Cupid's Invisible Hand:

Social Surplus and Identification in Matching Models

Alfred Galichon[1] **Bernard Salanié**[2]

May 10, 2014[3]

[1]Economics Department, Sciences Po, Paris and CEPR; e-mail: alfred.galichon@sciences-po.fr
[2]Department of Economics, Columbia University; e-mail: bsalanie@columbia.edu.

[3]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 and many seminar participants for 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 the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 313699, and from FiME, Laboratoire de Finance des Marchés de l'Energie.

Abstract

We investigate a model of one-to-one matching with transferable utility when some of the characteristics of the players are unobservable to the analyst. We allow for a wide class of distributions of unobserved heterogeneity, subject only to a separability assumption that very significantly extends Choo and Siow (2006). We first show that the stable matching maximizes a social gain function that trades off a sorting effect due to complementarities in observable characteristics, and a randomization effect caused by the presence of 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. If transfers are observed, then the pre-transfer utilities of both partners are also identified. We present a a discussion of computational issues, including an algorithm which can be extremely efficient in important instances. We conclude by discussing some empirical approaches suggested by these results.

Keywords: matching, marriage, assignment, hedonic prices.

JEL codes: C78, D61, C13.

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. Given transferable utilities, 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. These models have also been applied to competitive equilibrium with hedonic pricing (Chiappori, McCann and Nesheim, 2010) and the market for CEOs (Terviö, 2008 and Gabaix and Landier, 2008). We will show how our results can be used in these three contexts; but for concreteness, we often refer to partners as men and women in the 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 classical optimal assignment algorithms.

This sorting result is both simple and powerful; but its implications are also quite unrealistic and at variance with the data, in which matches are observed between partners with quite different characteristics. To account for this wider variety of matching patterns, one could introduce search frictions, as in Shimer and Smith (2000) or Jacquemet and Robin (2011). But the resulting model is hard to handle, and under some additional conditions it still implies assortative matching. An alternative solution consists in allowing the joint surplus of a match to incorporate latent characteristics—heterogeneity that is unobserved by the analyst. Choo and Siow (2006) have shown that 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. Then the usual apparatus of multinomial logit discrete choice models applies, linking marriage patterns to marital surplus in a very simple manner. Choo and Siow used this model to link the changes in gains to marriage and abortion laws; Siow and Choo (2006) 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, 2008) and to test for complementarities across partner educations (Siow, 2009); and, in a heteroskedastic version, to estimate the changes in the returns to education on the US marriage market (Chiappori, Salanié and Weiss, 2012).

We revisit here the theory of matching with transferable utilities in the light of Choo and Siow's insights; and we extend this framework to quite general distributions of unobserved variations in tastes. Our main contributions are threefold.

First, we show that the analysis can be carried much more generally outside of the very restrictive logit framework. We prove that the optimal matching in our generalized setting maximizes a very simple function: 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. 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, most quantities of interest can be obtained from derivatives of the terms that constitute generalized entropy. We show in particular that the joint surplus from matching is (minus) a derivative of the generalized entropy, computed at the observed matching. The expected and realized utilities of all groups of men and women follow just as directly. Moreover, if equilibrium transfers are observed, then we also identify the pre-transfer utilities on both sides of the market.

To prove these results, we use tools from convex analysis, and we construct the Legendre-

Fenchel transform of the expected utilities of agents. In independent work, Decker et al. (2012) proved the uniqueness of the equilibrium and derived some of its comparative static properties in the Choo and Siow multinomial logit framework. Our approach shows that the essence of these comparative static results holds beyond the logit framework. The first conclusion of our paper is thus that the most important structural implications of the Choo-Siow model are not a consequence of the logit framework, but hold under much more plausible assumptions on the unobserved heterogeneity.

Our second contribution is to delineate an empirical approach to parametric estimation in this class of models, using maximum likelihood. Indeed, our nonparametric identification results rely on the strong assumption that the distribution of the unobservables is known, while in practice the analyst will want to estimate its parameters; at the same time our results imply that the matching surplus cannot be simultaneously estimated with the distribution of the unobservable because there would be more parameters than cells in the data matrix. This suggests using a smaller number of parameters for the match surpluses. Maximum likelihood estimation is thus a natural recourse, which we investigate below. In practice, since evaluating the likelihood requires solving for the optimal matching, computational considerations loom large in matching models. We provide an efficient algorithm that maximizes the social surplus and computes the optimal matching, as well as the expected utilities in equilibrium. To do this, we adapt the Iterative Projection Fitting Procedure (known to some economists as RAS) to the structure of this problem, and we show that it is very stable and efficient. Finally, we discuss an alternative to the maximum likelihood, a simple moment matching estimator based on minimizing a generalized entropy among the matching distributions which fit a number of moments. This approach provides a very simple semi-parametric specification test.

Our third contribution is to revisit the original Choo and Siow dataset making use of the new possibilities allowed by this extended framework.

There are other approaches to estimating matching models with unobserved heterogeneity; see the handbook chapter by Graham (2011). Fox (2010) in particular exploits

a "rank-order property" and pools data across many similar markets; see Fox (2011) and Bajari and Fox (2013) for applications. More recently, Fox and Yang (2012) focus on identifying the complementarity between unobservable characteristics. A recent contribution by Menzel (2014) investigates the case when utility is assumed not transferable. We discuss the pros and cons of various methods in our conclusion.

Section 1 sets up the model and the notation. We prove our main results in Section 2, and we specialize them to leading examples in Section 3. 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 model. Finally, we present in Section 6 our IPFP algorithm, which greatly accelerates computations in important cases.

1 The Assignment Problem with Unobserved Heterogeneity

Throughout the paper, we maintain the basic assumptions of the transferable utility model 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", but our results are clearly not restricted to the marriage market.

Men are denoted by $i \in \mathcal{I}$ and women by $j \in \mathcal{J}$. A *matching* $(\tilde{\mu}_{ij})$ is a matrix such that $\tilde{\mu}_{ij} = 1$ if man i and woman j are matched, 0 otherwise. A matching is *feasible* if for every i and j,

$$\sum_{k \in \mathcal{J}} \tilde{\mu}_{ik} \leq 1 \text{ and } \sum_{k \in \mathcal{I}} \tilde{\mu}_{kj} \leq 1,$$

with equality for individuals who are married. Single individuals are "matched with 0": $\tilde{\mu}_{i0} = 1$ or $\tilde{\mu}_{0j} = 1$. For completeness, we should add the requirement that $\tilde{\mu}_{ij}$ is integral $(\tilde{\mu}_{ij} \in \{0, 1\})$. However it is known since at least Shapley and Shubik (1972) that this constraint is not binding, and we will omit it.

A hypothetical match between man i and woman j allows them to share a total utility $\tilde{\Phi}_{ij}$; the division of this total utility between them is done through utility transfers whose value is determined in equilibrium. Singles get utilities $\tilde{\Phi}_{i0}, \tilde{\Phi}_{0j}$. Following Gale and Shapley (1962) for matching with non-transferable utility, we focus on the set of *stable matchings*. A feasible matching is stable if there exists a division of the surplus in each realized match that makes it impossible for any man k and woman l to both achieve strictly higher utility by pairing up together, and for any agent to achieve higher utility by being single. More formally, let $\tilde{u}_i$ denote the utility man i gets in his current match; denote $\tilde{v}_j$ the utility of woman j. Then by definition $\tilde{u}_i + \tilde{v}_j = \tilde{\Phi}_{ij}$ if they are matched, that is if $\tilde{\mu}_{ij} > 0$; and $\tilde{u}_i = \tilde{\Phi}_{i0}$ (resp. $\tilde{v}_j = \tilde{\Phi}_{0j}$) if i (resp. j) is single. Stability requires that for every man k and woman l, $\tilde{u}_k \geq \tilde{\Phi}_{k0}$ and $\tilde{v}_l \geq \tilde{\Phi}_{0l}$, and $\tilde{u}_k + \tilde{v}_l \geq \tilde{\Phi}_{kl}$ for any potential match (k, l).

Finally, a *competitive equilibrium* is defined as a set of prices $\tilde{u}_i$ and $\tilde{v}_j$ and a feasible matching $\tilde{\mu}_{ij}$ such that

$$\tilde{\mu}_{ij} > 0 \text{ implies } j \in \arg\max_{j \in \mathcal{J} \cup \{0\}} \left(\tilde{\Phi}_{ij} - \tilde{v}_j \right) \text{ and } i \in \arg\max_{i \in \mathcal{I} \cup \{0\}} \left(\tilde{\Phi}_{ij} - \tilde{u}_i \right). \tag{1.1}$$

Shapley and Shubik showed that the set of stable matchings coincides with the set of competitive equilibria (and with the core of the assignment game); and that moreover, any stable matching achieves the maximum of the total surplus

$$\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \tilde{\nu}_{ij} \tilde{\Phi}_{ij} + \sum_{i \in \mathcal{I}} \tilde{\nu}_{i0} \tilde{\Phi}_{i0} + \sum_{j \in \mathcal{J}} \tilde{\nu}_{0j} \tilde{\Phi}_{0j}$$

over all feasible matchings $\tilde{\nu}$. The set of stable matchings is generically a singleton; on the other hand, the set of prices $\tilde{u}_i$ and $\tilde{v}_j$ (or, equivalently, the division of the surplus into $\tilde{u}_i$ and $\tilde{v}_j$) that support it is a product of intervals. This discrete setting was extended by Gretsky, Ostroy and Zame (1992) to a continuum of agents.

1.1 Observable characteristics

The analyst only observes some of the payoff-relevant characteristics that determine the surplus matrix $\tilde{\Phi}$. Following Choo and Siow, we assume that she 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 $y_j \in \mathcal{Y}$. Groups are defined by the intersection of 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.

As an example, observed groups $x, y = (E, R)$ may consist of education and income. Education could take values $E \in \{D, G\}$ (dropout or graduate), and income class R could take values 1 to n_R. Groups may also incorporate information that is sometimes available to the econometrician, such as physical characteristics, religion, and so on. In this paper we take the numbers of groups $|\mathcal{X}|$ and $|\mathcal{Y}|$ to be finite in number; we return to the case of continuous groups in the conclusion.

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, and as the problem is homogenous, we can assume that the total mass of individuals is equal to one. More formally, we assume:

Assumption 1 (Large Market)**.** *There is an infinite total number of individuals on the market. Letting n_x be the mass of men of group x, and m_y the mass of women of group y, the total mass of individuals is normalized to one, that is $\sum_x n_x + \sum_y m_y = 1$.*

One way to understand intuitively this assumption is to consider a sequence of large economies of total population of size N growing to infinity, that is

$$N = \sum_{x \in \mathcal{X}} N_x + \sum_{y \in \mathcal{Y}} M_y \to +\infty$$

while the proportion of each group remains constant, that is, the ratios $n_x = (N_x/N)$ and $m_y = (M_y/N)$ remain constant.

The effect of assuming an infinite number of individuals is that we will not have to worry about sampling issues when dealing with the distributions of the unobserved heterogeneity in Section 1.2. If the total number of individuals were finite, the distribution of the unobserved

heterogeneity of, say, women of a given observable group would be an empirical distribution affected by sample uncertainty.

Another benefit of Assumption 1 is that it mitigates concerns about agents misrepresenting their characteristics. There is almost always a profitable deviation in finite populations; but as shown by Gretsky, Ostroy and Zame (1999), the benefit from such manipulations goes to zero as the population is replicated. In the large markets limit, the Walrasian prices $\tilde{u}_i$ and $\tilde{v}_j$ become generically unique. We will therefore write "the equilibrium" in what follows.

The analyst does not observe some of the characteristics of the players, and she can only compute quantities that depend on the observed groups of the partners in a match. Hence she cannot observe $\tilde{\mu}$, and she must focus instead on the matrix of matches across groups (μ_{xy}). This is related to $(\tilde{\mu}_{ij})$ by

$$\mu_{xy} = \sum_{i,j} \mathbb{1}\left(x_i = x, y_j = y\right) \tilde{\mu}_{ij}.$$

The feasibility constraints on $\mu_{xy} \geq 0$ are $\mu \in \mathcal{M}\left(n, m\right)$, where $\mathcal{M}\left(n, m\right)$ (or $\mathcal{M}$ in the absence of ambiguity) is the set of $(|\mathcal{X}|\,|\mathcal{Y}| + |\mathcal{X}| + |\mathcal{Y}|)$ non-negative numbers (μ_{xy}) that satisfy the $(|\mathcal{X}| + |\mathcal{Y}|)$ following inequalities

$$\mathcal{M}\left(n, m\right) = \{\mu \geq 0 : \forall x \in \mathcal{X},\ \sum_{y \in \mathcal{Y}} \mu_{xy} \leq n_x\ ;\ \forall y \in \mathcal{Y},\ \sum_{x \in \mathcal{X}} \mu_{xy} \leq m_y\} \tag{1.2}$$

which simply means that the number of married men (women) of group x (y) is not greater than the number of men (women) of group x (y). Each element of $\mathcal{M}$ is called a "matching" as it defines a feasible set of matches (and singles). For notational convenience, we shall denote μ_{x0} the number of single men of group x and μ_{0y} the number of single women of group y, and

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

where $\mathcal{X}_0$ and $\mathcal{Y}_0$ are the set of marital choices that are available to male and female agents,

including singlehood. Obviously,

$$\mu_{x0} = n_x - \sum_{y \in \mathcal{Y}} \mu_{xy} \text{ and } \mu_{0y} = m_y - \sum_{x \in \mathcal{X}} \mu_{xy}.$$

1.2 Matching Surpluses

Several approaches can be used to take this model to the data. A computationally complex method would use a parametric specification for the surplus $\tilde{\Phi}_{ij}$ and solve the system of equilibrium equations (1.1). The set of maximizers at the solution of this system defines the stable matchings, and can be compared to the observed matching in order to derive a minimum distance estimator of the parameters. However, there are two problems with this approach: it is very costly, and it is not clear at all what drives identification of the parameters. The literature has instead attempted to impose identifying assumptions that allow for more transparent identification. We follow here the framework of Choo and Siow (2006). We will discuss other approaches in the conclusion, including those of Fox (2010) and Fox and Yang (2012).

Choo and Siow assumed that the utility surplus of a man i of group x (that is, such that $x_i = x$) who marries a woman of group y can be written as

$$\alpha_{xy} + \tau + \varepsilon_{iy}, \tag{1.3}$$

where α_{xy} is the systematic part of the surplus, and τ represents the utility transfer (possibly negative) that the man gets from his partner in equilibrium, and ε_{iy} is a standard type I extreme value random variation. 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 $y_j = y$ who marries a man of group x can be written as

$$\gamma_{xy} - \tau + \eta_{xj}, \tag{1.4}$$

where τ is the utility transfer she leaves to her partner. A woman of group y gets utility η_{0j} if she is single, that is we adopt normalization $\gamma_{0y} = 0$.

As shown in Chiappori, Salanié and Weiss (2012), the key assumption here is that the joint surplus created when a man i of group x marries a woman j of group y does not allow for interactions between their unobserved characteristics, conditional on (x, y). This leads us to assume:

Assumption 2 (Separability)**.** *There exists a vector Φ_{xy} such that the joint surplus from a match between a man i in group x and a woman j in group j is*

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

This assumption is reminiscent of the "pure characteristics" model of Berry and Pakes (2007). In Choo and Siow's formulation, the vector Φ is simply

$$\Phi_{xy} = \alpha_{xy} + \gamma_{xy},$$

which they call the *total systematic net gains to marriage*; and note that by construction, Φ_{x0} and Φ_{0y} are zero. It is easy to see that Assumption 2 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: $\tilde{\Phi}_{ij} + \tilde{\Phi}_{i'j'} = \tilde{\Phi}_{ij'} + \tilde{\Phi}_{i'j}$. Note that in this form it is clear that we need not adopt Choo and Siow's original interpretation of ε as a preference shock of the husband and η as a preference shock of the wife. To take an extreme example, we could equally have men who are indifferent over partners and are only interested in the transfer they receive, so that their ex post utility is τ; and women who also care about some attractiveness characteristic of men, in a way that may depend on the woman's group. The net utility of women of group y would be $\varepsilon_{iy} - \tau$; the resulting joint surplus would satisfy Assumption 2 and all of our results would apply[1]. In other words, there is no need to assume that the term ε_{iy_j} was "created" by man i, nor that the term η_{jx_i} was "created" by the woman j; it may perfectly be the opposite.

