n_x G_x$, where G_x is the expected value of the maximum utility of men of group x, we call the method based on (5.1) the *min-Emax* method. Problem (5.1) 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 how to use simulation and linear programming to approximate the problem. As (5.1)

is globally convex, a descent algorithm converges nicely under weak conditions; 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 (5.1) therefore replicates a Walrasian tâtonnement process; we don't need to be concerned about its convergence since global convexity guarantees it.

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 (3.6) 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 D.

5.2 IPFP

In some applications, the number of groups $|\mathcal{X}|$ and $|\mathcal{Y}|$ can be large and solving for equilibrium by minimizing (5.1) 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. Recall from Corollary 1 that the social welfare is the minimum of $S(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Phi}, \boldsymbol{r})$ defined in (1.16), and that it is achieved at the equilibrium expected utilities $\boldsymbol{u}, \boldsymbol{v}$. Denote $\boldsymbol{\mu}(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Phi}, \boldsymbol{r})$ the matching that maximizes $S(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Phi}, \boldsymbol{r})$. The optimality conditions with respect to u_x and v_y respectively are simply the adding-up constraints:

$$n_x = \sum_{y \in \mathcal{Y}} \mu_{xy}(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\Phi}, \boldsymbol{r}) + \mu_{x0}(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\Phi}, \boldsymbol{r}) \tag{5.2}$$

$$m_y = \sum_{x \in \mathcal{Y}} \mu_{xy}(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\Phi}, \boldsymbol{r}) + \mu_{0y}(\boldsymbol{u}, \boldsymbol{v}; \boldsymbol{\Phi}, \boldsymbol{r}). \tag{5.3}$$

We want to find the $\boldsymbol{u}$ and $\boldsymbol{v}$ that solve these equilibrium equations for given surplus $\boldsymbol{\Phi}$ and margins $\boldsymbol{r}$. To do so, we will adjust the prices alternatively on each side of this market. First we fix the prices (v_y) and we find the prices (u_x) such that the demands clear the markets for inputs $x \in \mathcal{X}$, that is, the system of equations (5.2) is satisfied. Then we fix these new prices (u_x) and we find the prices (v_y) such that the demands clear the markets for inputs $y \in \mathcal{Y}$, so that the system (5.3) is satisfied; 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[17]. We describe the algorithm in full detail in Appendix D, and we prove its convergence there.

Theorem 4. *Under Assumptions 1 and 2, the IPFP algorithm converges to the solution* $\boldsymbol{\mu}$ *of (1.13) and to the corresponding expected utilities.*

Take the multinomial logit Choo-Siow model of Example 1 for instance. 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 (3.4) can be rewritten as

$$\mu_{xy} = \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}\mu_{0y}}. \tag{5.4}$$

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:

$$\mu_{x0} + \left(\sum_{y\in\mathcal{Y}} \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{0y}}\right)\sqrt{\mu_{x0}} = n_x \tag{5.5}$$

$$\mu_{0y} + \left(\sum_{x\in\mathcal{X}} \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}}\right)\sqrt{\mu_{0y}} = m_y. \tag{5.6}$$

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 (5.5) iteratively. Starting from an arbitrary guess $\mu_{0y}^{(0)}$, at step $(2k+1)$ we find the following updating equation

$$\begin{cases} \mu_{x0}^{(2k+1)} = \left(\sqrt{n_x + \frac{A_x^2}{4}} - \frac{A_x}{2}\right)^2 \text{ with } A_x = \sum_{y\in\mathcal{Y}} \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{0y}^{(2k)}} \\ \mu_{0y}^{(2k+2)} = \left(\sqrt{m_y + \frac{B_y^2}{4}} - \frac{B_y}{2}\right)^2 \text{ with } B_y = \sum_{x\in\mathcal{X}} \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}^{(2k+1)}} \end{cases} \tag{5.7}$$

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

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

u_x and v_y. We also illustrate the algorithm for the nested logit model in Example 3 in Appendix C.1.2. We tested the performance of our proposed algorithms on an instance of the Choo and Siow model and we report the results in Appendix D. The IPFP algorithm is extremely fast compared to standard optimization or equation-solving methods. The min-Emax method of (5.1) is slower but still works very well for medium-size problems, and it is applicable to all separable models.

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 E. 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[18]. We use the "non-reform states" subsample, which has 224,068 observations representing 13.3m individuals.

Our Proposition 3 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 a more parsimonious model of surplus to the data,

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

using the semilinear model described in 4.2. This allows us to select a set of basis functions (ϕ_{xy}^k). We then fit alternative specifications to the data, using this set of basis functions and different distributions for the error terms.

6.1 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 $\phi_{xy}^1 \equiv 1$ and $\phi_{xy}^2 = D_{xy} \equiv \mathbb{1}(x \geq y)$, 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 1 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 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 $1 \leq n \leq 4$.

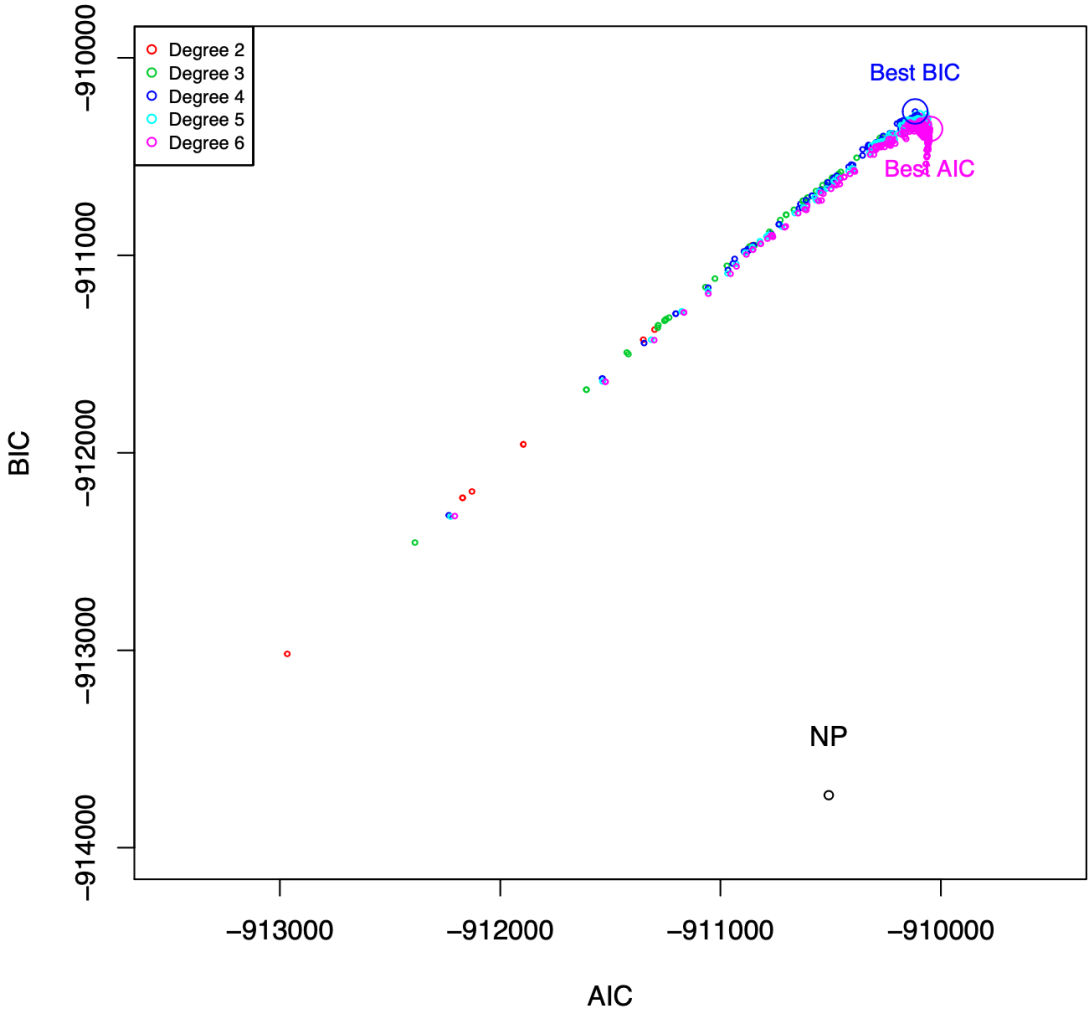


Figure 1: AIC and BIC Values for the Parametric and Nonparametric Choo and Siow Models

6.2 The Homoskedastic Logit Model

Table 4 in Appendix F 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 Proposition 4. 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. 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.

6.3 Beyond the Homoskedasticity: Heteroskedastic Logit Models

We now move to 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 section[19].

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[20]. As reported in Appendix F, adding heteroskedasticity across genders barely improves the fir, and deteriorates the BIC. 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 $\varepsilon_{iy} + \eta_{jx}$ with $\sigma_x \varepsilon_{iy} + \tau_y \eta_{jx}$, 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. Here again the entropy test does not reject the null of correct specification.

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[21], from 0.28 at age 18 to 0.72 at age 25. This is a large relative variation. It impacts directly 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 expected share of the

[19] 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.