[1]It is easy to see that in such a model, a man i who is married in equilibrium is matched with a woman in the group that values his attractiveness most, and he receives a transfer $\tau_i = \max_{y \in \mathcal{Y}} \varepsilon_{iy}$.

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, i.e. some unobserved preference of this man for some unobserved characteristics of that woman.

The basic problem we address in this paper is how we can identify (Φ_{xy}) (an array of unknowns of the same dimension) given the observation of (μ_{xy}) (an array of $|\mathcal{X}| \times |\mathcal{Y}|$ numbers). In order to study the relation between these two objects, we need to make assumptions on the distribution of the unobserved heterogeneity terms, which we now describe.

1.3 Unobserved Heterogeneity

In order to move beyond the multinomial logit setting of Choo and Siow, we allow for quite general distributions of unobserved heterogeneity in the following way:

Assumption 3 (Distribution of Unobserved Variation in Surplus)**.**

a) For any man i such that $x_i = x$, ε_{iy} is a $|\mathcal{Y}_0|$-dimensional random vector drawn from a zero-mean distribution $\mathbf{P}_x$;

b) For any woman j such that $y_j = y$, η_{xj} is a $|\mathcal{X}_0|$-dimensional random vector drawn from a zero-mean distribution $\mathbf{Q}_y$;

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

Parts (a) and (b) of Assumption 3 clearly constitute a substantial generalization with respect to Choo and Siow. This extends the logit framework in several important ways: it allows for different families of distributions, with any form of heteroskedasticity, and with any pattern of correlation across partner groups.

As will be clear from the examples below, and unlike the standard logit (i.i.d. extreme value) framework, Assumption 3 is flexible enough to allow for correlation between the utility shocks: in the present framework, one individual may have, for instance, correlated utility shocks for matching with partners of various education groups. The need to go beyond the logit framework has long been felt in Industrial Organization and in consumer demand theory, which has led to a large literature on Random Utility Models, initiated by McFadden's seminal work on Generalized Extreme Value theory (McFadden, 1978, see also Anderson et al., 1992 for a good exposition and applications). The present assumption is more general, as it does not require that the distribution of the terms ε_{iy} and η_{xj} should belong to the GEV class.

2 Social Surplus, Utilities, and Identification

We derive most of our results by considering the "optimal" matching, maximizing the total joint surplus, which is known since Shapley and Shubik (1972) to be equivalent to the equilibrium matching. As Choo and Siow remind us (p. 177): "A well-known property of transferable utility models of the marriage market is that they maximize the sum of marital output in the society". This is true when marital output is defined as it is evaluated by the participants: the market equilibrium in fact maximizes $\sum_{i,j} \tilde{\mu}_{ij} \tilde{\Phi}_{ij}$ over the set of feasible matchings $(\tilde{\mu}_{ij})$. A very naive evaluation of the sum of marital output, computed from the groups of partners only, would be

$$\sum_{xy} \mu_{xy} \Phi_{xy}, \tag{2.1}$$

but this is clearly misleading. Realized matches by nature have a value of the unobserved marital surplus $(\varepsilon_{iy} + \eta_{xj})$ that is more favorable than an unconditional draw; and as a

consequence, the equilibrium marriage patterns μ do not maximize $\sum_{xy} \mu_{xy} \Phi_{xy}$ over $\mathcal{M}$. In order to find the expression of the value function that μ maximizes, we need to account for terms that reflect the value of matching on unobservables.

2.1 Separability and Discrete Choice

We first argue that separability (Assumption 2) reduces the choice of partner to a one-sided discrete choice problem. To see this, note that by standard results in the literature (Shapley and Shubik, 1972), the equilibrium utilities solve the system of functional equations

$$\tilde{u}_i = \max_j \left(\tilde{\Phi}_{ij} - \tilde{v}_j \right) \text{ and } \tilde{v}_j = \max_i \left(\tilde{\Phi}_{ij} - \tilde{u}_i \right),$$

where the maximization includes the option of singlehood.

Focus on the first one. It states that the utility man i gets in equilibrium trades off the surplus his match with woman j creates and the share of the joint surplus he has to give her, which is given by her own equilibrium utility. Now use Assumption 2: for a man i in group x, $\tilde{\Phi}_{ij} = \Phi_{xy_j} + \varepsilon_{iy_j} + \eta_{xj}$, 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 $\tilde{u}_i = \max_y \{ \Phi_{xy} + \varepsilon_{iy} - \min_{j:y_j=y} \left(\tilde{v}_j - \eta_{xj} \right) \}$. 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:

Proposition 1. ***(Splitting the Surplus)***
Under Assumptions 2 and 3, there exist two vectors U_{xy} and V_{xy} such that $\Phi_{xy} = U_{xy} + V_{xy}$ and in equilibrium:

(i) Man i in group x achieves utility

$$\tilde{u}_i = \max_{y \in \mathcal{Y}_0} \left(U_{xy} + \varepsilon_{iy} \right)$$

and he matches with some woman whose group y achieves the maximum;

(ii) Woman j in group y achieves utility

$$\tilde{v}_j = \max_{x \in \mathcal{X}_0} \left(V_{xy} + \eta_{xj} \right)$$

and she matches with some man whose group x achieves the maximum.

This result, which will arise as a consequence of Theorem 1 below, also appears in Chiappori, Salanié and Weiss (2012), 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 they have idiosyncratic ε_{iy} 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.

The vectors U_{xy} and V_{xy} depend on all of the primitives of the model (the vector Φ_{xy}, the distributions of the utility shocks ε and η, and the number of groups n and m.) They are only a useful construct, and they should not be interpreted as utilities. As we will see in Section 2.3, there are at least three relevant definitions of utility, and U and V do not measure any of them.

2.2 Identification of discrete choice problems

In this section we deal with the problem of recovering the utilities U_{xy} from the choice probabilities $\mu_{y|x} = \mu_{xy}/n_x$, and we introduce a general methodology to do so based on "generalized entropy," a name which arises from reasons which will soon become clear. In the following, for any (A_{xy}) we denote $\mathbf{A}_{\mathbf{x}\cdot} = (A_{x1}, \ldots, A_{x|\mathcal{Y}|})$ and $\mathbf{A}_{\cdot\mathbf{y}} = (A_{1y}, \ldots, A_{|\mathcal{X}|y})$.

Consider a randomly chosen man in group x. His expected utility (conditional to belonging to this group) is

$$G_x(\mathbf{U}_{\mathbf{x}\cdot}) = \mathbb{E}_{\mathbf{P}_x} \max_{y \in \mathcal{Y}_0} (U_{xy} + \varepsilon_y), \tag{2.2}$$

where we set $U_{x0} = 0$ and the expectation is taken over the random vector $(\varepsilon_0, \ldots, \varepsilon_{|\mathcal{Y}|}) \sim \mathbf{P}_x$. First note that for any two numbers a, b and random variables (ε, η), the derivative of

$E\max(a+\varepsilon, b+\eta)$ with respect to a is simply the probability that $a+\varepsilon$ is larger than $b+\eta$. Applying this to the function G_x, we get

$$\frac{\partial G_x}{\partial U_{xy}}(\mathbf{U}_{\mathbf{x}\cdot}) = \Pr(U_{xy} + \varepsilon_{iy} \geq U_{xz} + \varepsilon_{iz} \text{ for all } z \in \mathcal{Y}_0).$$

But the right-hand side is simply the probability that a man of group x partners with a woman of group y; and therefore, for $x \in \mathcal{X}$, and $y \in \mathcal{Y}_0$

$$\frac{\partial G_x}{\partial U_{xy}}(\mathbf{U}_{\mathbf{x}\cdot}) = \frac{\mu_{xy}}{n_x} = \mu_{y|x}. \tag{2.3}$$

As the expectation of the maximum of linear functions of the (U_{xy}), G_x is a convex function of $\mathbf{U}_{\mathbf{x}\cdot}$. Now consider the function

$$G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = \max_{\tilde{\mathbf{U}}_{\mathbf{x}\cdot}=(\tilde{U}_{x1},\ldots,\tilde{U}_{x|\mathcal{Y}|})} \left(\sum_{y\in\mathcal{Y}} \mu_{y|x}\tilde{U}_{xy} - G_x(\tilde{\mathbf{U}}_{\mathbf{x}\cdot}) \right) \tag{2.4}$$

whenever $\sum_{y\in\mathcal{Y}} \mu_{y|x} \leq 1$, $G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = +\infty$ otherwise. Hence, the domain of G_x^* is the set of $\mu_{\cdot|x}$ which is the vector of choice probabilities of alternatives in $\mathcal{Y}$. Mathematically speaking, G_x^* is the *Legendre-Fenchel transform*, or *convex conjugate* of G_x. Like G_x and for the same reasons, it is a convex function. By the envelope theorem, at the optimum in the definition of G_x^*

$$\frac{\partial G_x^*}{\partial \mu_{y|x}}(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = U_{xy} \tag{2.5}$$

As a consequence, for any $y \in \mathcal{Y}$, U_{xy} is identified from $\boldsymbol{\mu}_{\cdot|\mathbf{x}}$, the observed matching patterns of men of group x. Going back to (2.4), convex duality implies that if $\boldsymbol{\mu}_{\cdot|\mathbf{x}}$ and $\mathbf{U}_{\mathbf{x}\cdot}$ are related by (2.3), then

$$G_x(\mathbf{U}_{\mathbf{x}\cdot}) = \sum_{y\in\mathcal{Y}} \mu_{y|x}U_{xy} - G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}). \tag{2.6}$$

The term $-G_x^*\left(\mu_{\cdot|x}\right)$ is simply the expectation of the utility shock for the preferred alternative associated with systematic probabilities U_{xy} which leads to the choice probabilities $\mu_{\cdot|x}$. Indeed, by first order conditions, the optimal U is such that $\mu_{y|x} = \partial G_x\left(U_{x\cdot}\right)/\partial U_{xy}$, thus U leads to the choice probabilities $\mu_{\cdot|x}$. Hence, letting Y_i^* be the optimal choice of

marital option y by a man of group x, one has

$$G_x(\mathbf{U}_{\mathbf{x}\cdot}) = \mathbb{E}\left[U_{xY_i^*} + \varepsilon_{iY_i^*}\right] = \sum_{y\in\mathcal{Y}} \mu_{y|x} U_{xy} + \mathbb{E}\left[\varepsilon_{iY_i^*}\right],$$

and, making use of (2.6),

$$-G_x^*\left(\boldsymbol{\mu}_{\cdot|\mathbf{x}}\right) = \mathbb{E}\left[\varepsilon_{iY_i^*}\right]. \tag{2.7}$$

We now provide a useful characterization of $-G_x^*\left(\mu_{\cdot|x}\right)$ using Optimal Transport theory, and show that the evaluation of this quantity as well as U_{xy} can be reformulated as an adjacent optimal matching problem.

Proposition 2. *(**General identification of the systematic surpluses**) Let $\mathcal{M}\left(\mu_{\cdot|x}, P_x\right)$ the set of probability distributions π of the random joint vector (Y,ε), where $Y \sim \mu_{\cdot|x}$ is a random element of $\mathcal{Y}_0$, and $\varepsilon \sim P_x$ is a random vector of $\mathbb{R}^{\mathcal{Y}_0}$. For $e \in \mathbb{R}^{\mathcal{Y}_0}$ and $y \in \mathcal{Y}_0$, let*

$$\Phi^h(y,e) = e_y.$$

Then $-G_x^*(\mu_{\cdot|x})$ is the value of the optimal matching problem between distribution $\mu_{\cdot|x}$ of Y and distribution P_x of ε, when the surplus is Φ^h. That is,

$$-G_x^*(\mu_{\cdot|x}) = \max_{\pi\in\mathcal{M}\left(\mu_{\cdot|x},P_x\right)} \mathbb{E}_\pi\left[\Phi^h(Y,\varepsilon)\right]. \tag{2.8}$$

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

Elaborating on this idea in the context of dynamic discrete games, Chiong, Galichon and Shum (2013) propose in ongoing work to discretize the distribution of ε and solve for the resulting linear program in order to identify the systematic part of the utilities.

2.3 Social surplus and its individual breakdown

We first give an intuitive derivation of our main result, Theorem 1 below. We define H_y similarly as G_x: a randomly chosen woman of group y expects to get utility

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

and the social surplus $\mathcal{W}$ is simply the sum of the expected utilities of all groups of men and women:

$$\mathcal{W} = \sum_{x \in \mathcal{X}} n_x G_x(\mathbf{U}_{\mathbf{x}\cdot}) + \sum_{y \in \mathcal{Y}} m_y H_y(\mathbf{V}_{\cdot \mathbf{y}}),$$

but by identity (2.6), we get

$$G_x(\mathbf{U}_{\mathbf{x}\cdot}) = \sum_{y \in \mathcal{Y}} \mu_{y|x} U_{xy} - G_x^* \left(\boldsymbol{\mu}_{\cdot|\mathbf{x}} \right) \text{ and } H_y(\mathbf{V}_{\cdot \mathbf{y}}) = \sum_{x \in \mathcal{X}} \mu_{x|y} V_{xy} - H_y^*(\boldsymbol{\mu}_{\cdot|\mathbf{y}}),$$

so summing over the total number of men and women, and using $U_{xy} + V_{xy} = \Phi_{xy}$, and defining

$$\mathcal{E}(\mu) := \sum_{x \in \mathcal{X}} n_x G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) + \sum_{y \in \mathcal{Y}} m_y H_y^*(\boldsymbol{\mu}_{\cdot|\mathbf{y}}), \tag{2.9}$$

we get an expression for the value of the total surplus:

$$\mathcal{W} = \sum_{x \in \mathcal{X}} n_x \underbrace{G_x(\mathbf{U}_{\mathbf{x}\cdot})}_{u_x} + \sum_{y \in \mathcal{Y}} m_y \underbrace{H_y(\mathbf{V}_{\cdot \mathbf{y}})}_{v_y} = \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \mu_{xy} \Phi_{xy} - \mathcal{E}(\mu).$$

The first part of this expression explains how the total surplus $\mathcal{W}$ is broken down at the individual level: the average expected equilibrium utility of men in group x is $u_x = G_x(\mathbf{U}_{\mathbf{x}\cdot})$, and similarly for women. The second part of this expression explains how the total surplus is broken down at the level of the couples. We turn this into a formal statement, which is proved in Appendix A.

Theorem 1. *(**Social and Individual Surpluses**) Under Assumptions 1, 2 and 3, the following holds:*

(i) the optimal matching μ maximizes the social gain over all feasible matchings $\mu \in \mathcal{M}$, that is

$$\mathcal{W}(\Phi, n, m) = \max_{\mu \in \mathcal{M}} \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \mu_{xy} \Phi_{xy} - \mathcal{E}(\mu). \tag{2.10}$$

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

$$\begin{aligned} \mathcal{W}(\Phi, n, m) \quad = \quad & \min_{U,V} \sum_{x \in \mathcal{X}} n_x G_x(U_{x\cdot}) + \sum_{y \in \mathcal{Y}} m_y H_y(V_{\cdot y}) \\ & s.t.\ U_{xy} + V_{xy} = \Phi_{xy}. \end{aligned} \tag{2.11}$$

(ii) 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 the U_{xy}'s are solution to (2.11).

(iii) The average expected utility of the men of group x is $u_x = G_x(\mathbf{U}_{\mathbf{x}\cdot})$.

(iv) Parts (ii) and (iii) transpose to the other side of the market with the obvious changes.

The right-hand side of equation (2.10) gives the value of the social surplus when the matching patterns are (μ_{xy}). The first term $\sum_{xy} \mu_{xy}\Phi_{xy}$ reflects "group preferences": if groups x and y generate more surplus when matched, then they should be matched with higher probability. On the other hand, the second and the third terms reflect 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 a higher probability. In the one-dimensional Beckerian example, a higher x or y could reflect higher education. If the marital surplus is complementary in the educations of the two partners, Φ_{xy} is supermodular and the first term is maximized when matching partners with similar education levels (as far as feasibility constraints allow.) But because of the dispersion of marital surplus that comes from the ε and η terms, it will be optimal to have some marriages between dissimilar partners.

To interpret the formula, start with the case when unobserved heterogeneity is dwarfed by variation due to observable characteristics: $\tilde{\Phi}_{ij} \simeq \Phi_{xy}$ if $x_i = x$ and $y_j = y$. Then we know that the observed matching μ must maximize the value in (2.1); but this is precisely what the more complicated expression in μ above boils down to if we scale up the values of Φ to infinity. If on the other hand data is so poor that unobserved heterogeneity dominates ($\Phi \simeq 0$), 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. For this reason, we call it the *generalized entropy* of the matching. In the intermediate case in which 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 (as in (2.1)) against matching on unobserved characteristics, as measured by the generalized entropy terms in $\mathcal{E}(\mu)$.

Theorem 1 is an equilibrium characterization result, which allows the analyst to predict the joint and individual shares of surplus at equilibrium. As we show in section 3, this can be done in closed form in a number of important cases. Note that there are three measures of surplus:

- ex ante utility u_x is the expected utility of a man, conditional on his being in group x. Part (iii) gives a very simple formula to compute it;
- ex interim utility, if we also condition on this man marrying a woman of group y, is

$$\mathbb{E}\left[U_{xy} + \varepsilon_{iy} | U_{xy} + \varepsilon_{iy} \geq U_{xz} + \varepsilon_{iz} \text{ for all } z \in \mathcal{Y}\right];$$

  this can be computed since the U_{xz}'s are identified from part (ii), although it may require simulation for general distributions;
- ex post utility $U_{xy} + \varepsilon_{iy}$ for these men, whose distribution can also be simulated.

In the special multinomial logit case studied by Choo and Siow, ex post utility is distributed as type I extreme value with mean $(-\log \frac{\mu_{x0}}{n_x})$, which is the common value u_x of ex ante and ex interim utility; but the three definitions give different results in general, as observed by de Palma and Kilani (2007).

2.4 Identification of matching surplus

There are two readings of Theorem 1, which are mathematically equivalent, but have very different practical purposes: one may use it to obtain the expression of μ as a function of Φ: this is an "equilibrium characterization" point of view. Conversely, one may use it to obtain the expression of Φ as a function of μ: this is an "identification" point of view. Our next

result, Theorem 2, illustrates the mathematical duality between the two points of view and applies it for identification purposes. Indeed, relations (2.12) allow to express μ as a function of U and V ("equilibrium characterization" point of view); they invert into relations (2.13) which allow to express U and V (and thus Φ) as a function of μ ("identification" point of view).

Note that the constraints associated to $\mu \in \mathcal{M}$ in (2.10) do not bind in the many datasets in which there are no empty cells: then $\mu_{xy} > 0$ for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and $\sum_{x \in \mathcal{X}} \mu_{xy} < n_x$, $\sum_{y \in \mathcal{Y}} \mu_{xy} < m_y$. In other words, μ then belongs to the interior of $\mathcal{M}$. It is easy to see that this must hold under the following assumption:

Assumption 4 (Full support)**.** *The distributions* $\mathbf{P}_x$ *and* $\mathbf{Q}_y$ *all have full support.*

Assumption 4 of course holds for the Choo and Siow model. It can be relaxed in the obvious way: all that matters is that the supports of the distributions are wide enough relative to the magnitude of the variations in the matching surplus. It is not essential to our approach; in fact, one of our leading examples in section 3 violates it. But it allows us to obtain very clean formulæ, as stated in the following theorem:

Theorem 2. *Under Assumptions 1, 2, 3 and 4:*

(i) U_{xy} *is identified by the equivalent set of relations*

$$\mu_{y|x} = \frac{\partial G_x}{\partial U_{xy}}(\mathbf{U}_{\mathbf{x}\cdot}) \text{ for } y \in \mathcal{Y}, \text{ or equivalently} \tag{2.12}$$

$$U_{xy} = \frac{\partial G_x^*}{\partial \mu_{y|x}}\left(\boldsymbol{\mu}_{\cdot|\mathbf{x}}\right) \text{ for } y \in \mathcal{Y}. \tag{2.13}$$

(ii) As a result, Φ_{xy} *is identified by*

$$\Phi_{xy} = \frac{\partial G_x^*}{\partial \mu_{y|x}}\left(\boldsymbol{\mu}_{\cdot|\mathbf{x}}\right) + \frac{\partial H_y^*}{\partial \mu_{x|y}}\left(\boldsymbol{\mu}_{\cdot|\mathbf{y}}\right), \tag{2.14}$$

that is

$$\Phi_{xy} = \frac{\partial \mathcal{E}}{\partial \mu_{xy}}(\mu). \tag{2.15}$$

Note that since the functions G_x^* and H_y^* are convex, they are differentiable almost everywhere—and under Assumption 4 they actually are differentiable everywhere.

The previous result does not assume that transfers are observed. When they are, the systematic parts of pre-transfer utilities (α, γ) are also observed. This case is unlikely to occur in the context of family economics, where the econometrician typically does not observe transfers between partners, but it is typically the case in other settings where matching theory has been successfully applied, as the CEO compensation literature, for instance, where the compensation amount is often available. In that case, $U_{xy} = \alpha_{xy} + \tau_{xy}$ and $V_{xy} = \gamma_{xy} - \tau_{xy}$, so the conjunction of the observation of τ along with the identification of $\Phi = U + V$ ensures there is a sufficient number of equations to identify α and γ separately. We state the following corollary:

Corollary 1. *Under Assumptions 1, 2, 3 and 4, denote (α, γ) the systematic parts of pre-transfer utilities and τ the transfers as in Section 1. Then α_{xy} and γ_{xy} are identified by*

$$\alpha_{xy} = \frac{\partial G_x^*}{\partial \mu_{y|x}}\left(\boldsymbol{\mu}_{\cdot|\mathbf{x}}\right) - \tau_{xy} \text{ and } \gamma_{xy} = \frac{\partial H_y^*}{\partial \mu_{x|y}}\left(\boldsymbol{\mu}_{\cdot|\mathbf{y}}\right) + \tau_{xy}.$$

Therefore if transfers τ_{xy} are observed, both pre-transfer utilities α_{xy} and γ_{xy} are also identified.

As a result of Proposition 2, all of the quantities in Theorem 1 can be computed by solving simple linear programming problems. This makes identification and estimation feasible in practice.

2.5 Comparative statics

In this section, we use the results of Theorem 1 to show that the comparative statics results of Decker et al. (2012) extend to our generalized framework. From the results of Section 2.3,

recall that $\mathcal{W}(\Phi, n, m)$ is given by the dual expressions

$$\mathcal{W}(\Phi, n, m) = \max_{\mu \in \mathcal{M}(n,m)} \sum_{xy} \mu_{xy} \Phi_{xy} - \mathcal{E}(\mu), \text{ and} \tag{2.16}$$

$$\mathcal{W}(\Phi, n, m) = \min_{U_{xy}+V_{xy}=\Phi_{xy}} \sum n_x G_x(U_{xy}) + \sum m_y H_y(V_{xy}) \tag{2.17}$$

As a result, note that by (2.16), $\mathcal{W}$ is a convex function of Φ, and by (2.17) it is a concave function of (n, m). By the envelope theorem in (2.16) and in (2.17), we get respectively

$$\frac{\partial \mathcal{W}}{\partial \Phi_{xy}} = \mu_{xy} \text{ and}$$

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

A second differentiation of $\partial \mathcal{W}/\partial n_x$ with respect to $n_{x'}$ yields

$$\frac{\partial u_x}{\partial n_{x'}} = \frac{\partial^2 \mathcal{W}}{\partial n_x \partial n_{x'}} = \frac{\partial u_{x'}}{\partial n_x} \tag{2.18}$$

(and similarly $\partial u_x/\partial m_y = \partial v_y/\partial n_x$ and $\partial v_y/\partial m_{y'} = \partial v_{y'}/\partial m_y$), which is the "unexpected symmetry" result proven by Decker et al. (2012), Theorem 2, for the multinomial logit Choo and Siow model: the variation in the systematic part of the surplus of individual of group x when the number of individuals of group x' varies by one unit equals the variation in the systematic part of the surplus of individual of group x' when the number of individuals of group x varies by one unit. Formula (2.18) shows that the result is valid quite generally in the framework of the present paper. The fact that $\mathcal{W}$ is a concave function of (n, m) implies that the matrix $\partial u_x/\partial n_{x'}$ is semidefinite negative; in particular, it implies that $\partial u_x/\partial n_x \leq 0$, which means that increasing the number of individuals of a given group cannot increase the individual welfare of individuals of this group.

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

$$\frac{\partial \mu_{xy}}{\partial n_{x'}} = \frac{\partial^2 \mathcal{W}}{\partial n_{x'} \partial \Phi_{xy}} = \frac{\partial u_{x'}}{\partial \Phi_{xy}} \tag{2.19}$$

which is proven (again in the case of the multinomial logit Choo and Siow model) in Decker et al. (2012), section 3. This means that 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

number of individuals of group x' on the number 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 number of men with a college degree causes the number 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, and therefore their share of surplus will decrease, hence $\partial u_{x'}/\partial\Phi_{xy} < 0$.

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

$$\frac{\partial\mu_{xy}}{\partial\Phi_{x'y'}} = \frac{\partial^2\mathcal{W}}{\partial\Phi_{xy}\partial\Phi_{x'y'}} = \frac{\partial\mu_{x'y'}}{\partial\Phi_{xy}}. \tag{2.20}$$

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.

3 Examples

3.1 A bestiary of models

While Proposition 2 and Theorem 1 provide a general way of computing surplus and utilities, they can often be derived in closed form. In all formulæ below, the proportions and numbers 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. In this section we will maintain Assumptions 1, 2.

Our first example is the classical multinomial logit model of Choo and Siow, which is

obtained as a particular case of the results in Section 2 when the $\mathbf{P}_x$ and $\mathbf{Q}_y$ distributions are iid standard type I extreme value:

Example 1 (Choo and Siow)**.** *Assume that* $\mathbf{P}_x$ *and* $\mathbf{Q}_y$ *are the distributions of i.i.d. standard type I extreme value random variables. Then*

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

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

$$\mathcal{E}(\mu)=\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}_0}}\mu_{xy}\log\mu_{y|x}+\sum_{\substack{y\in\mathcal{Y}\\ x\in\mathcal{X}_0}}\mu_{xy}\log\mu_{x|y}, \tag{3.2}$$

and surplus and matching patterns are linked by

$$\Phi_{xy}=2\log\mu_{xy}-\log\mu_{x0}-\log\mu_{0y}, \tag{3.3}$$

which is Choo and Siow's (2006) identification result. See Appendix B.1 for details.

Note that as announced after Theorem 1, the generalized entropy $\mathcal{E}$ boils down here to the usual definition of entropy. The multinomial logit Choo and Siow model is the simplest example which fits into McFadden's Generalized Extreme Value (GEV) framework, recalled in Appendix B. This framework includes most specifications used in classical discrete choice models. A simple variant of the Choo–Siow model is the heteroskedastic model considered by Chiappori, Salanié and Weiss (2012); it allows the scale parameters of the type I extreme value distributions to vary across genders or groups. Then $\mathbf{P}_x$ has a scale parameter σ_x and $\mathbf{Q}_y$ has a scale parameter τ_y; the expected utilities are $u_x = -\sigma_x\log\mu_{0|x}$ and $v_y = -\tau_y\log\mu_{0|y}$, and the general identification formula gives

$$\Phi_{xy}=(\sigma_x+\tau_y)\log\mu_{xy}-\sigma_x\log\mu_{x0}-\tau_y\log\mu_{0y}. \tag{3.4}$$

As a more complex example of a GEV distribution, we turn to a nested logit model.

Example 2 (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 $\mathbf{P}_x$ have cdf*

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

This is a particular case of the Generalized Extreme Value (GEV) framework described in Appendix B, with g defined there given by $g(z) = z_0 + \sum_s \left(\sum_e z_{se}^{1/\sigma_s}\right)^{\sigma_s}$. The numbers $1/\sigma_s$ describe the correlation in the surplus generated with partners of different education levels within social group s. Then (dropping the x indices for notational simplicity, so that for instance μ_s denotes the number of matches with women in social group s)

$$\begin{aligned} G(\mathbf{U}.) &= \log\left(1 + \sum_s \left(\sum_e \exp(U_{se}/\sigma_s)\right)^{\sigma_s}\right), \text{ and} \\ G^*(\boldsymbol{\mu}.) &= \mu_0 \log \mu_0 + \sum_s (1-\sigma_s)\,\mu_s \log \mu_s + \sum_s \sigma_s \sum_e \mu_{se} \log \mu_{se}. \end{aligned}$$

where μ_0 is again defined in (3.1). As in Example 1, the expected utility is $u = -\log \mu_0$.

If the heterogeneity structure is the same for all men and all women (with possibly different dispersion parameters σ for men and τ for women), then the expressions of $\mathcal{E}(\mu)$ and $\mathcal{W}(\mu)$ can easily be obtained. The social surplus from a match between a man of group $x = (s, e)$ and a woman of group $y = (s', e')$ is identified by

$$\Phi_{xy} = \log \frac{\mu_{xy}^{\sigma_{s'}^x + \tau_s^y}\, \mu_{x,s'}^{1-\sigma_{s'}^x}\, \mu_{s,y}^{1-\tau_s^y}}{\mu_{x0}\mu_{0y}}.$$

See Appendix B.2 for details.

Note that we recover the results of Example 1 when all σ parameters equal 1; also, if there is only one possible social status, then we recover the heteroskedastic model.

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

Example 3 (A mixture of logits)**.** *Take nonnegative numbers β_k^x such that $\sum_{k=1}^K \beta_k^x = 1$. Let the distribution $\mathbf{P}_x$ be a mixture of iid type I extreme value distributions with scale parameters σ_k^x, weighted by the probabilities β_k^x. Then*

$$G_x(\mathbf{U}_{\mathbf{x}\cdot}) = \sum_{k=1}^{K} \beta_k^x \sigma_k^x \log\left(1 + \sum_{y\in\mathcal{Y}} e^{U_{xy}/\sigma_k^x}\right) \tag{3.5}$$

and

$$G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = \min_{\sum_{k=1}^K \mu_y^k = \mu_{y|x}} \sum_{k=1}^{K} \sigma_k^x \left(\mu_0^k \log\frac{\mu_0^k}{\beta_k^x} + \sum_{y\in\mathcal{Y}} \mu_y^k \log\frac{\mu_y^k}{\beta_k^x}\right). \tag{3.6}$$

Then U_{xy} is given by $U_{xy} = \sigma_k^x \log(\mu_y^k/\mu_0^k)$, where $\left(\mu_y^k\right)$ is the minimizer of (3.6). See Appendix B.3 for details.