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

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

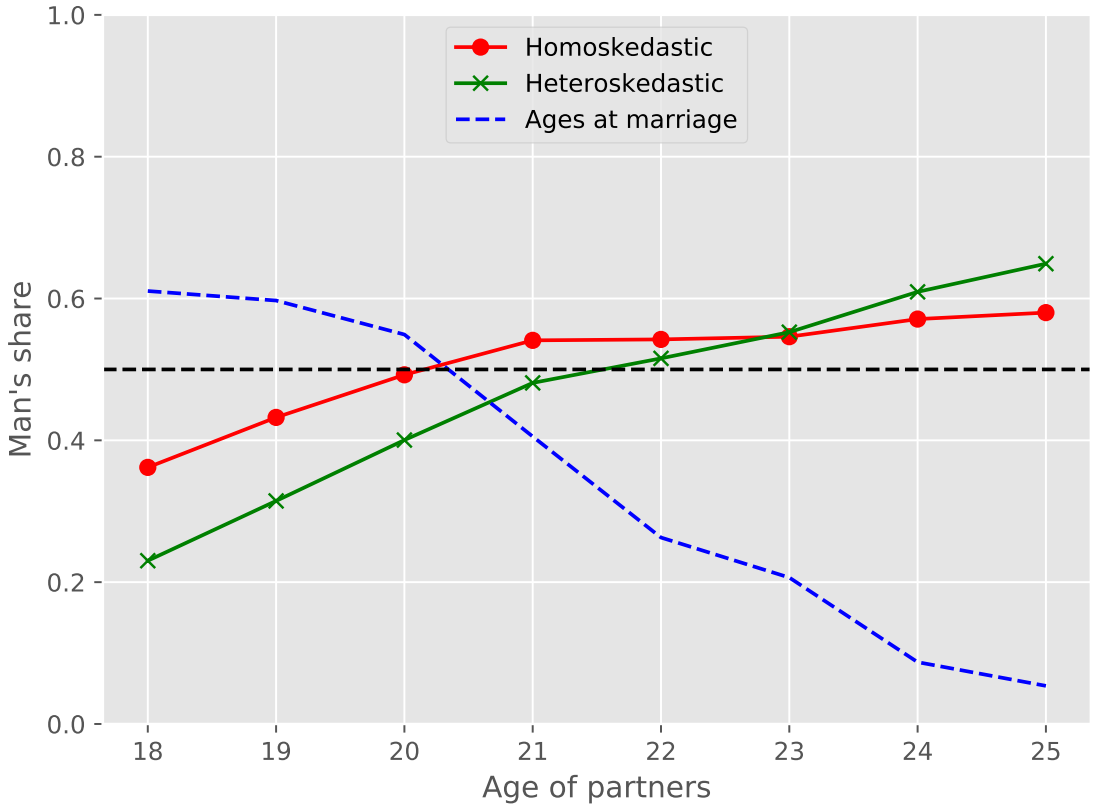


Figure 2: Men's Share of the Marriage Surplus in the Logit Models

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

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 2 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 expected utilities of men and women in marriage.

6.4 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[22]. 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)[23]. This specification belongs to the class of Generalized Extreme Values models that we discussed in Appendix C.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 give 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[24]. Still, they do suggest more subtle patterns of substitution between partners than the Choo and Siow model allows for. We illustrate this on Figures 4 and 5. Figure 4 for instance plots the "demand semi-elasticities": $\partial \log \mu_{t|x}/\partial V_y$ for men whose age x goes

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

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

[24] See Appendix F.

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:

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

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 4. 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 5 tells a similar story for women.

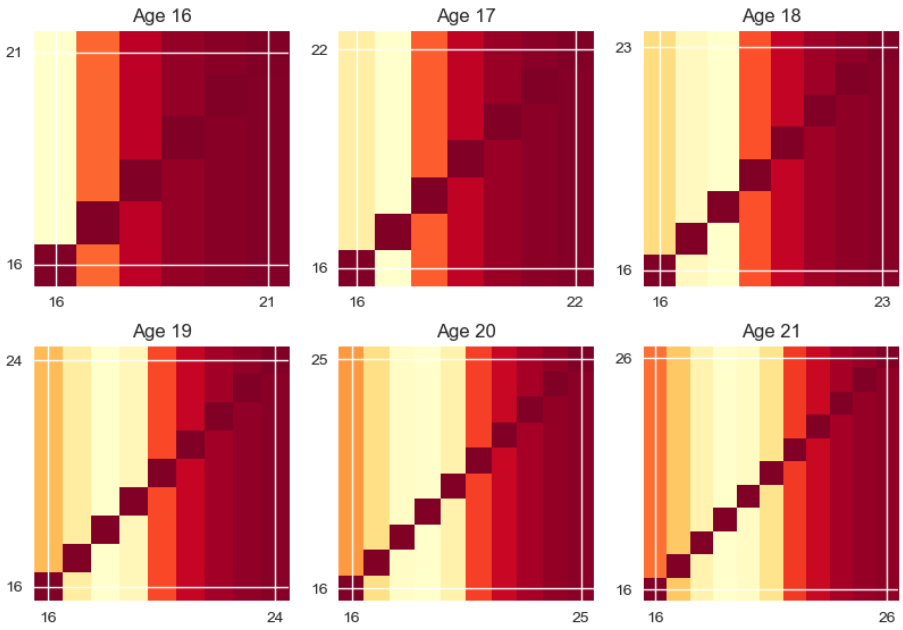
Concluding Remarks

We have left several interesting theoretical issues for future research. One such issue is the behavior of the finite population approximation of the model. We have worked in an idealized model with an infinite number of agents within each observable group. It is easy to adapt the proofs in Appendix A by replacing the distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ with the empirical distributions, and the expectation operators with the population means. E.g. $G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot})$ becomes

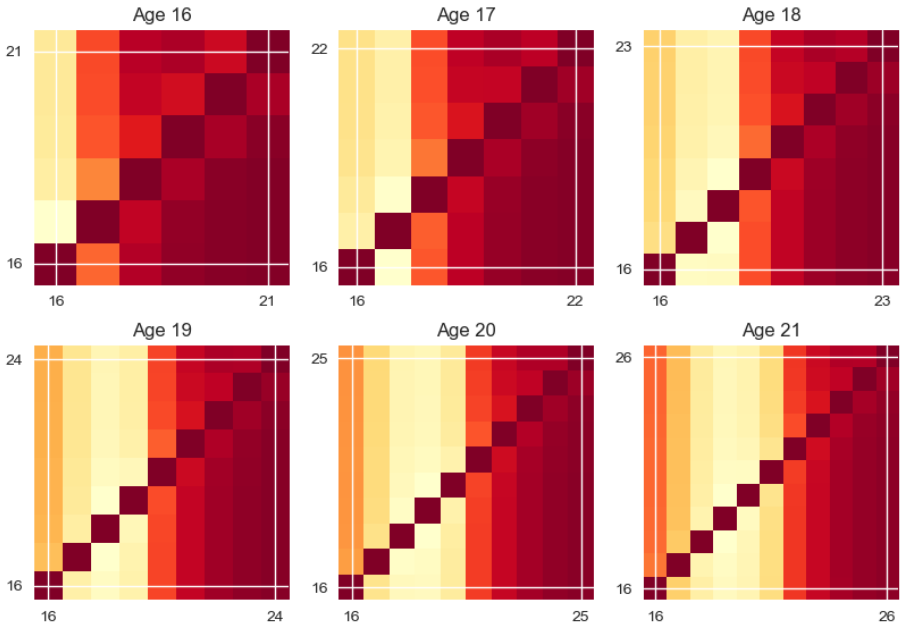
$$\frac{1}{n_x} \sum_{i:x_i=x} \max_{y \in \mathcal{Y}_0} \{U_{xy} + \varepsilon_{iy}\}.$$

This is still a convex function, but it has kinks; as a consequence, the derivatives become subgradients in items (i) and (ii) of Proposition 2, and the utilities are only partially identified from the matching patterns in Proposition 3. It is natural to expect that the estimated regions for the parameters shrink to their large population analogs. This goes beyond the scope of the present paper and is left as a conjecture. Likewise, we leave the characterization of the rate of convergence for future research.

Several assumptions made in our paper are tested on simulations by Chiappori, Nguyen, and Salanié (2019). They seek to quantify the bias induced by assuming separability in a model which is not by adding i.i.d. shocks in a Choo and Siow model, and find that the

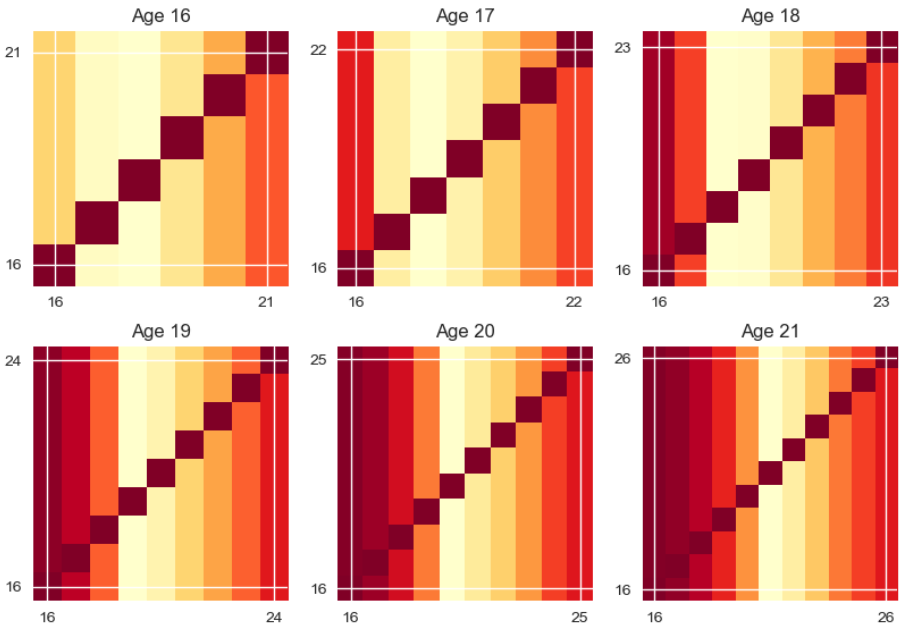


(a) Choo-Siow

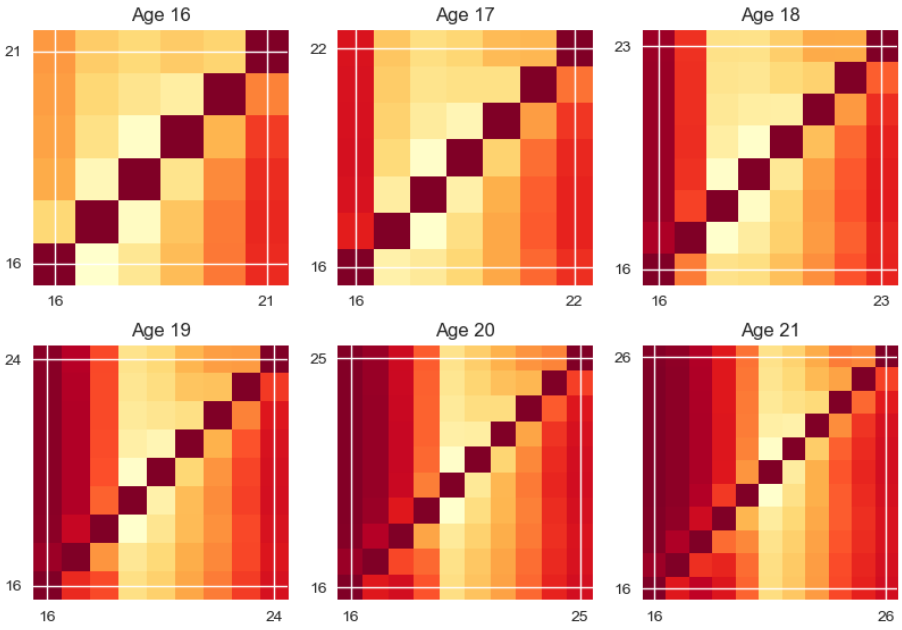


(b) FC-MNL

Figure 4: Semi-elasticities of substitution across partners: men



(a) Choo-Siow



(b) FC-MNL

Figure 5: Semi-elasticities of substitution across partners: women

resulting estimates of the surplus matrix are biased upwards, but the estimates of the surplus *complementarities* have very little bias, when one adds a large nonseparable heterogeneity. They also run simulations to assess the large market assumptions which suggest that for the Choo and Siow specification at least, the large markets assumption can be used safely even with populations of a few hundred individuals. We find these simulation results reassuring about the assumptions we have maintained in the present paper.

Finally, let us mention some extensions. On the methodological front, 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). Also, while the present paper operates under the maintained assumption that utility is fully transferable without frictions, this assumption can be relaxed; Galichon, Kominers, and Weber (2019) study models with imperfectly transferable utility and separable logit heterogeneity, while Galichon and Hsieh (2019) look at models with nontransferable utility and a similar form of heterogeneity.

References

Ackerberg, D., C. L. Benkard, S. Berry, and A. Pakes (2007): "Econometric Tools for Analyzing Market Outcomes," in *Handbook of Econometrics, vol. 6A*, ed. by J.-J. Heckman, and E. Leamer. North Holland.

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*, Forthcoming.

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

——— (1992): *Discrete Choice Theory of Product Differentiation*. MIT Press.

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., 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.

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 (2019): "Monge-Kantorovich Depth, Quantiles, Ranks and Signs," *Annals of Statistics*.

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), TBD.

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 (2019): "Like Attract Like: A Structural Comparison of Homogamy Across Same-Sex and Different-Sex Households," *Journal of Political Economy*, Forthcoming.

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.

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*, Forthcoming.

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): "IIA in Separable Matching Markets," Columbia University mimeo.

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.

Heckman, J.-J., R. Matzkin, and L. Nesheim (2010): "Nonparametric Identification and Estimation of Nonadditive Hedonic Models," *Econometrica*, 78, 1569–1591.

Hiriart-Urruty, J.-B., and C. Lemaréchal (2001): *Fundamental 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 (2017): "The Cobb Douglas Marriage Matching function: Marriage Matching with Peer and Scale Effects," University of Toronto mimeo.

Reiss, P., and F. Wolak (2007): "Structural Econometric Modeling: Rationales and Examples from Industrial Organization," in *Handbook of Econometrics, vol. 6A*, ed. by J.-J. Heckman, and E. Leamer. North Holland.

Rockafellar, R. T. (1970): *Convex Analysis*. Princeton University Press.

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.

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

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 Theorem 1

Replace the expression of G_x (1.3) in the formula for G_x^* (1.6) to obtain

$$-G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = \inf_{\tilde{\boldsymbol{U}}_{\boldsymbol{x}\cdot}} \left(-\sum_{y\in\mathcal{Y}_0} \mu_{y|x}\tilde{U}_{xy} + \boldsymbol{E}_{\boldsymbol{P}_x} \max_{y\in\mathcal{Y}_0} \left(\varepsilon_{iy} + \tilde{U}_{xy}\right) \right)$$

where the minimization is over $\tilde{\boldsymbol{U}}_{\boldsymbol{x}\cdot}$ such that $\tilde{U}_{x0} = 0$. The first term in the minimand can be seen as the expectation of the random variable $-\tilde{U}_{xY}$ under the distribution $Y \sim \mu_{Y|X=x}$. The second term can be rewritten as the expectation $\boldsymbol{E}_{\boldsymbol{P}_x}\tilde{u}_x(\boldsymbol{\varepsilon}_{\boldsymbol{i}\cdot})$ where $\boldsymbol{\varepsilon}_{\boldsymbol{i}}$ is drawn from the distribution $\boldsymbol{P}_x$ under the constraint that $\tilde{u}_x(\boldsymbol{\varepsilon}_{\boldsymbol{i}}) - \tilde{U}_{xy} \geq \varepsilon_{iy}$ for all y, with equality for at least one y. Therefore, setting $V_{xy} = -U_{xy}$, one has

$$\begin{aligned} -G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) &= \inf_{V_{x\cdot},\tilde{u}_x} \left(\boldsymbol{E}_{\mu_{Y|X=x}} V_{xY} + \boldsymbol{E}_{\boldsymbol{P}_x}\tilde{u}_x(\boldsymbol{\varepsilon}_{\boldsymbol{i}\cdot}) \right) \\ \text{s.t. } V_{x0} &= 0 \text{ and } \tilde{u}_x(\boldsymbol{\varepsilon}_{\boldsymbol{i}\cdot}) + V_{xy} \geq \varepsilon_{iy} \quad \forall \boldsymbol{\varepsilon}_{\boldsymbol{i}\cdot}, \forall y \in \mathcal{Y}_0 \end{aligned}$$

where the argument $\boldsymbol{\varepsilon}_{\boldsymbol{i}\cdot}$ ranges over all values in the support of $\boldsymbol{P}_x$. We recognize the value of the dual of a matching problem in which the margins are $\mu_{Y|x=x}$ and $\boldsymbol{P}_x$ and the surplus is ε_{iy}. By the equivalence of the primal and the dual, this yields expression (1.9).

A.2 Proof of Theorem 2

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

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

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

(i) 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

$$\begin{aligned} \mathcal{W} = \inf_{\tilde{u},\tilde{v}} \quad & \left(\int \tilde{u}(x,\varepsilon)\, d\tilde{n}(x,\varepsilon) + \int \tilde{v}(y,\eta)\, d\tilde{m}(y,\eta) \right) \\ \text{s.t.} \quad & \tilde{u}(x,\varepsilon) + \tilde{v}(y,\eta) \geq \Phi_{xy} + \varepsilon_y + \eta_x \ \ \forall (x,y,\varepsilon,\eta) \\ & \tilde{u}(x,\varepsilon) \geq \varepsilon_0 \ \ \forall (x,\varepsilon) \\ & \tilde{v}(y,\eta) \geq \eta_0 \ \ \forall (y,\eta). \end{aligned} \tag{A.1}$$

Fix any $\tilde{u}, \tilde{v}$ that satisfies all constraints in this program. For $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, define

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

Also define $U_{x0} = V_{0y} = 0$. Then $\tilde{u}(x,\varepsilon) \geq \max_{y \in \mathcal{Y}_0} \{U_{xy} + \varepsilon_y\}$ and $\tilde{v}(y,\eta) \geq \max_{x \in \mathcal{X}_0} \{V_{xy} + \eta_x\}$; and the first constraint in (A.1) 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

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

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

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

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,

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

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

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

which is expression (1.14). Introducing multipliers (μ_{xy}), this convex minimization problem can be expressed 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 (1.13):

$$\mathcal{W}(\boldsymbol{\Phi},\boldsymbol{r}) = \max_{\boldsymbol{\mu}\geq 0} \left(\sum_{xy} \mu_{xy}\Phi_{xy} - G^{*}\left(\boldsymbol{\mu},\boldsymbol{n}\right) - H^{*}\left(\boldsymbol{\mu},\boldsymbol{m}\right) \right).$$

A.3 Proof of Corollary 1

Recall from equation (1.13) that the equilibrium matching $\boldsymbol{\mu}$ maximizes $\sum_{x,y} \mu_{xy}\Phi_{xy} + \mathcal{E}(\boldsymbol{\mu},\boldsymbol{r})$ over μ in $\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$. While $\mathcal{E}$ is well-defined and concave, it is only strictly concave when $\boldsymbol{\mu}$ has the margins $\boldsymbol{r}$ (otherwise $\mathcal{E}$ is infinite). Take any strictly concave function $E(\boldsymbol{\mu};\boldsymbol{r})$ that extends $\mathcal{E}$, in the sense that $E(\boldsymbol{\mu};\boldsymbol{r}) = \mathcal{E}(\boldsymbol{\mu},\boldsymbol{r})$ whenever $\boldsymbol{\mu}$ has margins $\boldsymbol{n}$ and $\boldsymbol{m}$. There are many ways of doing it; indeed 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

$$K\left(\boldsymbol{\mu};\boldsymbol{r}\right) = \sum_{x} \left\{ A_x\left(\sum_{y} \mu_{xy} + \mu_{x0} \right) - A_x\left(n_x\right) \right\} + \sum_{y} \left\{ B_y\left(\sum_{x} \mu_{xy} + \mu_{0y} \right) - B_y\left(m_y\right) \right\}, \tag{A.2}$$

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 $\boldsymbol{\mu}$.