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[2]. Closer to our framework, hedonic models also build on idiosyncratic preferences for observed characteristics, on both sides of a match[3]. Our setup allows for such specifications. Assume for instance that men of group x care for a vector of observed characteristics of partners $\zeta_x(y)$, 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 for instance take $\mathbf{P}_x$ to be the distribution of $\boldsymbol{\zeta}_x(y) \cdot \boldsymbol{\varepsilon}_i$.

We investigate a particular case of this specification in the next example: the Random Scalar Coefficient (RSC) model, where the dimension of $\zeta_x(y)$ and $\boldsymbol{\varepsilon}_i$ is one. As we argue below, this assumption much simplifies the computations. Assuming further that the distribution of $\boldsymbol{\varepsilon}_i$ is uniform, one is led to what we call the Random Uniform Scalar Coefficient Model (RUSC). This last model has one additional advantage: it yields simple closed-form expressions, even though it does not belong to the Generalized Extreme Value (GEV) class.

[2]See the literature surveyed in Ackerberg et al (2007) or Reiss and Wolak (2007).

[3]See Ekeland et al (2004) and Heckman et al (2010).

Example 4 (Random [Uniform] Scalar Coefficient (RSC/RUSC) models)**.** *Assume that for each man i in group x,*

$$\varepsilon_{iy} = \varepsilon_i \zeta_x(y),$$

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_ε (which could also depend on x.) We call this model the Random Scalar Coefficient (RSC) model; and we show in Appendix B.4 that the entropy is

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

where $\bar{e}_x(y)$ is the expected value of ε on the interval $[a, b]$ defined by

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

and $\bar{f}_y(x)$ is defined similarly.

Assuming further that the ε_i are uniformly distributed over $[0, 1]$, we call this model the Random Uniform Scalar Coefficient (RUSC) model. In this case, simpler formulæ can be given. For any $x \in \mathcal{X}$, let S^x be the square matrix with elements $S^x_{yy'} = \max(\zeta_x(y), \zeta_x(y'))$ for $y, y' \in \mathcal{Y}_0$. Define T^x by $T^x_{yy'} = S^x_{y0} + S^x_{0y'} - S^x_{yy'} - S^x_{00}$, and let $\sigma^x_y = S^x_{00} - S^x_{y0}$.

Then G^*_x is quadratic with respect to $\boldsymbol{\mu}_{\cdot|\mathbf{x}}$:

$$G^*_x(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = \frac{1}{2}(\boldsymbol{\mu}_{\cdot|\mathbf{x}}{}' T^x \boldsymbol{\mu}_{\cdot|\mathbf{x}} + 2\sigma^x . \boldsymbol{\mu}_{\cdot|\mathbf{x}} - S^x_{00}).$$

If we now assume that preferences have such a structure for every group x of men and for every group y of women (so that $\eta_{xj} = \eta_j \xi_y(x)$), then the generalized entropy is quadratic in μ:

$$\mathcal{E}(\mu) = \frac{1}{2}(\mu' A \mu + 2B\mu + c),$$

where the expressions for A, B and c are given by (B.4)–(B.5) in Appendix B.4. As a consequence, the optimal matching solves a simple quadratic problem. See Appendix B.4 for details.

The structure of heterogeneity in the RUSC/RSC models is reminiscent of the one investigated in Ekeland et al. (2004) and Heckman et al. (2010), with continuous observed characteristics. In Ekeland et al. (2004), the distribution of the ε_i's is unknown, but identified from a separability assumption on the marginal willingness to pay. In contrast, closer to our paper is Heckman et al. (2010), where the distribution of the ε_i's is fixed and identification is obtained from a quantile transformation approach; however, in this setting, there is heterogeneity only on one side of the market.

3.2 Discussion

In spite of all its insights, the Choo-Siow multinomial logit framework carries a number of strong assumptions. This calls for caution when basing conclusions on it. To illustrate this point, we would like to show that some of the very strong conclusions are in fact dependent on the distributional assumptions made on the unobserved heterogeneity. The interest of our general framework is to show that the expected utilities can be a much richer function of observed matching patterns than in Choo and Siow's multinomial logit model.

- **Spillover effects**. Choo and Siow's original motivation was to generate a "marriage function with spillover effects" which takes care of substitution effects in a coherent way, in contrast with the previous demographic literature on marriage. This "matching function" is the map which takes the number of groups n_x and m_y as an input and returns the number of marriages μ_{xy} as an output. The "substitution effects" are expressed by constraint (3.1): if there are more marriages between group x and group y, there will be mechanically fewer marriages between groups x and y', and less marriages between groups x' and y. The explicit derivations in the above examples allow us to compare the influence that the numerical values of μ have on the surplus estimator Φ_{xy}, across the different models. This can be done by analyzing the term $\partial\Phi_{xy}/\partial\mu_{x'y'}$. In the case of Choo and Siow,

$$\Phi_{xy} = \log \frac{\mu_{xy}^2}{\left(n_x - \sum_{y' \in \mathcal{Y}} \mu_{xy'}\right)\left(m_y - \sum_{x' \in \mathcal{X}} \mu_{x'y}\right)}$$

so that Φ_{xy} is a function of μ_{xy}^2 and $\sum_{y' \neq y} \mu_{xy'}$ and $\sum_{x' \neq x} \mu_{x'y}$ only. Therefore if $y' \neq y'' \neq y$,

$$\frac{\partial \Phi_{xy}}{\partial \mu_{xy'}} = \frac{\partial \Phi_{xy}}{\partial \mu_{xy''}}. \tag{3.7}$$

To interpret this, start from a given matching μ which is rationalized by some surplus Φ, and suppose that a single man of group x marries a single woman of group $y' \neq y$. Then (3.7) tells us that our estimator of the surplus Φ_{xy} should change by exactly the same amount as if the single woman had been of any other group $y'' \neq y$, which seems counterintuitive.

This problematic finding comes from the assumption of independence of irrelevant alternatives (IIA) in the Choo-Siow model, just as restrictions on cross-elasticities obtained in multinomial logit models. The RUSC model is much better able to capture variation in cross-elasticities: the derivations in Appendix B.4 show that the effect of changes in observed matching patterns on the estimated surplus $\partial \mu_{xy} / \partial \Phi_{x'y'}$ allows for much richer effects than (3.7).

- **Comparative statics**. Interestingly, the comparative statics discussed in Section 2.5 have explicit expressions in some cases. Take relation (2.18) for instance, which expresses that the derivative of the expected utility u_x of men of group x with respect to the number of men of group x' coincides with the derivative of $u_{x'}$ with respect to n_x. For the Choo and Siow multinomial logit model investigated in Decker et al. (2012), this derivative is a complicated term. In the RUSC model of Example 4, the derivative is given by (B.7):

$$\frac{\partial u_x}{\partial n_{x'}} = \frac{\partial u_{x'}}{\partial n_x} = \frac{1}{n_x^2 n_{x'}^2} \mu' R^{xx'} \mu$$

where $R^{xx'}$ is a matrix whose expression is given in (B.8) of Appendix B.4. Similarly, (2.19) and (2.20) are explicit and given respectively by (B.9) and (B.11).

4 Parametric Inference

Theorem 1 shows that, given a specification of the distribution of the unobserved heterogeneities $\mathbf{P}_x$ and $\mathbf{Q}_y$, any model that satisfies assumptions 1, 2, and 3 is nonparametrically identified from the observation of a single market. There is therefore no way to test separability using only data on one market. When multiple markets with identical Φ_{xy}, $\mathbf{P}_x$ and $\mathbf{Q}_y$ are observed, then the model is nonparametrically overidentified given a fixed specification of $\mathbf{P}_x$ and $\mathbf{Q}_y$. The flexibility allowed by Assumption 3 can then be used to infer information about these distributions.

In the present paper, we are assuming that a single market is being observed. While the formula in Theorem 1(i) gives a straightforward nonparametric estimator of the systematic surplus function Φ, with multiple surplus-relevant observable groups it will be very unreliable. Even our toy education/income example of Section 1.1 already has $4n_R^2$ cells; and realistic applications will require many more. In addition, we do not know the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$. Both of these remarks point towards the need to specify a parametric model in most applications. Such a model would be described by a family of joint surplus functions Φ_{xy}^λ and distributions $\mathbf{P}_x^\lambda$ and $\mathbf{Q}_y^\lambda$ for λ in some finite-dimensional parameter space Λ.

We observe a sample of $\hat{N}_{ind}$ individuals; $\hat{N}_{ind} = \sum_x \hat{N}_x + \sum_y \hat{M}_y$, where $\hat{n}_x$ (resp. $\hat{m}_y$) denotes the number of men of group x (resp. women of group y) in the sample. We let $\hat{n}_x = \hat{N}_x / \hat{N}_{ind}$ and $\hat{m}_y = \hat{M}_y / \hat{N}_{ind}$ the rescaled number of individuals. Let $\hat{\mu}$ the observed matching; we assume that the data was generated by the parametric model above, with parameter vector λ_0.

Recall the expression of the social surplus:

$$\mathcal{W}(\Phi^\lambda, \hat{n}, \hat{m}) = \max_{\mu \in \mathcal{M}(\hat{n}, \hat{m})} \{ \sum_{x,y} \mu_{xy} \Phi_{xy}^\lambda - \mathcal{E}^\lambda(\mu) \}$$

Let μ^λ be the optimal matching. Of course, computing μ^λ is a crucial issue. We will show in Section 6 how it can be computed, in some cases very efficiently. For now we focus on statistical inference on λ. We propose two methods: a very general Maximum Likelihood

method, and a more restrictive moment-based method.

4.1 Trade-off between observable and unobservable dimensions

In Theorem 2, we have kept fixed distributions for the unobservable heterogeneity terms $\mathbf{P}_x$ and $\mathbf{Q}_y$, and we have answered with formula (2.15) the question raised at the end of Section 1.2: how can we achieve identification of Φ_{xy} (an array of $|\mathcal{X}| \times |\mathcal{Y}|$ unknowns) given the observation of μ_{xy} (an array of $|\mathcal{X}| \times |\mathcal{Y}|$ observations)? Of course, fixing the distribution of the unobserved heterogeneity terms is a strong assumption, while we do not require full nonparametric identification of Φ. If we are content with a parametric form of Φ whose parameter has dimensionality lower than $|\mathcal{X}| \times |\mathcal{Y}|$, we get degrees of freedom which we can use for inference on the distributions $\mathbf{P}_x$ and $\mathbf{Q}_y$, appropriately parameterized.

For example, if $\mathcal{X}$ and $\mathcal{Y}$ are finite subsets of $\mathbb{R}^d$, we could have a semiparametric specification, in the spirit of Ekeland et al. (2004) $\Phi(x, y) = \phi_1(y) + y'\phi_2(x)$, where ϕ_1 is a function from $\mathcal{Y}$ to $\mathbb{R}$, and ϕ_2 is a function from $\mathcal{X}$ to $\mathbb{R}^d$. With this assumption, Φ would become an object of dimension $|\mathcal{Y}| + d \times |\mathcal{X}|$, instead of $|\mathcal{X}| \times |\mathcal{Y}|$ in the nonparametric case. The degrees of freedom gained by imposing the semi-parametric specification of Φ can be used for inference purpose on the distribution of the unobservable heterogeneity terms.

4.2 Maximum Likelihood estimation

In this section we will use Conditional Maximum Likelihood (CML) estimation, where we condition on the observed margins $\hat{n}_x$ and $\hat{m}_y$. For each man of group x, the log-likelihood of marital choice is $\sum_{y \in \mathcal{Y}_0} (\hat{\mu}_{xy}/\hat{n}_x) \log(\mu^{\lambda}_{xy}/\hat{n}_x)$, and a similar expression holds for each woman of group y. Under Assumptions 1, 2 and 3,, the choice of each individual is stochastic in that it depends on his vector of unobserved heterogeneity, and these vectors are independent across men and women. Hence the log-likelihood of the sample is the sum of the individual

log-likelihood elements:

$$\begin{aligned}\log L(\lambda) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}_0}\hat{\mu}_{xy}\log\frac{\mu^{\lambda}_{xy}}{\hat{n}_x} + \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}_0}\hat{\mu}_{xy}\log\frac{\mu^{\lambda}_{xy}}{\hat{m}_y} \\ &= 2\sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}}\hat{\mu}_{xy}\log\frac{\mu^{\lambda}_{xy}}{\sqrt{\hat{n}_x\hat{m}_y}} + \sum_{x\in\mathcal{X}}\hat{\mu}_{x0}\log\frac{\mu^{\lambda}_{x0}}{\hat{n}_x} + \sum_{y\in\mathcal{Y}}\hat{\mu}_{0y}\log\frac{\mu^{\lambda}_{0y}}{\hat{m}_y}.\end{aligned} \tag{4.1}$$

The Conditional Maximum Likelihood Estimator $\hat{\lambda}^{MLE}$ given by the maximization of $\log L$ is consistent, asymptotically normal and asymptotically efficient under the usual set of assumptions.

Example 2 continued. *In the Nested Logit model of Example 2, where the group of men and women are respectively (s_x, e_x) and (s_y, e_y), one can take $\sigma^{s_x e_x}_{s_y}$ and $\sigma^{s_y,e_y}_{s_x}$ as parameters. Assume that there are N_s social categories and N_e classes of education. There are $N_s^2 \times N_e^2$ equations, so one can parameterize the surplus function Φ^{θ} by a parameter θ of dimension less than or equal to $N_s^2 \times N_e^2 - 2N_s^2 \times N_e$. Letting $\lambda = \left(\sigma^{s_x e_x}_{s_y}, \sigma^{s_y,e_y}_{s_x}, \theta\right)$, μ^{λ} is the solution in M to the system of equations*

$$\Phi^{\theta}_{xy} = \log\frac{\mu_{xy}^{\sigma^x_{s'}+\tau^y_s}\,\mu_{x,s'}^{1-\sigma^x_{s'}}\,\mu_{s,y}^{1-\tau^y_s}}{(n_x - \sum_y \mu_{xy})(m_y - \sum_x \mu_{xy})},\ \forall x\in\mathcal{X},\ y\in\mathcal{Y}$$

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

In some cases, the expression of the likelihood μ^{λ} can be obtained in closed form. This is the case in the Random Uniform Scalar Coefficient model:

Example 4 continued. *Assume that the data generating process is the RUSC model of Example 4. We parameterize Φ, $\zeta_x(.)$, and $\zeta_y(.)$ by a parameter vector $\lambda \in \mathbb{R}^K$, hence parameterizing S and T and thus A and B. If the solution is interior, then the optimal matching is given by $\mu^{\lambda} = (A^{\lambda})^{-1}(\mathbf{\Phi}^{\lambda} - B^{\lambda})$, and the log-likelihood can be deduced by (4.1).*

Maximum likelihood estimation has many advantages: (i) it allows for joint parametric estimation of the surplus function and of the unobserved heterogeneities; (ii) it enjoys desirable statistical properties in terms of statistical efficiency; (iii) its asymptotic properties are well-known. However, there is no guarantee that the log-likelihood shall be a concave function in general, and hence maximization of the likelihood may lead to practical problems in some situations. In some of these cases, an alternative method, based on moments, is available. This method is detailed in the next section.

4.3 Moment-based estimation: The Linear Model

The previous analysis involving maximum likelihood has one shortcoming: there is no guarantee that the log-likelihood is a convex function, and so, if no proper care is taken, the maximization of the log-likelihood may be trapped in a local maximum. Under additional assumptions, we shall describe a method based on moments which is computationally very efficient.