The program (1.13) can be rewritten as

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \sum_{x,y\in\mathcal{X}\times\mathcal{Y}} \mu_{xy}\Phi_{xy} + E(\boldsymbol{\mu};\boldsymbol{r}) \\ s.t. \quad & \mu_{x0} + \sum_{y\in\mathcal{Y}} \mu_{xy} = n_x \\ & \mu_{0y} + \sum_{x\in\mathcal{X}} \mu_{xy} = m_y. \end{aligned} \tag{A.3}$$

Denote a_x and b_y the multipliers of the constraints. The Lagrangian of (A.3) can be written as

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

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 1, this gives $a_x = u_x$; and the μ's are the corresponding matching patterns.

A.4 Proof of Corollary 2

The convexity of $\mathcal{W}$ w.r.t. Φ follows immediately from (1.13); the concavity of $\mathcal{W}$ w.r.t. $(\boldsymbol{r})$ similarly follows from (1.14). 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 $\boldsymbol{r} = (\boldsymbol{n},\boldsymbol{m})$.

A.5 Proof of Proposition 1

We proved part (i) and $U_{xy} + V_{xy} = \Phi_{xy}$ when proving Theorem 2. 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 (1.4). Part (iii) is similar.

A.6 Proof of Proposition 2

Given that $\boldsymbol{P}_x$ has full support, for all $\boldsymbol{U}_{x\cdot}$ and for each y the probability that $U_{xy}+\varepsilon_y$ reaches the maximum in the definition of $G_x(\boldsymbol{U}_{x\cdot})$ is positive. But by the Daly-Zachary-Williams theorem, this probability is the derivative of G_x wrt U_{xy}, and it equals $\mu_{y|x}$, proving (i).

Since $\mu_{0|x} > 0$, we have $\sum_{y\in\mathcal{Y}} \mu_{y|x} < 1$ and we can neglect the feasibility constraints in (1.6). Applying the envelope theorem gives us $U_{xy} = \left(\partial G_x^*/\partial \mu_{y|x}\right)(\boldsymbol{\mu}_{\cdot|x})$ for the value of U_{xy} that achieves the optimum, which proves (ii). Point (iii) follows by the fact that $U_{xy} = \left(\partial G_x^*/\partial \mu_{y|x}\right)(\boldsymbol{\mu}_{\cdot|x})$ is equivalent to $\boldsymbol{U}_{x\cdot}$ being the minimizer of the strictly convex function $\tilde{\boldsymbol{U}}_{x\cdot} \longmapsto G_x\left(\tilde{\boldsymbol{U}}_{x\cdot}\right) - \sum_{y\in\mathcal{Y}} \mu_{y|x}\tilde{U}_{xy}$, that is

$$\min_{\tilde{\boldsymbol{U}}_{x\cdot}:\tilde{U}_{x0}=0} \int \max_{y\in\mathcal{Y}_0} \left\{\tilde{U}_{xy} + \varepsilon_y\right\} d\boldsymbol{P}_x(\varepsilon) - \sum_{y\in\mathcal{Y}} \mu_{y|x}\tilde{U}_{xy}$$

which gives (2.3), QED.

A.7 Proof of Proposition 3

Part (i) follows from Proposition 2 (ii). And since the $\boldsymbol{\mu}$'s are the multipliers in (1.14) and they are all positive, the constraints must be saturated, proving (ii).

A.8 Proof of Theorem 3

(i) The moment-matching estimator $\hat{\boldsymbol{\lambda}}$ solves (4.4), so that $\sum_{x,y} \hat{\mu}_{xy}\phi_{xy}^k = \left(\partial \mathcal{W}/\partial \lambda_k\right)(\boldsymbol{\Phi}^{\hat{\boldsymbol{\lambda}}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$; but

$$\mathcal{W}(\boldsymbol{\Phi}^{\boldsymbol{\lambda}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}) = \max_{\boldsymbol{\mu}} \left(\sum_{x,y} \mu_{xy}\Phi_{xy}^{\boldsymbol{\lambda}} + \mathcal{E}(\boldsymbol{\mu})\right) = \sum_{x,y} \mu_{xy}^{\boldsymbol{\lambda}}\Phi_{xy}^{\boldsymbol{\lambda}} + \mathcal{E}(\boldsymbol{\mu}^{\boldsymbol{\lambda}})$$

and $\Phi_{xy}^{\boldsymbol{\lambda}} = \sum_k \lambda_k \phi_{xy}^k$. Therefore by the Envelope Theorem, $\left(\partial \mathcal{W}/\partial \lambda_k\right)(\boldsymbol{\Phi}^{\hat{\boldsymbol{\lambda}}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}) = \sum_{x,y} \mu_{xy}^{\hat{\boldsymbol{\lambda}}}\phi_{xy}^k$.

(ii) Given (1.13), the program (4.4) can be rewritten as

$$\max_{\boldsymbol{\lambda}\in\mathbb{R}^K} \min_{\boldsymbol{\mu}\in\mathcal{M}(\hat{\boldsymbol{n}},\hat{\boldsymbol{m}})} \left(\sum_k \lambda_k \sum_{x,y} \left(\hat{\mu}_{xy} - \mu_{xy}\right)\phi_{xy}^k - \mathcal{E}\left(\boldsymbol{\mu}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}\right)\right).$$

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 $\sum_{x,y}(\hat{\mu}_{xy}-\mu_{xy})\,\phi_{xy}^k \neq 0$ for some k; then minimizing over $\boldsymbol{\lambda}$ gives $-\infty$. Therefore these equalities must hold at the optimum, and $\boldsymbol{\mu}$ minimizes $\mathcal{E}$ over the set of $\boldsymbol{\mu} \in \mathcal{M}(\hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$ such that $\sum_{x,y}(\hat{\mu}_{xy}-\mu_{xy})\,\phi_{xy}^k = 0$ for all k.

(iii) Given (1.14), the program in Theorem 3 can be rewritten as

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^k}\left(\mathcal{W}(\boldsymbol{\Phi}^{\boldsymbol{\lambda}},\hat{\boldsymbol{r}}) - \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}}\hat{\mu}_{xy}\Phi_{xy}^{\boldsymbol{\lambda}}\right)$$

which is just the definition of the moment matching estimator in (4.5). QED

A.9 Proof of Proposition 4

Denote $\hat{\boldsymbol{\lambda}} := \hat{\boldsymbol{\lambda}}^{MM}$. Since $\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}}$ maximizes $\sum_{x,y}\mu_{xy}\Phi_{xy}^{\hat{\boldsymbol{\lambda}}} + \mathcal{E}(\boldsymbol{\mu}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$ wrt $\boldsymbol{\mu}$,

$$\mathcal{E}\left(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}\right) - \mathcal{E}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}\right) \geq \sum_{x,y}\left(\hat{\mu}_{xy} - \mu_{xy}^{\hat{\boldsymbol{\lambda}}}\right)\Phi_{xy}^{\hat{\boldsymbol{\lambda}}}$$

and the inequality is strict unless $\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}} = \hat{\boldsymbol{\mu}}$, since $\mathcal{E}$ is strictly concave in $\boldsymbol{\mu}$. But the RHS is zero by definition of the moment-matching estimator $\hat{\boldsymbol{\lambda}}$. Therefore $\mathcal{E}(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}) \geq \mathcal{E}\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}\right)$, with equality if and only if $\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}} = \hat{\boldsymbol{\mu}}$. Finally, item (iii) of Theorem 3 shows that $\mathcal{E}(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}) = \mathcal{E}_{\max}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}})$ if the model is well-specified.

A.10 Proof of Proposition 5

Letting $\mu_{x0}^{\lambda} = n_x \exp(-u_x)$, $\mu_{0y}^{\lambda} = m_y \exp(-v_y)$ and $\mu_{xy}^{\lambda} = \sqrt{n_x m_y}\exp\left(\frac{\Phi_{xy}^{\lambda} - u_x - v_y}{2}\right)$, the optimality conditions with respect to u, v and λ imply respectively that μ_{xy}^{λ} has the right x and y margins, and the right moments k, that is

$$\begin{cases} \sum_{y\in\mathcal{Y}}\mu_{xy}^{\lambda} + \mu_{x0}^{\lambda} = n_x \\ \sum_{x\in\mathcal{X}}\mu_{xy}^{\lambda} + \mu_{0y}^{\lambda} = m_y \\ \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}}\mu_{xy}^{\lambda}\phi_{xy}^k = \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}}\hat{\mu}_{xy}\phi_{xy}^k \end{cases}.$$

A.11 Proof of Theorem 4

We start by extending the generalized entropy $\mathcal{E}$ to a strictly concave function E as explained in A.3. For notational simplicity, we now drop the arguments $\boldsymbol{r}$ and $\boldsymbol{\Phi}$. Corollary 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:

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

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

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

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 expected 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 $\mathcal{M}\left(\boldsymbol{n}\right)$ and $\mathcal{M}\left(\boldsymbol{m}\right)$ by

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

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

$$\boldsymbol{\mu}^{(2k+1)} = \arg\min_{\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{n})} D\left(\boldsymbol{\mu}, \boldsymbol{\mu}^{(2k)}\right) \text{ and } \boldsymbol{\mu}^{(2k+2)} = \arg\min_{\boldsymbol{\mu} \in \mathcal{M}(\boldsymbol{m})} D\left(\boldsymbol{\mu}, \boldsymbol{\mu}^{(2k+1)}\right). \tag{A.4}$$

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 (1.13).

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.2). In practice, good judgement should be exercised, as the choice of an extension E that makes it easy to solve the systems in A.4 is crucial for the performance of the algorithm.

B Additional Elements

This section is dedicated to additional discussions, supplementary results, and interpretations that are not covered in the main text.

B.1 The Separability assumption

Recall from assumption 1 that we have assumed that the individual matching surplus $\tilde{\Phi}$ is separable in the sense that in the terms of equation (1.2), one has $\tilde{\Phi}_{ij} = \Phi_{xy} + \varepsilon_{iy} + \eta_{xj}$. This form is a consequence of the stronger assumptions imposed by Choo and Siow (2006) who assumed that the utility surplus of a man i of group x who marries a woman of group y can be written as

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

where α_{xy} is the "systematic" part of the surplus; τ represents the utility transfer (possibly negative) that the husband gets from his partner in equilibrium; and ε_{iy} is a standard type I extreme value random term. If such a man remains single, he gets utility ε_{i0}; that is to say, the systematic utilities of singles α_{x0} are normalized to zero. Similarly, the utility of a woman j of group y who marries a man of group x can be written as

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

where τ is the utility transfer she leaves to her partner. Again, we normalize $\gamma_{0y} = 0$.

It is easy to see that Assumption 1 is equivalent to specifying that 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 we shuffle partners: $\tilde{\Phi}_{ij} + \tilde{\Phi}_{i'j'} = \tilde{\Phi}_{ij'} + \tilde{\Phi}_{i'j}$.

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. 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. Note that since there is a continuum of women in each group y, but only one man i, man i must capture all joint surplus if he marries a woman of group y in equilibrium: his net utility must be ε_{iy}, and hers is zero. In other words, this man will receive a transfer $\tau_i = \max_{y \in \mathcal{Y}} \varepsilon_{iy}$, which must depend 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 ε and η terms are observed by all agents.

While separability is a restrictive assumption, it allows for "matching on unobservables" : when the analyst observes a woman of group y matched with a man of group x, it may be 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.

B.2 Logit paradoxes

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 Gumbel-distributed random vectors, hence appending a logit structure to the matching model in consideration. This brings familiar features of the logit model, and in particular, the Independence of Irrelevant Alternatives (IIA) axiom.

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 several definitions and draw out their implications in Galichon and Salanié (2019). Still, it is not hard to construct illustrations similar to Debreu's example with 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. We will propose in section 3.2.2 a new Random Scalar Coefficient model that would indeed leave matching patterns and utilities unchanged.

B.3 Additional comparative statics results

The results of Theorem 2 can be used to show that the comparative statics results of Decker, Lieb, McCann, and Stephens (2012) extend beyond the logit model to our generalized framework, beyond those stated in subsection 2.3. Many of these results are collected in Galichon and Salanié (2017), but we recall some here for completeness. From the results of Section 1.3, recall that $\mathcal{W}(\mathbf{\Phi}, \boldsymbol{r})$ is given by the dual expressions

$$\mathcal{W}(\mathbf{\Phi}, \boldsymbol{r}) = \max_{\mu \in \mathcal{M}(\boldsymbol{r})} \left(\sum_{xy} \mu_{xy} \Phi_{xy} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{r}) \right), \text{ and} \tag{B.3}$$

$$\mathcal{W}(\mathbf{\Phi}, \boldsymbol{r}) = \min_{U_{xy}+V_{xy}=\Phi_{xy}} \left(\sum n_x G_x(U_{xy}) + \sum m_y H_y(V_{xy}) \right); \tag{B.4}$$

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 (2.7), the cross-derivative of $\mathcal{W}$ with respect to $n_{x'}$ and Φ_{xy} yields

$$\frac{\partial \mu_{xy}}{\partial n_{x'}} = \frac{\partial^2 \mathcal{W}}{\partial n_{x'} \partial \Phi_{xy}} = \frac{\partial u_{x'}}{\partial \Phi_{xy}} \tag{B.5}$$

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

$$\frac{\partial \mu_{xy}}{\partial \Phi_{x'y'}} = \frac{\partial^2 \mathcal{W}}{\partial \Phi_{xy} \partial \Phi_{x'y'}} = \frac{\partial \mu_{x'y'}}{\partial \Phi_{xy}}. \tag{B.6}$$

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 effect on marriages between groups x and y. In that case marriages (x, y) and (x', y') are complements. We emphasize here that all the comparative statics derived in this section hold in *any* model satisfying our assumptions.

B.4 Geometric interpretation of the estimation procedure

The approach to inference we describe in section 4.2 has a simple geometric interpretation. In this appendix, we fix the distributions $\boldsymbol{P}_x$ and a specification $(\phi_{xy}^k)_{k=1,...,K}$ 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:

$$\mathcal{F} = \left\{ \left(C^1, ..., C^K\right) : C^k = \sum_{xy} \mu_{xy} \phi_{xy}^k, \ \boldsymbol{\mu} \in \mathcal{M}\left(\hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}\right) \right\}$$

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

Figure 6 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

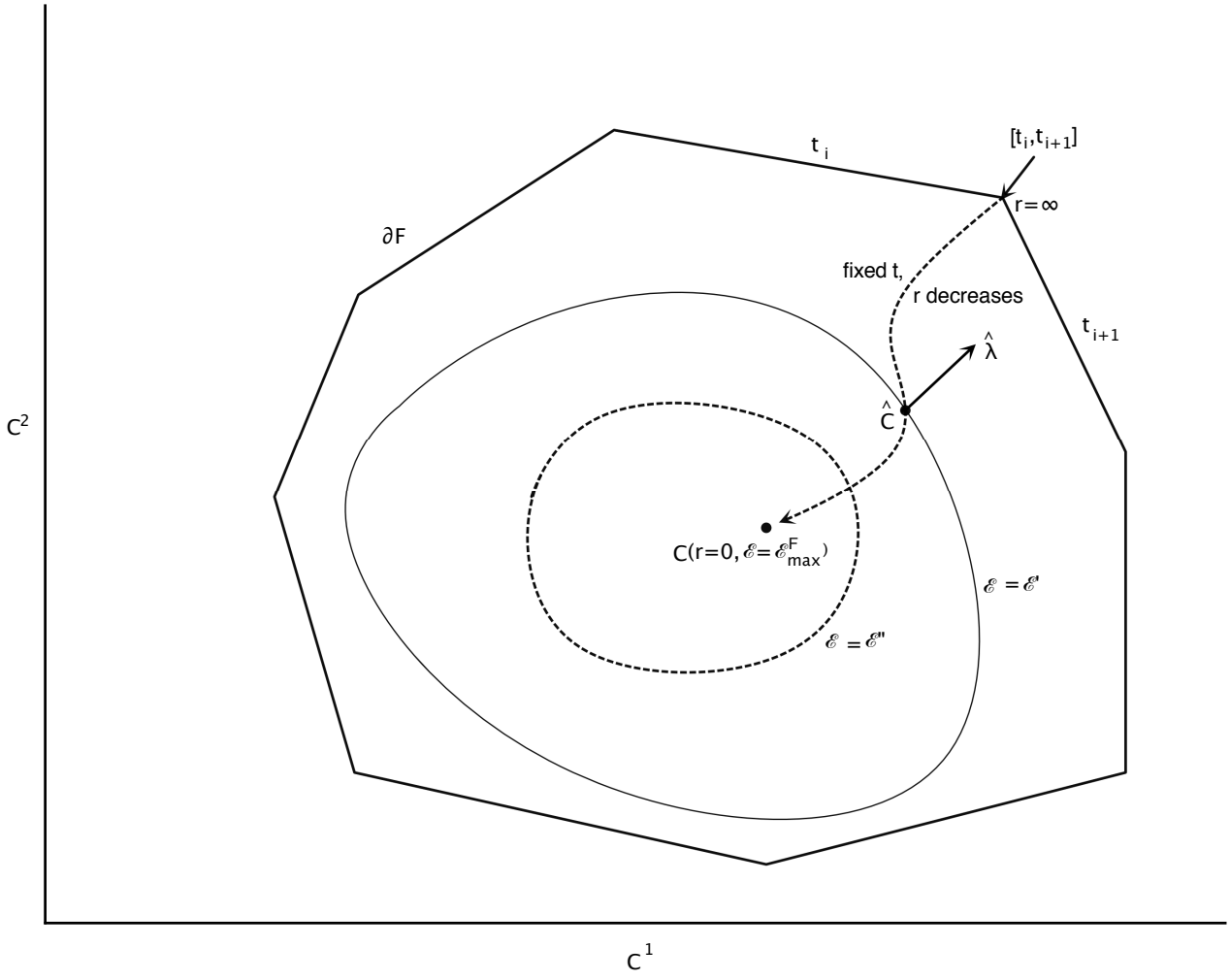


Figure 6: The covariogram and related objects

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 6, 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 3(ii) 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 6.