In this section we shall impose two strong assumptions. First, we shall assume that the distribution of the unobservable heterogeneity is known and fixed, so that we won't parameterize the distribution of the unobservable heterogeneity. Next, we shall assume that the surplus can be linearly parameterized by

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

where the parameter $\lambda \in \mathbb{R}^K$ and the sign of each λ_k is unrestricted, and where ϕ_{xy}^1,..., ϕ_{xy}^K are K (known) *basis surplus vectors* which are linearly independent: no linear combination of these vectors is identically equal to zero. We call this specification the "linear model" because the surplus depends linearly on the parameters. Quite obviously, if the set of basis surplus vectors is large enough, this specification covers the full set without restriction; however, parsimony is often valuable in applications. Note that the linearity of Φ^{λ} with respect to λ implies that $\mathcal{W}\left(\Phi^{\lambda}, n, m\right)$ is convex with respect to λ.

Return to the education/income example of Section 1.1, where $x, y = (E, R)$ consists of education and income; education takes values $E \in \{D, G\}$ (dropout or graduate), and income class R takes values 1 to n_R. Then we could for instance assume that a match between man i and woman j creates a surplus that depends on whether partners are matched on both education and income dimensions. The corresponding specification would have basis functions like $\mathbf{1}(E_x = E_y = e)$ and $\mathbf{1}(R_x = R_y = r)$, along with "one-sided" basis functions to account for different probabilities of marrying: $\mathbf{1}(R_x = r, E_x = e)$ and $\mathbf{1}(R_y = r, E_y = e)$, so that

$$\Phi_{xy}^{\lambda} = \sum_{e} \lambda_e \mathbf{1}(E_x = E_y = e) + \sum_{r} \lambda_r \mathbf{1}(R_x = R_y = r)$$
$$+ \sum_{r'e'} \lambda_{r'e'} \mathbf{1}(R_x = r', E_x = e') + \sum_{r''e''} \lambda_{r''e''} \mathbf{1}(R_y = r'', E_y = e'')$$

This specification only has $(5n_R + 2)$ parameters, to be compared to $4n_R^2$ for an unrestricted specification (where for instance the matching surplus of a man in income class 3 with a woman in income class 2 would also depend on both of their education levels). With more, multi-valued criteria the reduction in dimensionality would be much larger. It is clear that the relative importance of the λ's reflects the relative importance of the criteria. They indicate how large the systematic preference for complementarity of incomes of partners is relative to the preference for complementarity in educations.

For any feasible matching μ, we define the associated *comoments*

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

In the case of the education/income example above, the empirical comoment associated to basis function $\mathbf{1}(E_x = E_y = D)$ is $\sum_{x,y} \mu_{xy} \mathbf{1}(E_x = E_y = D)$, which is the number of couples where partners are both dropouts.

The estimator we propose in this section consists in looking for a parameter vector λ which is such that the comoments predicted by the model with parameter value λ coincide

with the empirical comoments. To do this, introduce the *Moment Matching estimator* as the value $\hat{\lambda}^{MM}$ of the parameter vector solution to the following expression

$$\hat{\lambda}^{MM} := \arg\max_{\lambda \in \mathbb{R}^k} \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \hat{\mu}_{xy} \Phi^{\lambda}_{xy} - \mathcal{W}\left(\Phi^{\lambda}, n, m\right), \tag{4.3}$$

whose objective function is concave, because, as mentioned above, $\mathcal{W}\left(\Phi^{\lambda}, n, m\right)$ is convex with respect to λ, and Φ^{λ}_{xy} is linear.

Theorem 3. *Under Assumptions 1, 2 and 3, assume the distributions of the unobserved heterogeneity terms* $\mathbf{P}_x$ *and* $\mathbf{Q}_y$ *are known. Then:*

(i) The Moment Matching estimator is characterized by the fact that the predicted comoments coincide with the observed comoments, that is, equality $C^k(\hat{\mu}) = C^k(\mu^{\lambda})$ *holds for all* k *whenever* $\lambda = \hat{\lambda}^{MM}$.

(ii) Equivalently, the Moment Matching estimator $\hat{\lambda}^{MM}$ *is the vector of Lagrange multipliers of the moment constraints in the program*

$$\mathcal{E}_{\min}(\hat{\mu}) = \min_{\mu \in \mathcal{M}} \left\{ \mathcal{E}(\mu) : C^k(\mu) = C^k(\hat{\mu}), \forall k \right\}. \tag{4.4}$$

Therefore the Moment Matching estimator matches the observed comoments to those that are predicted by the model.

Example 1 continued. *Fix the distributions of the unobservable heterogeneities to be type I extreme value distributed as in the multinomial logit Choo-Siow setting, and assume that surplus function* Φ^{λ}_{xy} *is linearly parameterized by a vector* $\lambda \in \mathbb{R}^K$*, as in (4.2). Then the log-likelihood can be written as*

$$\log L(\lambda) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \hat{\mu}_{xy} \Phi^{\lambda}_{xy} - \mathcal{W}(\lambda). \tag{4.5}$$

Therefore in this setting the Conditional Maximum Likelihood estimator and the Moment Matching estimator are equivalent, that is $\hat{\lambda}^{MM} = \hat{\lambda}^{MLE}$*. They consist in the maximization of the map* $\lambda \to \sum_{k,x,y} \lambda_k \hat{\mu}_{xy} \phi^k_{xy} - \mathcal{W}(\lambda)$*, which is smooth and strictly concave.*

The fact that $\hat{\lambda}^{MM}$ and $\hat{\lambda}^{MLE}$ coincide in the multinomial logit Choo-Siow setting is quite particular to that setting. It is not the case in other models, such as the RUSC model for instance. In fact, the RUSC model is interesting to study as one can obtain an explicit expression of $\hat{\lambda}^{MM}$ in the common case when no cell is empty ($\mu_{xy} > 0$ for all (x, y)):

Example 4 continued. *Assume that the data generating process is the RUSC model of Example 4, where we fix $\zeta_x(.)$, and $\zeta_y(.)$, and where Φ^{λ}_{xy} is linearly parameterized by a vector $\lambda \in \mathbb{R}^K$ as in (4.2). Assume further that all μ's are positive. Then*

$$\mathcal{W}(\lambda) = \frac{1}{2}((\boldsymbol{\phi}.\boldsymbol{\lambda} - B)' A^{-1} (\boldsymbol{\phi}.\boldsymbol{\lambda} - B) - c)$$

where $\boldsymbol{\phi} = \left(\phi^k_{xy}\right)_{xy,k}$ is to be understood as a matrix, and $\boldsymbol{\lambda} = (\lambda_k)_k$ as a vector. As a consequence, the Moment Matching estimator is a simple affine function of the observed comoments: $\hat{\boldsymbol{\lambda}}^{MM} = \left(\boldsymbol{\phi}' A^{-1} \boldsymbol{\phi}\right)^{-1} \left(C(\hat{\mu}) + \boldsymbol{\phi}' A^{-1} B\right)$.

Note that Part (ii) of Theorem 3 is useful to provide a very simple semiparametric specification test. Compare the *actual* value $\mathcal{E}(\hat{\mu})$ of the entropy associated to the empirical distribution to the value $\mathcal{E}_{\min}(\hat{\mu})$ of the program (4.4). By definition of $\mathcal{E}_{\min}$, one has $\mathcal{E}(\hat{\mu}) \geq \mathcal{E}_{\min}(\hat{\mu})$. However, these two values coincide if and only if there is a value λ of the parameter such that $\Phi_\lambda = \Phi$. We state this in the following proposition:

Proposition 3. ***(Semiparametric specification testing)*** *Under Assumptions 1, 2 and 3, assume that the distributions of the unobserved heterogeneity terms $\mathbf{P}_x$ and $\mathbf{Q}_y$ are known. Then $\mathcal{E}(\hat{\mu}) \geq \mathcal{E}_{\min}(\hat{\mu})$, with equality if and only if there is a value λ of the parameter such that $\Phi_\lambda = \Phi$.*

5 Empirical Application

[TO BE ADDED]

6 Computation

Maximizing the conditional likelihood requires computing the optimal matching μ^λ for a large number of values of λ. But the optimal matching will be a large-dimensional object in realistic applications; and it is itself the maximizer of $\mathcal{W}$ in (2.10). It is therefore crucial to be able to compute μ^λ efficiently. We show here how the Iterative Projection Fitting Procedure (IPFP) often provides a solution to this problem.

Take the multinomial logit Choo-Siow model of Example 1 for instance. Fix a value of λ and drop it from the notation: let the joint surplus function be Φ, with optimal matching μ. Formula (3.3) can be rewritten as

$$\mu_{xy} = \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}\mu_{0y}}. \tag{6.1}$$

As noted by Decker et al. (2012), we could just plug this into the feasibility constraints $\sum_y \mu_{xy} + \mu_{x0} = \hat{n}_x$ and $\sum_x \mu_{xy} + \mu_{0y} = \hat{m}_y$ and solve for the numbers of singles μ_{x0} and μ_{0y}. Unfortunately, the resulting equations are still high-dimensional and highly nonlinear, which makes them hard to handle. Even proving the uniqueness of the solution to this system of equations is a hard problem.

On the other hand, to find a feasible solution of (3.3), we could start from an infeasible solution and project it somehow on the set of feasible matchings $\mathcal{M}(\hat{n}, \hat{m})$. Moreover, IPFP was precisely designed to find projections on intersecting sets of constraints, by projecting iteratively on each constraint[4]. The intuition of the method is straightforward. Assume that there exists a convex function $E(\mu)$ defined for any $\mu = \left(\mu_{xy}, \mu_{x0}, \mu_{0y}\right) \geq 0$, and such that $E(\mu_{xy}, n_x - \sum_y \mu_{xy}, m_y - \sum_x \mu_{xy}) = \mathcal{E}\left(\mu_{xy}\right)$, and E is almost everywhere strictly convex and smooth. Problem (2.10) rewrites as the maximization of $\sum_{x\in\mathcal{X},y\in\mathcal{Y}} \mu_{xy}\Phi_{xy} - E(\mu)$ over the set of vectors $\mu \geq 0$ satisfying the constraints on the margins $\sum_{y\in\mathcal{Y}_0} \mu_{xy} = n_x$ and $\sum_{x\in\mathcal{X}_0} \mu_{xy} = m_y$. Introducing u_x and v_y the Lagrange multipliers of the constraints $\mu \in \mathcal{M}$

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

yields

$$\max_{\mu\geq 0}\min_{u,v}\sum_{x\in\mathcal{X}}u_x(n_x-\sum_{y\in\mathcal{Y}_0}\mu_{xy})+\sum_{y\in\mathcal{Y}}v_y(m_y-\sum_{x\in\mathcal{X}_0}\mu_{xy})+\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\mu_{xy}\Phi_{xy}-E(\mu) \quad (6.2)$$

whose first order conditions are $\partial E/\partial\mu_{xy}=\Phi_{xy}-u_x-v_y$, $\partial E/\partial\mu_{x0}=-u_x$, and $\partial E/\partial\mu_{0y}=-v_y$.

However, instead of computing the full problem (6.2), we shall solve iteratively: at step $2k+1$ the minmax problem with u and μ as variables keeping v fixed ($=v^{2k}$), that is

$$\min_{u}\max_{\mu\geq 0}\sum_{x\in\mathcal{X}}u_x(n_x-\sum_{y\in\mathcal{Y}_0}\mu_{xy})-\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}_0}v_y^{2k}\mu_{xy}+\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\mu_{xy}\Phi_{xy}-E(\mu) \quad (6.3)$$

and, at step $2k+2$, the minmax problem with v and μ as variables keeping u fixed ($=u^{2k+1}$), that is

$$\min_{v}\max_{\mu\geq 0}-\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}_0}u_x^{2k+1}\mu_{xy}+\sum v_y(m_y-\sum_{x\in\mathcal{X}_0}\mu_{xy})+\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\mu_{xy}\Phi_{xy}-E(\mu). \quad (6.4)$$

This leads us to the following algorithm.

Algorithm 1 (Iterative Projection Fitting Procedure)**.**

Step 0*. Start with any initial choice of* $\left(u^0,v^0\right)$ *and set* $k=0$*.*

Step $2k+1$*. Keep* v^{2k} *fixed and look for* u *and* μ *solution to (6.3). By F.O.C.,*

$$\frac{\partial E(\mu)}{\partial\mu_{xy}}=\Phi_{xy}-u_x-v_y^{2k}\ ;\ \frac{\partial E(\mu)}{\partial\mu_{x0}}=-u_x\ ;\ \frac{\partial E(\mu)}{\partial\mu_{0y}}=-v_y^{2k} \quad (6.5)$$

and $\sum_{y\in\mathcal{Y}_0}\mu_{xy}=n_x$*. Call* u^{2k+1} *and* μ^{2k+1} *the solutions to this problem.*

Step $2k+2$*. Keep* u^{2k+1} *fixed and look for* v *and* μ *such that (6.4) which yields F.O.C.*

$$\frac{\partial E(\mu)}{\partial\mu_{xy}}=\Phi_{xy}-u_x^{2k+1}-v_y\ ;\ \frac{\partial E(\mu)}{\partial\mu_{x0}}=-u_x^{2k+1}\ ;\frac{\partial E(\mu)}{\partial\mu_{0y}}=-v_y \quad (6.6)$$

and $\sum_{x\in\mathcal{X}_0}\mu_{xy}=m_y$*. Call* v^{2k+2} *and* μ^{2k+2} *the solutions. If* μ^{2k+2} *is close enough to* μ^{2k+1}*, then take* $\mu=\mu^{2k+2}$ *to be the optimal matching and stop; otherwise add one to* k *and go to step 2k+1.*

Note that the algorithm can be interpreted as a Walrasian tâtonnement process where the prices of the x and the y are moved iteratively in order to adjust supply to demand on each side of the market. We prove in Appendix A that:

Theorem 4. *The algorithm converges to the solution μ of (2.10).*

As remark of importance, note that there are many possible ways of extending $\mathcal{E}$ (which is defined only on $\mathcal{M}$) to the entire space of $\mu \geq 0$. In practice, good judgement should be exercised, as the choice of E extending $\mathcal{E}$ is crucial for good practical performance of the algorithm.

Example 1 continued. *To illustrate, take the multinomial logit Choo and Siow model from Example 1. Here, we take $E(\mu) = \sum_{x\in\mathcal{X}}\sum_{x\in\mathcal{Y}_0}\mu_{xy}\log\mu_{xy} + \sum_{x\in\mathcal{X}_0}\sum_{x\in\mathcal{Y}}\mu_{xy}\log\mu_{xy}$, and we have $\partial E/\partial\mu_{xy} = 2 + 2\log\mu_{xy}$, $\partial E/\partial\mu_{x0} = 1 + \log\mu_{x0}$, and $\partial E/\partial\mu_{0y} = 1 + \log\mu_{0y}$. Start with $u_0 = v_0 = 0$. At step $2k+1$, keep v^{2k} fixed, and look for u and μ satisfying equations (6.5), which yields $\mu_{xy}^{2k+1} = (\mu_{x0}^{2k+1}\mu_{0y}^{2k})^{1/2}\exp(\Phi_{xy}/2)$, so that*

$$\mu_{x0}^{2k+1} + \sqrt{\mu_{x0}^{2k+1}}\sqrt{\mu_{0y}^{2k}}\exp\left(\frac{\Phi_{xy}}{2}\right) = n_x \tag{6.7}$$

while at step 2k+2 do the converse.

According to computational experiments we ran, the convergence of this algorithm is extremely fast compared to standard optimization methods. The results of our computational experiment (and benchmark with other methods) are reported in Appendix D. We next illustrate the algorithm in the nested logit case.