C Explicit examples of random utility structure

C.1 The Generalized Extreme Value Framework

Consider a family of functions $g_x : \mathbb{R}^{\mathcal{Y}_0} \to \mathbb{R}$ that (i) are positive homogeneous of degree one; (ii) go to $+\infty$ whenever any of their arguments goes to $+\infty$; (iii) are such that their partial derivatives (outside of $\mathbf{0}$) at any order k have sign $(-1)^k$; (iv) are such that the functions defined by $F(w_0, ..., w_J) = \exp(-g_x(e^{-w_0}, ..., e^{-w_J}))$ are multivariate cumulative distribution functions, associated to a distribution which we denote $\boldsymbol{P}_x$. Then introducing utility shocks $\boldsymbol{\varepsilon}_x \sim \boldsymbol{P}_x$, we have by a theorem of McFadden (1978):

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

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

If $\sum_{y\in\mathcal{Y}_0} p_y = 1$, then $G^*_x(\boldsymbol{p}) = \left(\log g_x\left(e^{\boldsymbol{w}^x(\boldsymbol{p})}\right) + \gamma\right) - \sum_{y\in\mathcal{Y}_0} p_y w^x_y(\boldsymbol{p})$, where for $x \in \mathcal{X}$, the vector $w^x(p)$ is a solution to the system of equations

$$p_y = \left(\partial \log g_x/\partial w^x_y\right)\left(e^{\boldsymbol{w}^x}\right) \quad \text{for} \quad y \in \mathcal{Y}_0. \tag{C.2}$$

The generalized entropy arising from the heterogeneity on the men side is

$$G^*(\boldsymbol{\mu}) = \sum_{x\in\mathcal{X}} \{n_x \log g_x \left(e^{\boldsymbol{w}^x(\boldsymbol{\mu}_{x\cdot}/n_x)}\right) - \sum_{y\in\mathcal{Y}_0} \mu_{xy} w_y^x \left(\boldsymbol{\mu}_{x\cdot}/n_x\right)\} + C \tag{C.3}$$

where $C = \gamma \sum_{x\in\mathcal{X}} n_x$. Applying the envelope theorem, the derivative of this expression with respect to μ_{xy} $(x, y \geq 1)$ is $-w_y^x \left(\boldsymbol{\mu}_{x\cdot}/n_x\right)$. Therefore

$$U_{xy} = w_y^x \left(\boldsymbol{\mu}_{x\cdot}/n_x\right). \tag{C.4}$$

C.1.1 Example 1: The logit model of Choo and Siow

Claims of Section 3.1. With centered standard type I extreme value iid distributions $G(-\gamma, 1)$, the function g_x in (C.1) is $g_x(\boldsymbol{t}) = \exp(-\gamma)(1 + \sum_{y\in\mathcal{Y}} t_y)$. The expected utility is $G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \log(1 + \sum_{y\in\mathcal{Y}} \exp(U_{xy}))$, and the maximum in the program that defines $G_x^*(\boldsymbol{\mu}_{\cdot|x})$ is achieved by $U_{xy} = \log(\mu_{y|x}/\mu_{0|x})$. This yields

$$G_x^*(\boldsymbol{\mu}_{\cdot|x}) = \sum_{y\in\mathcal{Y}} \mu_{y|x} \log \frac{\mu_{y|x}}{\mu_{0|x}} - \log\left(1 + \sum_{y\in\mathcal{Y}} \frac{\mu_{y|x}}{\mu_{0|x}}\right) = \mu_{0|x} \log(\mu_{0|x}) + \sum_{y\in\mathcal{Y}} \mu_{y|x} \log \mu_{y|x}$$

which gives equation (3.3). Equation (3.4) obtains by straightforward differentiation.

(This also follows from our formulæ for GEV: here $\left(\partial \log g_x/\partial w_y^x\right)\left(e^{w^x}\right) = \exp(w_y^x)/(1 + \sum_{y\in\mathcal{Y}} \exp(w_y^x))$ in (C.2), which gives $w_y^x(p) = \log(p_y/p_0)$.)

We know from corollary 1 that the value of the social welfare is given by

$$\min_{\boldsymbol{a},\boldsymbol{b}} \left(\sum_{x\in\mathcal{X}} n_x a_x + \sum_{y\in\mathcal{Y}} m_y b_y + S(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Phi}, \boldsymbol{r})\right),$$

where $S(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Phi}, \boldsymbol{r})$ is the maximum value of

$$E(\boldsymbol{\mu}, \boldsymbol{r}) + \sum_{x,y} \mu_{xy}(\Phi_{xy} - a_x - b_y) - \sum_x \mu_{x0} a_x - \sum_y \mu_{0y} b_y.$$

Easy calculations for the Choo and Siow model give

$$S(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Phi}, \boldsymbol{r}) = F(\boldsymbol{a}, \boldsymbol{b}; \boldsymbol{\Phi}, \boldsymbol{r}).$$

Taking first-order conditions gives the formulæ for μ_{x0}, μ_{0y}, and μ_{xy} in the text.

C.1.2 The nested logit model

For concreteness, we develop a very simple two-level nested logit model; generalizations are immediate.

Example 3 (A two-level nested logit model)**.** *Suppose for instance that men of a given group x are concerned about the social group of their partner and her education, so that $y = (s, e)$. We can allow for correlated preferences by modeling this as a nested logit in which educations are nested within social groups. Let $\boldsymbol{P}_x$ have cdf*

$$F(\boldsymbol{w}) = \exp\left(-\exp(-w_0) - \sum_s \left(\sum_e \exp(-w_{se}/\sigma_s)\right)^{\sigma_s}\right)$$

This is a particular case of the Generalized Extreme Value (GEV) framework described in (C.1), with $g_x(z) = z_0 + \sum_s \left(\sum_e z_{se}^{1/\sigma_s}\right)^{\sigma_s}$. The numbers $1/\sigma_s$ describe the correlation in the surplus generated with partners of different education levels within social group s.

Identifying Utilities and Surplus. For simplicity, we center the type I extreme value distributions. Consider a man of a group x (the x indices will be dropped for convenience, so that for instance μ_s denotes the mass of matches with women in social group s). By (C.1), the expected utility of this man is

$$G(\boldsymbol{U}.) = \log\left(1 + \sum_s \left(\sum_e e^{U_{se}/\sigma_s}\right)^{\sigma_s}\right), \tag{C.5}$$

hence, by (C.2), it follows that $\mu_{se}/\mu_0 = (\sum_e e^{U_{se}/\sigma_s})^{\sigma_s - 1} e^{U_{se}/\sigma_s}$, where μ_0 is again defined in (3.1). Thus $\log(\mu_s/\mu_0) = \sigma_s \log(\sum_e \exp(U_{se}/\sigma_s))$, and therefore $U_{se} = \log(\mu_s/\mu_0) + \sigma_s \log(\mu_{se}/\mu_s)$. Now by (C.3) (applied to men of group x only)

$$\begin{aligned} G^*(\boldsymbol{\mu}.) &= \sum \mu_{se} U_{x,se} - \log\left(1 + \sum_s \left(\sum_e e^{U_{se}/\sigma_s}\right)^{\sigma_s}\right) \\ &= \mu_0 \log \mu_0 + \sum_s (1 - \sigma_s)\, \mu_s \log \mu_s + \sum_{s,e} \sigma_s \mu_{se} \log \mu_{se}. \end{aligned}$$

As in Example 1, the expected utility is $u = -\log \mu_0$.

If the nested logit applies for men of group x with parameters (σ_s^x) and for women of group y with parameters $(\tau_{s'}^y)$, we can write $U_{x,se} = \log(\mu_{x,s}/\mu_{x,0}) + \sigma_s^x \log(\mu_{x,se}/\mu_{x,s})$ and $V_{y,s'e'} = \log(\mu_{s',y}/\mu_{0,y}) + \tau_{s'}^y \log(\mu_{s'e',y}/\mu_{s',y})$. Adding up gives the formula for the surplus from a match between a man of group $x = (s', e')$ and a woman of group $y = (s, e)$:

$$\Phi_{xy} = \log \frac{\mu_{xy}^{\sigma_s^x + \tau_{s'}^y} \mu_{x,s}^{1-\sigma_s^x} \mu_{s',y}^{1-\tau_{s'}^y}}{\mu_{x0}\mu_{0y}}. \tag{C.6}$$

Note that we recover the results of Example 1 when all σ and τ parameters equal 1; and if there is only one possible social group s, s', then we recover the heteroskedastic model of Chiappori, Salanié, and Weiss (2017).

Maximum Likelihood Estimation. In this nested logit model, where the groups of men and women are respectively (s_x, e_x) and (s_y, e_y), one can take $\sigma_{s_y}^{s_x,e_x}$ and $\sigma_{s_x}^{s_y,e_y}$ as parameters. Assume that there are N_s social categories and N_e classes of education. This gives us $N_s^2 \times N_e^2$ equations, which leaves at most $(N_s^2 \times N_e^2 - 2N_s^2 \times N_e)$ degrees of freedom to parameterize the surplus function $\mathbf{\Phi}^\theta$ with a parameter vector θ. Letting $\boldsymbol{\lambda} = \left(\sigma_{s_y}^{s_x,e_x}, \sigma_{s_x}^{s_y,e_y}, \theta\right)$, $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ solves the system of equations

$$\Phi_{xy}^\theta = \log \frac{\mu_{xy}^{\sigma_s^x + \tau_{s'}^y} \mu_{x,s}^{1-\sigma_s^x} \mu_{s',y}^{1-\tau_{s'}^y}}{(n_x - \sum_y \mu_{xy})(m_y - \sum_x \mu_{xy})}, \ \forall x \in \mathcal{X}, \ y \in \mathcal{Y}$$

and the log-likelihood can be deduced by (4.1).

Using IPFP to compute the equilibrium. Now assume for simplicity that there is only one social group, so the model boils down to a heteroskedastic logit model with scale parameters σ^x and τ^y. Recall the equilibrium formula:

$$\mu_{xy} = \mu_{x0}^{\frac{\sigma_x}{\sigma_x+\tau_y}} \mu_{0y}^{\frac{\tau_y}{\sigma_x+\tau_y}} \exp\left(\frac{\Phi_{xy}}{\sigma_x + \tau_y}\right).$$

At step $(2k+1)$, keep μ_{0y} fixed and solve for μ_{x0} such that

$$n_x = \mu_{x0} + \sum_{y \in \mathcal{Y}} \mu_{x0}^{\frac{\sigma_x}{\sigma_x+\tau_y}} \mu_{0y}^{\frac{\tau_y}{\sigma_x+\tau_y}} \left(\frac{\Phi_{xy}}{\sigma_x + \tau_y}\right); \tag{C.7}$$

and at step $2k+2$, keep μ_{x0} fixed and solve for μ_{0y} such that

$$m_y = \mu_{0y} + \sum_{x \in \S} \mu_{0y}^{\frac{\tau_y}{\sigma_x+\tau_y}} \mu_{x0}^{\frac{\sigma_x}{\sigma_x+\tau_y}} \left(\frac{\Phi_{xy}}{\sigma_x + \tau_y}\right). \tag{C.8}$$

Note that steps (C.7) and (C.8) only require inverting a continuous and increasing real function of one variable, and are hence very cheap computationally. This idea can be extended to the fully general nested logit at the cost of having to invert systems of equations whose number of variables is the size N_e of the nests.

C.1.3 The mixture of logits model

Our next example considers a more complex but richer specification. It approximates the distribution of unobserved heterogeneities through a mixture of logits whose location, scale and weights may depend on the observed group:

Example 4 (A mixture of logits)**.** *Assume $\boldsymbol{P}_x$ is a mixture of i.i.d. centered type I extreme value distributions of scale parameters σ_k^x with weights β_k^x for $k = 1, \ldots, K$. By linearity, the ex-ante indirect utility of man of group x is the weighted sum of the corresponding ex-ante indirect utilities computed in Example 1, that is $G_x(\boldsymbol{U}_{x\cdot}) = \sum_k \beta_k^x G_{xk}(\boldsymbol{U}_{x\cdot})$, where $G_{xk}(\boldsymbol{U}_{x\cdot}) = \sigma_k^x \log(1 + \sum_{y\in\mathcal{Y}} e^{U_{xy}/\sigma_k^x})$:*

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \sum_{k=1}^{K} \beta_k^x \sigma_k^x \log\left(1 + \sum_{y\in\mathcal{Y}} e^{U_{xy}/\sigma_k^x}\right). \tag{C.9}$$

By the results of Example 1, $G_{xk}^*(\boldsymbol{\mu}) = \sigma_k^x \sum_{y\in\mathcal{Y}_0} \mu_y \log \mu_y$.

It follows from standard results in convex analysis (see e.g. Rockafellar (1970, section 20)) that:

- *the Legendre-Fenchel transform of a sum of functions is the infimum-convolution of the transforms of the functions in the sum, so that*

$$(f_1 + \ldots + f_K)^*(p) = \inf_{p^1+\ldots+p^K=p} \left(f_1^*(p^1) + \ldots + f_K^*(p^K)\right);$$

- *The transform of $x \to \beta f(x)$ is $\beta f^*\left(\frac{p}{\beta}\right)$ if β is a positive scalar.*

As a result,

$$G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = \inf_{\forall y\in\mathcal{Y}_0, \sum_{k=1}^K \mu_y^k = \mu_{y|x}} \sum_{k=1}^{K} \sigma_k^x \left(\mu_0^k \log \frac{\mu_0^k}{\beta_k^x} + \sum_{y\in\mathcal{Y}} \mu_y^k \log \frac{\mu_y^k}{\beta_k^x}\right) \tag{C.10}$$

and $U_{xy} = \sum_{k=1}^{K} \sigma_k^x \log(\mu_y^k/\mu_0^k)$*, where* $\left(\mu_y^k\right)$ *is the minimizer in (C.10).*

C.1.4 The FC-MNL Model

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

Example 5 (FC-MNL)**.** *The function* g_x *takes the following form:*

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

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*; and the* $\boldsymbol{b}$ *matrix could depend on* x*. 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_x a τ-homogeneous function, rather than 1-homogeneous. This is a harmless normalization. It gives

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

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

$$\mu_{y|x} = \frac{1}{g_x} \exp(U_{xy}/\sigma) \sum_{y'\in\mathcal{Y}_0} b_{y,y'} \left(\frac{\exp(U_{xy}/\sigma) + \exp(U_{xy'}/\sigma)}{2} \right)^{\tau\sigma-1}.$$

It is apparent from the formulæ that the "cross-price elasticities" (the dependence of $\boldsymbol{\mu}_{\cdot|x}$ on $\boldsymbol{U}_{x\cdot}$ 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.

C.2 A non-GEV Model: Random Scalar Coefficients

Claims of Section 3.2. From Proposition 2, $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = -\max_{\pi\in\mathcal{M}_x} \boldsymbol{E}_\pi\left[\zeta_x(Y)\varepsilon\right]$, where π has margins F_ε and $\mu(Y|x=x)$. Since the function $(\varepsilon,\zeta) \longrightarrow \varepsilon\zeta$ is supermodular, the optimal matching must be positively assortative: larger ε's must be matched with y's with larger values of the index $\zeta_x(y)$. For each x, the values of $\zeta_x(y)$ are distinct and we let $\zeta_{(1)} < \ldots < \zeta_{(|\mathcal{Y}|+1)}$ denote the ordered values of distinct values of $\zeta_x(y)$ for $y \in \mathcal{Y}_0$; the value $\zeta_{(k)}$ occurs with probability

$$\Pr\left(\zeta_x(Y) = \zeta_{(k)}|x\right) = \sum_{\zeta_x(y)=\zeta_{(k)}} \mu_{y|x}. \tag{C.11}$$

By positive assortative matching, there exists a sequence $\varepsilon_{(0)} = \inf\varepsilon < \varepsilon_{(1)} < \ldots < \varepsilon_{(|\mathcal{Y}|)} < \varepsilon_{(|\mathcal{Y}|+1)} = \sup\varepsilon$ such that ε matches with a y with $\zeta_x(y) = \zeta_{(k)}$ if and only if $\varepsilon \in [\varepsilon_{(k-1)}, \varepsilon_{(k)}]$. Since probability must be conserved, the sequence is constructed recursively by

$$F_\varepsilon\left(\varepsilon_{(k)}\right) - F_\varepsilon\left(\varepsilon_{(k-1)}\right) = \sum_{\zeta_x(y)=\zeta_{(k)}} \mu_{y|x}, \tag{C.12}$$

giving $F_\varepsilon\left(\varepsilon_{(k)}\right) = \sum_{\zeta_x(y)\leq\zeta_{(k)}} \mu_{y|x}$. As a result, $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = -\sum_{1\leq k\leq|\mathcal{Y}|+1} \zeta_{(k)} e_k$, where $e_k = \int_{\varepsilon_{(k-1)}}^{\varepsilon_{(k)}} \varepsilon f(\varepsilon)\, d\varepsilon = \left(F(\varepsilon_{(k)}) - F(\varepsilon_{(k-1)})\right)\bar{e}_k$, with $\bar{e}_k$ defined as the conditional mean of ε in the interval $\left[\varepsilon_{(k-1)}, \varepsilon_{(k)}\right]$; then $n_x G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = -n_x \sum_{1\leq k\leq|\mathcal{Y}|+1} \zeta_{(k)} \sum_{\zeta_x(y)=\zeta_{(k)}} \mu_{y|x}\bar{e}_k = \sum_y \mu_{xy}\bar{e}_{K(y)}$, with $K(y)$ the value of k such that $\zeta_x(y) = \zeta_{(k)}$. Therefore the entropy is

$$\mathcal{E} = -\sum_{xy} \mu_{xy}\left(\zeta_x(y)\bar{e}_x(y) + \xi_y(x)\bar{f}_y(x)\right), \tag{C.13}$$

where $\bar{e}_x(y)$ is the expected value of ε on the interval $[a,b]$, with a and b the real numbers such that

$$F_x(a) = \sum_{z|\zeta_x(z)<\zeta_x(y)} \mu_{z|x} \text{ and } F_x(b) = \sum_{z|\zeta_x(z)\leq\zeta_x(y)} \mu_{z|x},$$

and $\bar{f}_y(x)$ is defined similarly on the other side of the market.

Note that even if ε_i has full support, the distribution $\mathbb{P}_x$ won't since its components are perfectly correlated (or anti-correlated). This allows for "zero cells" $\mu_{xy} = 0$, which are often encountered in data. If in addition the distribution of ε_i is uniform, we obtain

what we call the Random Uniform Scalar Coefficient Model (RUSC). This last model yields simple closed-form expressions for interior solutions, even though it does not belong to the Generalized Extreme Value (GEV) class.

D Computational Methods and Benchmarks

Section 5 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 of section 3.1, and IPFP.

In the second part of this appendix, we show how linear programming techniques can be used to solve and estimate a model with discretized error distributions.

D.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 $\{1, \ldots, 100\}$; and for each (x, y) match we draw $\Phi_{xy}/2$ from $\mathcal{N}(0, 1)$.

D.1.1 Minpack

We applied the solver Minpack to the system of $(\mathcal{X}| + |\mathcal{Y})$ equations that characterizes the optimal matching (see section 5). Minpack is probably the most-used solver in scientific applications; it underlies many statistical and numerical packages.

D.1.2 Min-Emax

The min-Emax method we described in section 5 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[25] to obtain the solution.

D.1.3 Minimizing F

Formula (3.6) provides us with an alternative method that works on the smaller object (u_x, v_y) of group expected utilities. Here again we used the Knitro optimizer.

D.1.4 IPFP

Finally and as shown in section 5, the IPFP iterations for the Choo and Siow model are given by quadratic equations in only one unknown, the square root of the numbers of singles. These equations are easily solved in closed form. The pseudo-code in Algorithm 1 gives a detailed implementation.