Example 2 continued *Consider the Nested Logit model of Example 2, and 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 which comes from (3.4)*

$$\mu_{xy} = \mu_{x0}^{\frac{\sigma_x}{\sigma_x+\tau_y}}\mu_{0y}^{\frac{\tau_y}{\sigma_x+\tau_y}}\exp\frac{\Phi_{xy}}{\sigma_x+\tau_y}$$

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

$$n_x = \mu_{x0} + \sum_{y \in \mathcal{Y}} \mu_{x0}^{\frac{\sigma_x}{\sigma_x+\tau_y}} \mu_{0y}^{\frac{\tau_y}{\sigma_x+\tau_y}} \exp \frac{\Phi_{xy}}{\sigma_x + \tau_y} \tag{6.8}$$

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

$$m_y = \mu_{0y} + \sum_{x \in \S} \mu_{0y}^{\frac{\tau_y}{\sigma_x+\tau_y}} \mu_{x0}^{\frac{\sigma_x}{\sigma_x+\tau_y}} \exp \frac{\Phi_{xy}}{\sigma_x + \tau_y}. \tag{6.9}$$

Note that steps (6.8) and (6.9) only require inverting a continuous and increasing real function of one variable, and are hence extremely 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 depends on the size of the nests.

Concluding Remarks

As mentioned earlier, several other approaches to estimating matching models with heterogeneity exist. One could directly specify the equilibrium utilities of each man and woman, as Hitsch, Hortacsu and Ariely (2010) did in a non-transferable utility model. Under separability, this would amount to choosing a distribution $\mathbf{P}_x$ and a parametrization λ of U and fitting the multinomial choice model where men maximize $U_{xy}(\lambda) + \varepsilon_{iy}$ over their marital options $y \in \mathcal{Y}_0$. The downside is that unlike the joint surplus, the utilities U and V are not primitive objects; and it is very difficult to justify a specification of equilibrium utilities.

An alternative class of approaches pools data from many markets in which the surplus from a match is assumed to be the same. Fox (2010) starts from the standard monotonicity property of single-agent choice models, in which under very weak assumptions, the probability of choosing an alternative increases with its mean utility. By analogy, he posits a "rank-order property" for matching models with transferable utility: 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. Unlike the results we derived from the multinomial logit Choo-Siow framework, the rank-order property is not implied by any theoretical

model we know of. In our framework, it holds only when the generalized entropy is a constant function, that is when there is no matching on unobservable characteristics. The attraction of the identification results based on the rank-order property, on the other hand, is that they extend easily to models with many-to-one or many-to-many matching.

It is worthwhile noting that Fox and Yang (2012) take an approach that is somewhat dual to ours: while we use separability to restrict the distribution of unobserved heterogeneity so that we can focus on the surplus over observables, they restrict the latter in order to recover the distribution of complementarities across unobservables. To do this, they rely on pooling data across many markets; in fact given the very high dimensionality of unobservable shocks, their method, while very ingenious, has yet to be tested on real data.

We have left some interesting theoretical issues for future research. One such issue, for instance, 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; however, when there is a finite number of agents in each group, the surplus function $\tilde{\Phi}_{ij} = \Phi(x_i, y_j) + \varepsilon_{iy} + \eta_{xj}$ becomes stochastic, and it is easy to see from the proof in the Appendix that Theorem 1 is still valid with G_x and H_y replaced by $\hat{G}_x$ and $\hat{H}_y$ where

$$\hat{G}_x(\mathbf{U}_{\mathbf{x}\cdot}) = \frac{1}{n_x} \sum_{i:x_i=x} \max_{y \in \mathcal{Y}_0} \{U_{xy} + \varepsilon_{iy}\} \text{ and } \hat{H}_y(\mathbf{V}_{\cdot\mathbf{y}}) = \frac{1}{m_y} \sum_{j:y_j=y} \max_{x \in \mathcal{X}_0} \{V_{xy} + \eta_{xj}\}$$

While the pointwise convergence $\hat{G}_x(\mathbf{U}_{\mathbf{x}\cdot}) \to G_x(\mathbf{U}_{\mathbf{x}\cdot})$ and $\hat{H}_y(\mathbf{V}_{\cdot\mathbf{y}}) \to H_y(\mathbf{V}_{\cdot\mathbf{y}})$ as the number of individuals gets large follows from the law of large numbers, it is natural to expect that the solutions $\hat{\mu}$ and $(\hat{U}, \hat{V})$ of the finitely sampled primal and dual problems converge to their large population analogs[5]. This goes beyond the scope of the present paper and is left as a conjecture. Likewise, exploration of the rate of convergence is left for future research.

To conclude, let us emphasize the wide applicability of the methods introduced in the present paper, and the potential for extensions.

[5]What is needed is to show that the gradient of the sum of the Legendre transforms of the $\hat{G}_x$ and the $\hat{H}_y$ maps converges to its population analog.

On the applied front, the estimators introduced in this paper provide a tractable parametric estimator of the matching surplus and can be put to work in many applied settings. Outside of the marriage market, Guadalupe et al. (2013) apply it to international trade; Bojilov and Galichon (2013) to the labor market.

On the methodological front, a challenge is to extend the logit setting of Choo and Siow to the case where the observable characteristics of the partners are possibly continuous. This issue is addressed by Dupuy and Galichon (2013) using the theory of extreme value processes; they also propose a test of the number of relevant dimensions for the matching problem. In some cases, closed-form solutions exist: see Bojilov and Galichon (2013).

While the framework we used here is bipartite, one-to-one matching, our results open the way to possible extensions to other matching problems. Among these, the "roommate problem" drops the requirement that the two partners of a match are drawn from distinct populations. Chiappori, Galichon and Salanié (2013) have shown that this problem is in fact isomorphic, in a large population, to an associated bipartite matching problem; as a consequence, the empirical tools from the present paper can be extended to the study of the roommate problem. Although an extension to situations of "one-to-many matching" where one entity on one side of the market (such as a firm) may match with several agents on the other (such as employees) seems less direct, it is likely that the present approach would be useful. It may also be insightful in the study of trading on networks, when transfers are allowed (thus providing an empirical counterpart to Hatfield and Kominers, 2012, and Hatfield et al., 2013). Finally, the approach proposed in Proposition 2 to identify utilities in discrete choice problems has nothing specific to the matching setting; they are applied in Chiong, Galichon and Shum (2013) in order to provide identification in dynamic discrete choice problems in very general situations–in particular, outside of the GEV framework commonly used in these problems.

Appendix

A Proofs

A.1 Proof of Proposition 2

Replace the expression of G_x (2.2) in the formula for G_x^* (2.4). It follows

$$\begin{aligned} G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) &= -\min_{\tilde{\mathbf{U}}_{\mathbf{x}\cdot}}\{\mathbb{E}_{\mathbf{P}_x}\max_{y\in\mathcal{Y}}\left(\tilde{U}_{xy}+\varepsilon_{iy},\varepsilon_{i0}\right)-\sum_{y\in\mathcal{Y}}\mu_{y|x}\tilde{U}_{xy}\} \\ &= -\min_{\bar{\mathbf{U}}_{\mathbf{x}\cdot}}\{\sum_{y\in\mathcal{Y}_0}\mu_{y|x}\bar{U}_{xy}+\mathbb{E}_{\mathbf{P}_x}\max_{y\in\mathcal{Y}_0}\left(\varepsilon_{iy}-\bar{U}_{xy}\right)\} \end{aligned}$$

where $\bar{U}_{xy}=-\tilde{U}_{xy}$ and $\bar{U}_{x0}=0$ in the second line. The first term in the minimand is the expectation of $\bar{\mathbf{U}}_{\mathbf{x}\cdot}$ under the distribution $\mu_{Y|X=x}$; therefore this can be rewritten as

$$G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}})=-\min_{\bar{U}_{xy}+\bar{W}_x(\boldsymbol{\varepsilon}_{\mathbf{i}\cdot})\geq\varepsilon_{iy}}\{E_{\mu_{Y|X=x}}\bar{U}_{xY}+\mathbb{E}_{\mathbf{P}_x}\bar{W}_x(\varepsilon_{i\cdot})\}$$

where the minimum is taken over all pairs of functions $(\bar{\mathbf{U}}_{\mathbf{x}\cdot},\bar{W}_x(\varepsilon_{i\cdot}))$ that satisfy the inequality. We recognize the value of the dual of a matching problem in which the margins are $\mu_{Y|X=x}$ and $\mathbf{P}_x$ and the surplus is ε_{iy}. By the equivalence of the primal and the dual, this yields Expression (2.8).

A.2 Proof of Theorem 1

In the proof we denote $\tilde{n}(x,\varepsilon)$ the distribution of (x,ε) where the distribution of x is n, and the distribution of ε conditional on x is $\mathbf{P}_x$; formally, for $S\subseteq\mathcal{X}\times\mathbb{R}^{\mathcal{Y}_0}$, we get

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

(i) By the dual formulation of the matching problem (see Gretsky, Ostroy and Zame, 1992), the market equilibrium assigns utilities $\tilde{u}(x,\varepsilon)$ to man i such that $x_i=x$ and $\varepsilon_i=\varepsilon$ and $\tilde{v}(y,\eta)$ to woman j such that $y_j=y$ and $\eta_j=\eta$ so as to solve

$$\mathcal{W}=\min\left(\int\tilde{u}(x,\varepsilon)\,d\tilde{n}(x,\varepsilon)+\int\tilde{v}(y,\eta)\,d\tilde{m}(y,\eta)\right)$$

where the minimum is taken under the set of constraints $\tilde{u}(x,\varepsilon)+\tilde{v}(y,\eta)\geq\Phi_{xy}+\varepsilon_y+\eta_x$, $\tilde{u}(x,\varepsilon)\geq\varepsilon_0$, and $\tilde{v}(y,\eta)\geq\eta_0$. For $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, introduce

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

so that $\tilde{u}(x,\varepsilon)=\max_{y\in\mathcal{Y}}\left\{U_{xy}+\varepsilon_y,\varepsilon_0\right\}$ and $\tilde{v}(y,\eta)=\max_{x\in\mathcal{X}}\left\{V_{xy}+\eta_x,\eta_{0j}\right\}$. Then $\mathcal{W}$ minimizes $\int\max_{y\in\mathcal{Y}}\left\{U_{xy}+\varepsilon_y,\varepsilon_0\right\}d\tilde{n}(x,\varepsilon)+\int\max_{x\in\mathcal{X}}\left\{V_{xy}+\eta_x,\eta_0\right\}d\tilde{m}(y,\eta)$ over U and V subject to constraints $U_{xy}+V_{xy}\geq\Phi_{xy}$. Assign non-negative multipliers μ_{xy} to these constraints. By convex duality, we can rewrite

$$\mathcal{W}=\max_{\mu_{xy}\geq0}\left(\sum_{\substack{x\in\mathcal{X}\\y\in\mathcal{Y}}}\mu_{xy}\Phi_{xy}\qquad-\max_{U_{xy}}\left\{\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}U_{xy}-\int\max_{y\in\mathcal{Y}}\left\{U_{xy}+\varepsilon_y,\varepsilon_0\right\}d\tilde{n}(x,\varepsilon)\right\}\right.$$
$$\left.-\max_{V_{xy}}\left\{\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}V_{xy}-\int\max_{x\in\mathcal{X}}\left\{V_{xy}+\eta_x,\eta_0\right\}d\tilde{m}(y,\eta)\right\}\right).$$

Now, $\int\max_{y\in\mathcal{Y}}\left\{U_{xy}+\varepsilon_y,\varepsilon_0\right\}d\tilde{n}(x,\varepsilon)=\sum_x n_x\mathbb{E}_{\mathbf{P}_x}[\max_{y\in\mathcal{Y}}U_{xy}+\varepsilon_y,\varepsilon_0]=n_xG_x(\mathbf{U}_{\mathbf{x}\cdot})$, where $\mathbb{E}_{\mathbf{P}_x}$ denotes the expectation over the population of men in group x, and where we have invoked Assumption 1 in order to replace the sum by an expectation. Adding the similar expression for women, we get that $\mathcal{W}$ is the maximum over $\mu\geq0$ of $\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}\Phi_{xy}-A(\mu)-B(\mu)$, where $A(\mu)=\max_{(U_{xy})}\{\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}U_{xy}-\sum_{x\in\mathcal{X}}n_xG_x(\mathbf{U}_{\mathbf{x}\cdot})\}$, and B has a similar expression involving H and m instead of G and n. Consider the term with first subscript x in $A(\mu)$. It is $n_x(\sum_{y\in\mathcal{Y}}\mu_{y|x}U_{xy}-G_x(\mathbf{U}_{\mathbf{x}\cdot}))$, that is n_x times the Legendre transform of G evaluated at $\boldsymbol{\mu}_{\cdot|\mathbf{x}}$, so we can rewrite $A(\mu)$ and $B(\mu)$ in terms of the Legendre-Fenchel transforms:

$$A(\mu)=\sum_{x\in\mathcal{X}}n_xG_x^*\left(\boldsymbol{\mu}_{\cdot|\mathbf{x}}\right)\text{ and }B(\mu)=\sum_{y\in\mathcal{Y}}m_yH_y^*\left(\boldsymbol{\mu}_{\cdot|\mathbf{y}}\right).$$

Expression (2.10) follows, and points (ii), (iii) and (iv) are then deduced immediately.

A.3 Proof of Theorem 2

If Assumption 4 holds for $\mathbf{P}_x$, then the function G_x is increasing in each of its arguments; since its derivatives are the probabilities $\mu_{y|x}$ at the optimum, they must be positive. More-

over, $G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}})$ would be infinite if $\sum_y \mu_{y|x}$ were to equal one; and that is not compatible with optimality. We can therefore neglect the feasibility constraints (2.10). By the first order conditions in the program defining A in the proof of Theorem 1 above, one gets $U_{xy} = \left(\partial G_x^*/\partial \mu_{y|x}\right)(\boldsymbol{\mu}_{\cdot|x})$ which is (2.13). The envelope theorem in the same program gives us (2.12), which proves (i). Similarly, one gets $V_{xy} = \left(\partial H_y^*/\partial \mu_{x|y}\right)(\boldsymbol{\mu}_{\cdot|y})$ which, by summation and using the fact that $\Phi_{xy} = U_{xy} + V_{xy}$, yields (2.14), proving (ii).

A.4 Proof of Corollary 1

The result follows from the fact that $U_{xy} = \alpha_{xy} + \tau_{xy}$ and $V_{xy} = \gamma_{xy} - \tau_{xy}$; thus if U_{xy} and V_{xy} are identified and τ_{xy} is observed, then α and γ are identified by $\alpha_{xy} = U_{xy} - \tau_{xy}$ and $\gamma_{xy} = V_{xy} + \tau_{xy}$.

A.5 Proof of Theorem 3

(i) The Moment Matching estimator $\hat{\lambda}$ is solution to problem (4.3). Hence, by F.O.C. $\hat{\lambda}$ satisfies $\sum_{x,y} \hat{\mu}_{xy} \Phi_{xy}^k = \partial \mathcal{W}/\partial \lambda_k(\Phi^{\hat{\lambda}}, n, m)$; but by the Envelope Theorem, $\partial \mathcal{W}/\partial \lambda_k(\Phi^{\hat{\lambda}}, n, m) = \sum_{x,y} \mu_{xy}^{\hat{\lambda}} \Phi_{xy}^k$.