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*

$X \leftarrow size(\boldsymbol{n})$

$Y \leftarrow size(\boldsymbol{m})$

$\boldsymbol{K} \leftarrow \exp(\boldsymbol{\Phi}/2)$

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

$\delta \leftarrow \infty, i \leftarrow 0$

$\boldsymbol{T} \leftarrow 0_Y$

while $\delta > \tau$ *and* $i < I$ ***do***

$\quad \boldsymbol{S} \leftarrow \boldsymbol{K}\boldsymbol{T}$ ▷ *Project on* $\boldsymbol{n}$ *margins*

$\quad \boldsymbol{t} \leftarrow \left(\sqrt{\boldsymbol{S}^2 + 4\boldsymbol{n}} - \boldsymbol{S}\right)/2$

$\quad S \leftarrow \boldsymbol{K}'\boldsymbol{t}$ ▷ *Project on* $\boldsymbol{m}$ *margins*

$\quad \boldsymbol{T} \leftarrow \left(\sqrt{\boldsymbol{S}^2 + 4\boldsymbol{m}} - \boldsymbol{S}\right)/2$

$\quad \delta_1 \leftarrow \max |\boldsymbol{t}^2 + \boldsymbol{t} \odot \boldsymbol{K}\boldsymbol{T} - \boldsymbol{n}|$ ▷ *Error on* $\boldsymbol{n}$ *margins*

$\quad \delta_2 \leftarrow \max |\boldsymbol{T}^2 + \boldsymbol{T} \odot \boldsymbol{K}'\boldsymbol{t} - \boldsymbol{m}|$ ▷ *Error on* $\boldsymbol{m}$ *margins*

$\quad \delta \leftarrow \max(\delta_1, \delta_2)$

$\quad i \Leftarrow i + 1$

end while

if $i \geq I$ ***then***

$\quad$ *Failed to achieve requested precision*

else

$\quad \boldsymbol{\mu} \leftarrow \boldsymbol{K} \odot (\boldsymbol{t} \otimes \boldsymbol{T})$ ▷ $\otimes$ *denotes outer product and* $\odot$ *element-wise product*

end if

D.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 7. Each panel gives the distribution of CPU times for one of the four methods, in the form of a Tukey box-and-whiskers graph[26].

[26]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 the circles plot all points beyond that.

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.

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

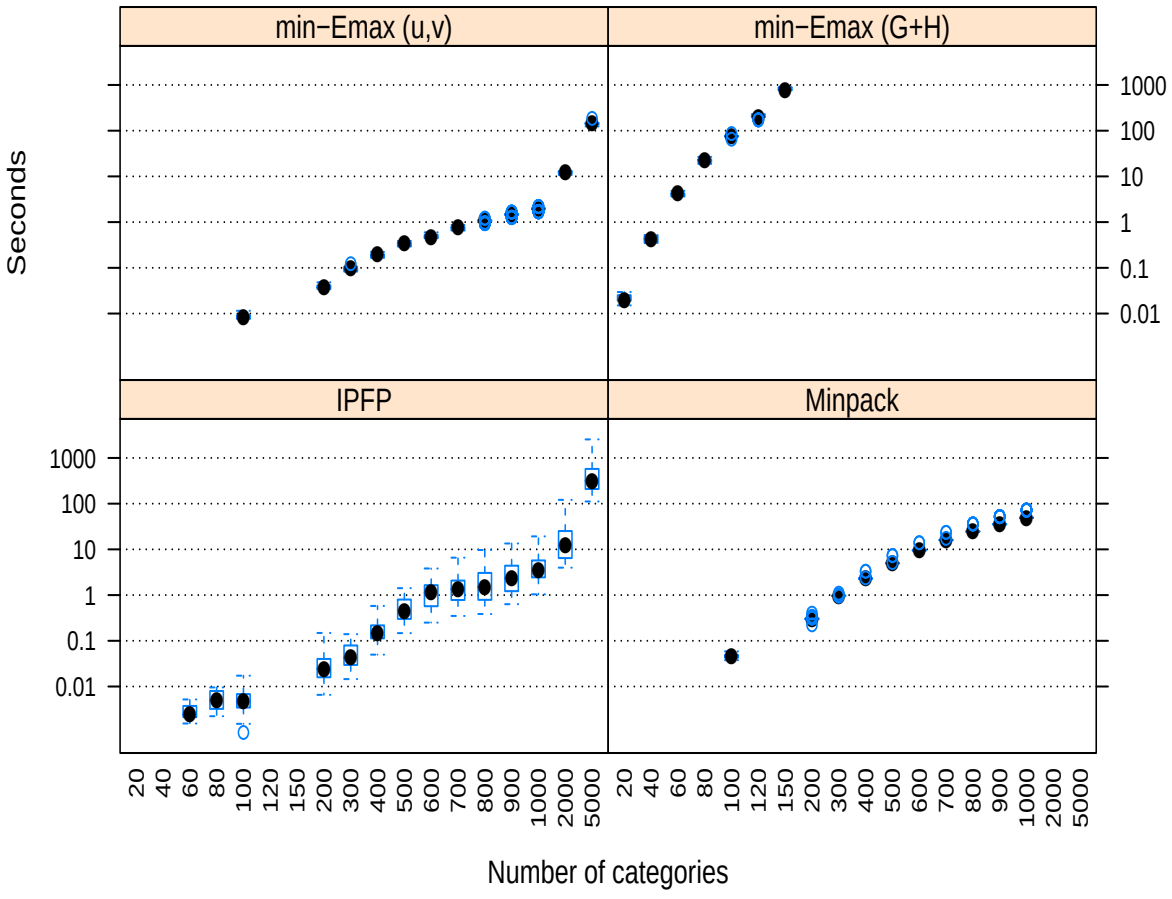


Figure 7: Solving for the optimal matching

D.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 $(\varepsilon_y^{xk})_{y\in\mathcal{Y}_0,k=1,\dots,K_x}$ 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,

$$u_x^k \geq U_{xy}+\varepsilon_y^{xk} \quad \forall y \quad \text{and} \quad u_x^k \geq \varepsilon_0^{xk} \tag{D.1}$$

$$v_y^l \geq V_{xy}+\eta_x^{yl} \quad \forall x \quad \text{and} \quad v_y^l \geq \eta_0^{yl}. \tag{D.2}$$

It follows from (1.14) 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 (D.1) and (D.2) 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 (2019, Theorem 3.1). Hence the approach described in this subsection is applicable to any separable model.

D.2.2 Computing the moment-matching estimator

The linear programming approach of D.2.1 can be extended in order to compute the moment-matching estimator in the semilinear models of section 4.2. Equation (4.4) 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 (D.1) and (D.2), 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 expected 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, which is straightforward in logit-type models.

E The application to Choo and Siow's data

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.

E.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.

E.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[27].
>
> [...]
>
> 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.

[27] 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.

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.

E.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—CS 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.

E.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	Increase in population
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	Increase in population
(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**

E.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[28]. The CS specification by construction rules out zero probability 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 (as with the RSC models of section 3.2.2) and/or because their supports are bounded. Such models may generate numerical and statistical difficulties, however. There exist simple separable models that yield well-behaved matching patterns while still accommodating zero cells. It is easy to see for instance that any RSC model whose error terms ε_i and η_j have a bounded support and densities that vanish at the bounds generates a continuously differentiable $\boldsymbol{\mu}(\boldsymbol{\lambda})$.

F Detailed Estimation Results

F.1 Selecting the Basis Functions

All specifications we use in our model selection are 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 ($D = 0$).

We estimate a set of models (M, N, M_D, N_D) defined as follows: in addition to D and

[28] 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.

to the constant, they also include all terms $x^m Y^n$ for $1 \leq m \leq M$ and $1 \leq n \leq N$, with interactions with the D age difference dummy for degrees less than $1 \leq M_D \leq M$ and $1 \leq N_D \leq N$. Such a model has $(M+1)(N+1)+(M_D+1)(N_D+1)$ basis functions. The nonparametric model (NP in what follows) has as many as there are matching cells, that is $(40-16)^2 = 625$.

We take $M = N = 6$, which gives us $(6-1)^4 = 625$ parametric models. The least complex has two basis functions (1 and D) and the richest one has 98. We estimate them with the moment matching method described in 4.2, which is much faster than maximum likelihood. After a model estimation has converged, we compute its log-likelihood and the values of the Akaike and Bayesian Information Criteria (AIC and BIC.)

Figure 1 in the text shows the values of the AIC (on the horizontal axis) and of the BIC (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. The best models are $(M, N, M_D, N_D) = (6, 5, 2, 5)$ by AIC standards, and $(M, N, M_D, N_D) = (2, 4, 2, 4)$ with BIC.

The 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 and improved the AIC slightly. While the AIC criterion subtracts 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 rewards parsimony 6.2 times more. As a result, the BIC-selected model only has 30 coefficients, of which 28 are significant at 5%. 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 $1 \leq n \leq 4$.

F.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.

F.2.1 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[29]. 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

$$\sigma_x = \exp(\sigma_1 x + \ldots + \sigma_p x^p)$$
$$\tau_y = \exp(\tau_0 + \tau_1 y + \ldots + \tau_q y^q)$$

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

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

Of all such specifications for $0 \leq p \leq 4$ and $1 \leq q \leq 4$, This specification yields a noticeable improvement in the fit: +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.

F.2.2 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.

F.2.3 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 6.4. The maximum likelihood estimates for the distributional parameters[30] are in Table 6.

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

F.3 Random Scalar Coefficient Models

We dedicated a lot of effort to several variants of the RSC model. Unfortunately, we were unable to successfully fit any of them. We found it quite difficult to guess good initial values for the parameters of the scale factors $\zeta_x(y)$ and $\xi_y(x)$. While it usually is tempting to start with reasonable constant values for heteroskedasticity factors, it cannot be done here since the model is only well-defined when ζ_x varies with y and ξ_y varies with x. In addition, some simulated matching patterns become zero when the inequalities that define them are incompatible, even with full support error distributions. This makes the optimization hard to manage.

	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

	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

	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