(ii) Program (4.3) can be rewritten as

$$\max_{\lambda \in \mathbb{R}^k} \min_{\mu \in \mathcal{M}} \sum_k \lambda_k \sum_{x,y} \left(\hat{\mu}_{xy} - \mu_{xy}\right) \Phi_{xy}^k + \mathcal{E}(\mu)$$

that is μ minimizes $\mathcal{E}(\mu)$ over the set of $\mu \in \mathcal{M}$ such that $\sum_{x,y} \left(\hat{\mu}_{xy} - \mu_{xy}\right) \Phi_{xy}^k = 0$.

A.6 Proof of Proposition 3

Since $\mu^{\hat{\lambda}}$ maximizes $\mathcal{W}$ when $\lambda = \hat{\lambda}$, $\sum_{x,y} \hat{\mu}_{xy} \Phi_{xy}^{\hat{\lambda}} - \mathcal{E}(\hat{\mu}) \leq \sum_{x,y} \mu_{xy}^{\hat{\lambda}} \Phi_{xy}^{\hat{\lambda}} - \mathcal{E}\left(\mu^{\hat{\lambda}}\right)$, and, since $\mathcal{E}$ is strictly convex in μ, equality holds if and only if $\mu^{\hat{\lambda}} = \hat{\mu}$. But equality $\sum_{x,y} \hat{\mu}_{xy} \Phi_{xy}^{\hat{\lambda}} = \sum_{x,y} \mu_{xy}^{\hat{\lambda}} \Phi_{xy}^{\hat{\lambda}}$ holds by construction, hence $\mathcal{E}(\hat{\mu}) \geq \mathcal{E}(\mu^{\hat{\lambda}})$ with equality if and only if $\mu^{\hat{\lambda}} = \hat{\mu}$.

A.7 Proof of Theorem 4

The proof uses results in Bauschke and Borwein (1997), which builds on Csiszar (1975). The map $\mu \to E(\mu)$ is essentially smooth and essentially strictly convex; hence it is a "Legendre function" in their terminology. Introduce D the associated "Bregman divergence" as

$$D(\mu;\nu) = E(\mu) - E(\nu) - \langle \nabla E(\nu), \mu - \nu \rangle,$$

and introduce the linear subspaces $\mathcal{M}(n)$ and $\mathcal{M}(m)$ by

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

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

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

By Theorem 8.4 of Bauschke and Borwein, the iterated projection algorithm converges[6] to the projection μ of $\mu^{(0)}$ on $\mathcal{M}(n,m)$, which is also the maximizer μ of (2.10).

B Explicit examples

The Generalized Extreme Values 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 $\mathbf{P}_x$. Then introducing utility shocks $\varepsilon_x \sim \mathbf{P}_x$, we have by a theorem of McFadden (1978):

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

[6] In the notation of their Theorem 8.4, the hyperplanes (C_i) are $\mathcal{M}(p)$ and $\mathcal{M}(q)$; and the Bregman/Legendre function f is our ϕ.

where γ is the Euler constant $\gamma \simeq 0.577$. Therefore, if $\sum_{y\in\mathcal{Y}_0} p_y = 1$, then $G_x^*(p) = \sum_{y\in\mathcal{Y}_0} p_y w_y^x(p) - \left(\log g_x\left(e^{w^x(p)}\right) + \gamma\right)$, where for $x \in \mathcal{X}$, the vector $w^x(p)$ is a solution to the system of equations $p_y = \left(\partial \log g_x/\partial w_y^x\right)\left(e^{w^x}\right)$ for $y \in \mathcal{Y}_0$. Hence, the part of the expression of $\mathcal{E}(\mu)$ arising from the heterogeneity on the men side is

$$\sum_{x\in\mathcal{X}} \{n_x \log g_x\left(e^{w^x(\mu_{x\cdot}/n_x)}\right) - \sum_{y\in\mathcal{Y}_0} \mu_{xy} w_y^x(\mu_{x\cdot}/n_x)\} + C$$

where $C = \gamma \sum_{x\in\mathcal{X}} n_x$. The derivative of this expression with respect to μ_{xy} $(x, y \geq 1)$ is $-w_y^x(\mu/n)$.

B.1 Derivations for Example 1

Claims of Section 3.1. With type I extreme value iid distributions, the expected utility is $G_x(\mathbf{U}_{\mathbf{x}\cdot}) = \log(1+\sum_{y\in\mathcal{Y}} \exp(U_{xy}))$, 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 (3.2). Equation (3.3) obtains by straightforward differentiation.

Claims of Section 4.3. We can rewrite L as

$$\log L(\lambda) = \sum_{x,y} \hat{\mu}_{xy} \log \frac{\left(\mu_{xy}^\lambda\right)^2}{\mu_{x0}^\lambda \mu_{0y}^\lambda} + \sum_{x\in\mathcal{X}} \hat{n}_x \log \frac{\mu_{x0}^\lambda}{\hat{n}_x} + \sum_{y\in\mathcal{Y}} \hat{m}_y \log \frac{\mu_{0y}^\lambda}{\hat{m}_y} = \sum_{x,y} \hat{\mu}_{xy}\Phi_{xy}^\lambda - \mathcal{W}(\lambda),$$

which establishes (4.5). Now by the envelope theorem, $\partial\mathcal{W}/\partial\lambda = \sum_{x,y} \mu_{xy}^\lambda \partial\Phi_{xy}^\lambda/\partial\lambda$ since the entropy term does not depend on λ in the multinomial logit Choo and Siow model; this proves that $\hat{\lambda}^{MM} = \hat{\lambda}^{MLE}$.

B.2 Derivations for Example 2

Consider a man of a group x; and as in the text, drop the x indices. By (B.1), the expected utility of this man is $G(\mathbf{U}\cdot) = \log(1 + \sum_s (\sum_e e^{U_{se}/\sigma_s})^{\sigma_s})$, hence, by (2.3), it

follows that $\mu_{se}/\mu_0 = (\sum_e e^{U_{se}/\sigma_s})^{\sigma_s - 1} e^{U_{se}/\sigma_s}$. 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 (2.6),

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

Now if the nested logit applies for men of group x with parameters $(\sigma^x_{s'})$ and for women of group y with parameters (τ^y_s), we can write $U_{x,s'e'} = \log(\mu_{x,s'}/\mu_{x0}) + \sigma^x_{s'} \log(\mu_{x,s'e'}/\mu_{x,s'})$ and $V_{se,y} = \log(\mu_{s,y}/\mu_{0y}) + \tau^y_s \log(\mu_{se,y}/\mu_{s,y})$. Adding up gives the formula for Φ_{xy} in the text. To obtain the expected utilities, we just substitute in the expression of $G(U)$ the values of U_{se}.

B.3 Derivations for Example 3

When $\mathbf{P}_x$ is a mixture of i.i.d. Gumbel distributions of scale parameters σ^x_k with weights β^x_k, 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 $G_x(\mathbf{U}_{x\cdot}) = \sum_k \beta^x_k G_{xk}(\mathbf{U}_{x\cdot})$, where $G_{xk}(\mathbf{U}_{x\cdot}) = \sigma^x_k \log(1 + \sum_{y \in \mathcal{Y}} e^{U_{xy}/\sigma^x_k})$. Still from the results of Example 1, $G^*_{xk}(\boldsymbol{\mu}) = \sigma^x_k \sum_{y \in \mathcal{Y}_0} \mu_y \log \mu_y$. By standard results in Convex Analysis (see e.g. Rockafellar 1970, section 20), the convex conjugate of a sum of functions is the infimum-convolution of the conjugates of the functions in the sum. The convex conjugate of $\mathbf{U}_{x\cdot} \to \beta^x_k G_{xk}(\mathbf{U}_{x\cdot})$ is $f^*(\mu^k_\cdot) = \beta^x_k G^*_{xk}\left(\frac{\mu^k_\cdot}{\beta^x_k}\right)$; thus (3.6) follows. $\mu_{y|x}$ obtains by straightforward differentiation of (3.5). Finally, it follows from the properties of the conditional logit model that the log odds ratio $\log(\mu^k_y/\mu^k_0)$ must coincide with U_{xy}/σ^x_k, QED.

B.4 Derivations for Example 4

Claims of Section 3.1. From Proposition 2, $G^*_x(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = -\max_{\pi \in \mathcal{M}_x} \mathbb{E}_\pi[\zeta_x(Y)\varepsilon]$, 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

$$\Pr(\zeta_x(Y) = \zeta_{(k)}|x) = \sum_{\zeta_x(y)=\zeta_{(k)}} \mu_{y|x}. \tag{B.2}$$

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 $\varepsilon \in [\varepsilon_{(k-1)}, \varepsilon_{(k)}]$; and since probability is conserved, the sequence is constructed recursively by

$$F_\varepsilon\left(\varepsilon_{(k)}\right) - F_\varepsilon\left(\varepsilon_{(k-1)}\right) = \sum_{\zeta_x(y)=\zeta_{(k)}} \mu_{y|x}, \tag{B.3}$$

giving $F_\varepsilon\left(\varepsilon_{(k)}\right) = \sum_{\zeta_x(y)\leq\zeta_{(k)}} \mu_{y|x}$; and as a result, $G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = -\sum_{1\leq k\leq|\mathcal{Y}|+1} \zeta_{(k)} e_k$, 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 interval $\left[\varepsilon_{(k-1)}, \varepsilon_{(k)}\right]$; then $-n_x G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = n_x \sum_{1\leq k\leq|\mathcal{Y}|+1} \zeta_{(k)} \sum_{\zeta_x(y)=\zeta_{(k)}} \mu_{y|x} \bar{e}_k = \sum_y \mu_{xy} \bar{e}_{K(y)}$, with $K(y)$ the value of k such that $\zeta_x(y) = \zeta_{(k)}$; in the main text we use the notation $\bar{e}_x(y) = \bar{e}_{K(y)}$.

When ε is distributed uniformly over $[0,1]$, (B.3) becomes $\varepsilon_{(k)} = \sum_{\zeta_x(y)\leq\zeta_{(k)}} \mu_{y|x}$, and $\mathbb{E}\left[\varepsilon \mathbf{1}(\varepsilon \in [\varepsilon_{(k-1)}, \varepsilon_{(k)}])\right] = \left(\varepsilon_{(k)} - \varepsilon_{(k-1)}\right)\left(\varepsilon_{(k)} + \varepsilon_{(k-1)}\right)/2$, we obtain

$$\mathbb{E}\left[\varepsilon \mathbf{1}(\varepsilon \in [\varepsilon_{(k-1)}, \varepsilon_{(k)}])\right] = \sum_{y|\zeta_x(y)=\zeta_{(k)}} \mu_{y|x} \Big(\sum_{y'|\zeta_x(y')<\zeta_x(y)} \mu_{y'|x} + \frac{1}{2} \sum_{y'|\zeta_x(y')=\zeta_x(y)} \mu_{y'|x} \Big).$$

Summing up over $k = 1, \ldots, |\mathcal{Y}| + 1$, we get $G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = -\frac{1}{2} \sum_{y,y'\in\mathcal{Y}_0} S_{yy'}^x \mu_{y|x} \mu_{y'|x}$,where $S_{yy'}^x = \max(\zeta_x(y), \zeta_x(y'))$.

Therefore, using $\mu_{0|x} = 1 - \sum_{y\in\mathcal{Y}} \mu_{y|x}$, we obtain

$$G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = -\frac{1}{2} \Big(\sum_{y,y'\in\mathcal{Y}} \left(S_{yy'}^x - S_{y0}^x - S_{0y'}^x + S_{00}^x\right) \mu_{y|x} \mu_{y'|x} + 2 \sum_{y\in\mathcal{Y}} (S_{y0}^x - S_{00}^x) \mu_{y|x} + S_{00}^x \Big).$$

Now define a matrix T^x and a vector σ^x by $T_{yy'}^x = S_{y0}^x + S_{0y'}^x - S_{yy'}^x - S_{00}^x$ and $\sigma_y^x = S_{00}^x - S_{y0}^x$; this gives $G_x^*(\boldsymbol{\mu}_{\cdot|\mathbf{x}}) = \frac{1}{2}\left(\boldsymbol{\mu}_{\cdot|\mathbf{x}}{}' T^x \boldsymbol{\mu}_{\cdot|\mathbf{x}} + 2\sigma^x . \boldsymbol{\mu}_{\cdot|\mathbf{x}} - S_{00}^x\right)$.

Introducing

$$A_{xy,x'y'} = \frac{1}{n_x} 1\left\{x = x'\right\} T^x_{yy'} + \frac{1}{m_y} 1\left\{y = y'\right\} T^y_{xx'} \tag{B.4}$$

$$B_{xy} = \sigma^x_y + \sigma^y_x \text{ and } c = -\sum_{x\in\mathcal{X}} n_x S^x_{00} - \sum_{y\in\mathcal{Y}} m_y S^y_{00} \tag{B.5}$$

leads to $\mathcal{E}(\mu) = (\mu' A\mu + 2B.\mu + c)/2$.where μ is the vector of μ_{xy} for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Claims of Section 3.2. Note that μ is determined by

$$\mathcal{W} = \max_{\mu\in\mathcal{M}(n,m)} \mathbf{\Phi}.\mu - \frac{1}{2}(\mu' A\mu + 2B.\mu + c)$$

where $\mathbf{\Phi}.\mu$ is the vector product $\sum_{xy} \mu_{xy}\Phi_{xy}$. Hence, if μ is interior, i.e. if there are no empty cells, the solution is given by $\mu = A^{-1}(\mathbf{\Phi} - B)$ and $\mathcal{W} = \frac{1}{2}((\mathbf{\Phi} - B)' A^{-1}(\mathbf{\Phi} - B) - c)$, where the invertibility of A follows from the fact that for each x, the values of $\zeta_x(y)$, $y \in \mathcal{Y}_0$ have been assumed to be distinct. One has $\partial A^{-1}/\partial n_x = -A^{-1}(\partial A/\partial n_x)A^{-1}$ and $\partial A/\partial n_x = -M^x/n_x^2$, where

$$M^x_{x'y,x''y'} = 1\left\{x = x' = x''\right\} T^x_{yy'} \tag{B.6}$$

hence, a calculation shows that $\partial\mathcal{W}/\partial n_x = (\mathbf{\Phi} - B)' A^{-1} M^x A^{-1}(\mathbf{\Phi} - B)/(2n_x^2)$, thus

$$\frac{\partial^2\mathcal{W}}{\partial n_x \partial n_{x'}} = \frac{1}{n_x^2 n_{x'}^2} \mu' R^{xx'} \mu \tag{B.7}$$

where

$$R^{xx'} = -n_x 1\left\{x = x'\right\} M^x + \frac{M^{x'} A^{-1} M^x + M^x A^{-1} M^{x'}}{2}. \tag{B.8}$$

Now, (2.19) yields

$$\frac{\partial u_{x'}}{\partial \Phi_{xy}} = \frac{\partial^2\mathcal{W}}{\partial n_{x'}\partial\Phi_{xy}} = \frac{1}{n_{x'}^2} Z^{x'}_{xy} \tag{B.9}$$

where

$$Z^{x'} = A^{-1} M^{x'} \mu \tag{B.10}$$

and it is recalled that the expression for M^x is given in (B.6). Finally (2.20) yields

$$\frac{\partial \mu_{xy}}{\partial \Phi_{x'y'}} = A^{-1}_{xy,x'y'}. \tag{B.11}$$

C Geometric interpretation

Our approach to inference has a simple geometric interpretation. Consider the set of comoments associated to every feasible matching

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

This is a convex polyhedron, which we call the *covariogram*; and if the model is well-specified the covariogram must contain the observed matching $\hat{\mu}$. For any value of the parameter vector λ, the optimal matching μ^λ generates a vector of comoments C^λ that belongs to the covariogram; and it also has an entropy $\mathcal{E}^\lambda \equiv \mathcal{E}(\mu^\lambda)$. We already know that this model is just-identified from the comoments: the mapping $\lambda \longrightarrow C^\lambda$ is invertible on the covariogram. Denote $\lambda(C)$ its inverse. The corresponding optimal matching has entropy $\mathcal{E}_r\left(C\right) = \mathcal{E}^{\lambda(C)}$. The level sets of $\mathcal{E}_r\left(.\right)$ are the isoentropy curves in the covariogram; they are represented on Figure 1. The figure assumes $K = 2$ dimensions; then λ can be represented in polar coordinates as $\lambda = r\exp(it)$. For $r = 0$, the model is uninformative and entropy is highest; the matching is random and generates comoments C_0. At the other extreme, the boundary ∂F of the covariogram corresponds to $r = \infty$. Then there is no unobserved heterogeneity and generically over t, the comoments generated by λ must belong to a finite set of vertices, so that λ is only set-identified. As r decreases for a given t, the corresponding comoments follow a trajectory indicated by the dashed line on Figure 1, from the boundary ∂F to the point C_0. At the same time, the entropy $\mathcal{E}^\lambda$ increases, and the trajectory crosses contours of higher entropy ($\mathcal{E}^{'}$ then $\mathcal{E}^{''}$ on the figure.) The CML Estimator $\hat{\lambda}$ could also be obtained by taking the normal to the isoentropy contour that goes through the observed comoments $\hat{C}^k = C^k\left(\hat{\mu}\right)$, as shown on Figure 1. Indeed, the estimator $\hat{\lambda}^{MM}$ of the parameter vector is given by the gradient of $\mathcal{E}_r\left(.\right)$ at the point $\hat{C}$, that is $\partial\mathcal{E}_r(\hat{C})/\partial C^k = \hat{\lambda}^{MM}_k$.

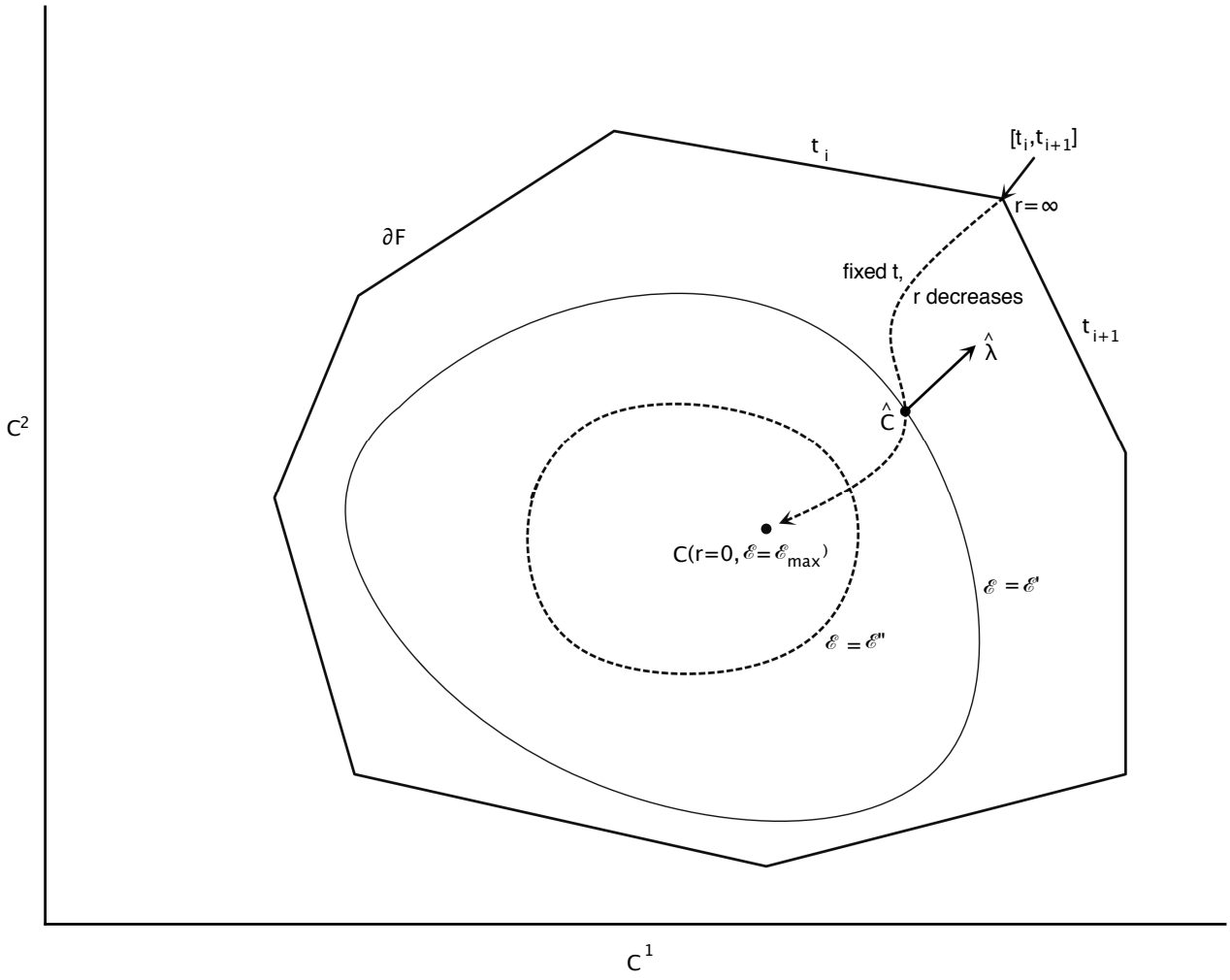

Figure 1: The covariogram and related objects

D Computational experiments

Equation (6.7) is a quadratic equation in only one unknown, $\sqrt{\mu_{x0}^{2k+1}}$; as such it can be solved in closed form.The convergence is extremely fast. We tested it on a simulation in which we let the number of categories $|\mathcal{X}| = |\mathcal{Y}|$ increase from 100 to 1,000. For each of these ten cases, we draw 50 samples, with the n_x and m_y drawn uniformly in $\{1, \ldots, 100\}$; and for each (x, y) match we draw Φ_{xy} from $\mathcal{N}(0, 1)$. To have a basis for comparison, we also ran two nonlinear equation solvers on the system of $(\mathcal{X}| + |\mathcal{Y})$ equations

$$a_x^2 + a_x \left(\sum_y \exp(\Phi_{xy}/2) b_y \right) = n_x \tag{D.1}$$

and

$$b_y^2 + b_y \left(\sum_x \exp(\Phi_{xy}/2) a_x \right) = m_y, \tag{D.2}$$

which characterizes the optimal matching with $\mu_{xy} = \exp(\Phi_{xy}/2)\sqrt{\mu_{x0}\mu_{0y}}$, $\mu_{x0} = a_x^2$, and $\mu_{0y} = b_y^2$ (see Decker et al. (2012)).

To solve system (D.1)–(D.2), we used both Minpack and Knitro. Minpack is probably the most-used solver in scientific applications, and underlies many statistical and numerical packages. Knitro[7] is a constrained optimization software; but minimizing the function zero under constraints that correspond to the equations one wants to solve has become popular recently.

For all three methods, we used C/C^{++} programs, run on a single processor of a Mac desktop. We set the convergence criterion for the three 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 2. Each panel gives the distribution of CPU times over 50 samples (20 for Knitro) for the ten experiments, in the form of a Tukey box-and-whiskers graph[8].

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

[8] The box goes from the first to the third quartile; the horizontal bar is at the median; the lower (resp.

The performance of IPFP stands out clearly—note the different vertical scales. While IPFP has more variability than Minpack and Knitro (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 the fastest sample under Minpack, and fifteen times smaller than the fastest time with Knitro. This is all the more remarkable that we fed the code for the Jacobian of the system of equations into Minpack, and for both the Jacobian and the Hessian into Knitro.

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.

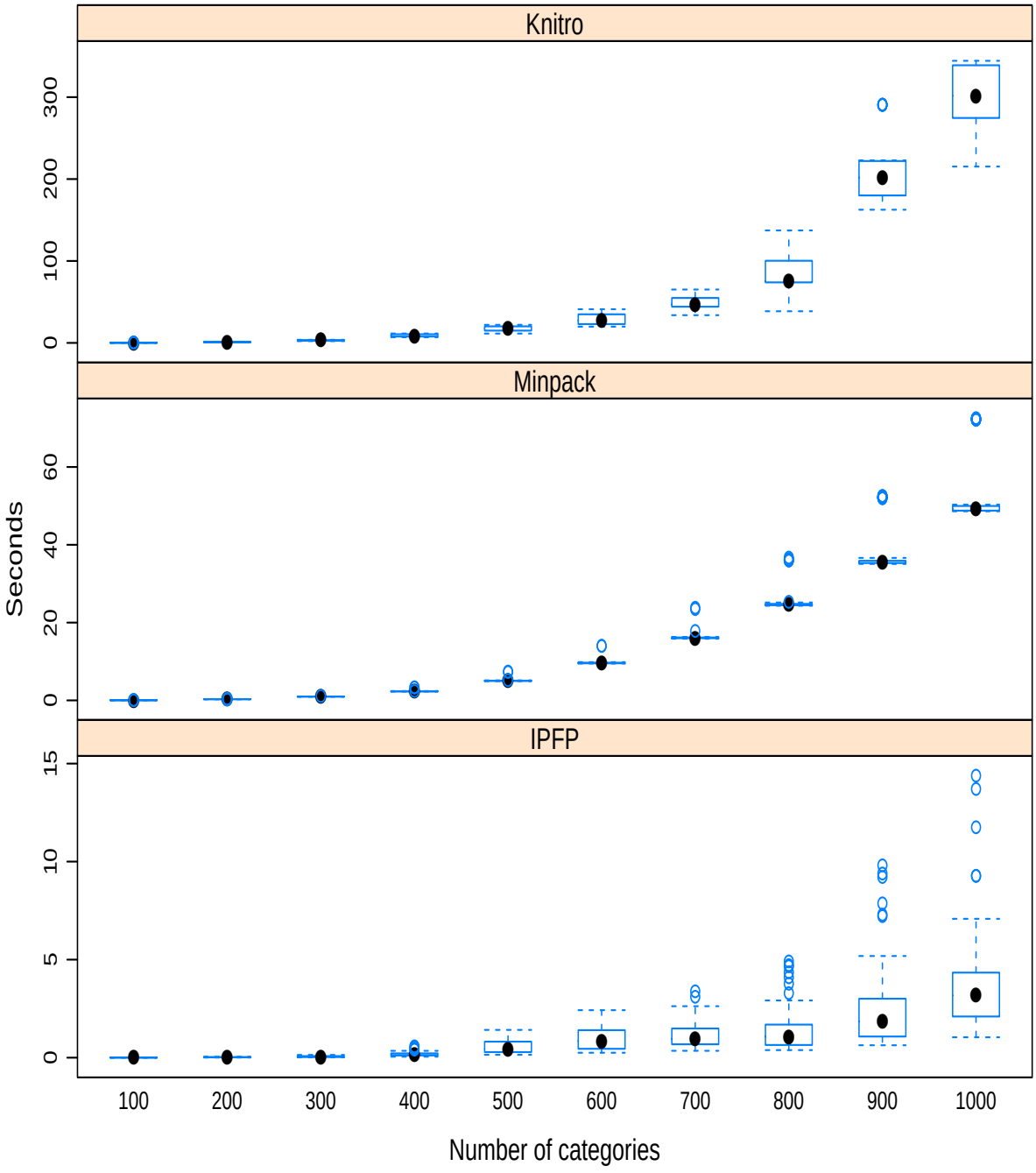

Figure 2: Solving for the optimal matching

References

[1] Ackerberg, D., C. Lanier Benkard, S. Berry, and A. Pakes (2007): "Econometric Tools for Analyzing Market Outcomes", chapter 63 of the *Handbook of Econometrics, vol. 6A*, J.J. Heckman and E. Leamer eds, North Holland.

[2] Anderson, S., A. de Palma, A., and J.-F. Thisse (1992): *Discrete Choice Theory of Product Differentiation*, MIT Press.

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

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

[5] Becker, G. (1973): "A Theory of Marriage, part I," *Journal of Political Economy*, 81, pp. 813–846.

[6] Berry, S. and Pakes, A. (2007): "The pure characteristics demand model". *International Economic Review* 48 (4), pp. 1193–1225.

[7] Bojilov, R., and A. Galichon (2013): "Closed-form solution for multivariate matching," mimeo.

[8] Botticini, M., and A. Siow (2008): "Are there Increasing Returns in Marriage Markets?," mimeo.

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

[10] Chiappori, P.-A., A. Galichon, and B. Salanié (2013): "The Roommate Problem is More Stable than You Think," mimeo.

[11] 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.

[12] Chiappori, P.-A., B. Salanié, and Y. Weiss (2012): "Partner Choice and the Marital College Premium," mimeo.

[13] Chiong, K., A. Galichon, and M. Shum (2013): "Estimating dynamic discrete choice models via convex analysis," mimeo.

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

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

[16] de Palma, A., and K. Kilani (2007): "Invariance of Conditional Maximum Utility," *Journal of Economic Theory*, 132, 137–146.

[17] 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.

[18] Dupuy, A. and A. Galichon (2013): "Personality traits and the marriage market," mimeo.

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

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

[21] Fox, J. (2011): "Estimating Matching Games with Transfers," mimeo.

[22] Fox, J., and C. Yang (2012): "Unobserved Heterogeneity in Matching Games," mimeo.

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

[24] Gale, D., and L. Shapley (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69, 9–14.

[25] Galichon, A., and B. Salanié (2010): "Matching with Tradeoffs: Revealed Preferences over Competing Characteristics," Discussion Paper 7858, CEPR.

[26] 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.

[27] Gretsky, N., J. Ostroy, and W. Zame (1992): "The nonatomic assignment model," *Economic Theory*, 2(1), 103–127.

[28] Gretsky, N., J. Ostroy, and W. Zame (1999): "Perfect competition in the continuous assignment model," *Journal of Economic Theory*, 88, 60–118.

[29] Guadalupe, M., V. Rappoport, B. Salanie and C. Thomas (2013): "The Perfect Match: Assortative Matching in International Acquisitions," mimeo.

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

[31] Hatfield, J. W., S. D. Kominers, A. Nichifor, M. Ostrovsky, and A. Westkamp (2011): "Stability and competitive equilibrium in trading networks," mimeo.

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

[33] Hitsch, G., A. Hortacsu, and D. Ariely (2010): "Matching and Sorting in Online Dating," *American Economic Review*, 100, 130–163.

[34] Jacquemet, N., and J.-M. Robin (2011): "Marriage with Labor Supply," mimeo.

[35] McFadden, D. (1978): "Modelling the Choice of Residential Location," in A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull (eds.), *Spatial interaction theory and planning models*, 75-96, North Holland: Amsterdam.

[36] Menzel, K. (2014): "Large Matching Markets as Two-Sided Demand Systems," working paper.

[37] Reiss, P. and F. Wolak (2007): "Structural Econometric Modeling: Rationales and Examples from Industrial Organization", chapter 64 of the *Handbook of Econometrics, vol. 6A*, J.-J. Heckman and E. Leamer eds, North Holland.

[38] Rockafellar, R.T. (1970). *Convex Analysis*. Princeteon University Press.

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

[40] Shimer, R., and L. Smith (2000): "Assortative matching and Search," *Econometrica*, 68, 343–369.

[41] Siow, A. (2008). "How does the marriage market clear? An empirical framework." *The Canadian Journal of Economics* 41 (4), pp. 1121–1155.

[42] Siow, A. (2009): "Testing Becker's Theory of Positive Assortative Matching," mimeo.

[43] Siow, A., and E. Choo (2006): "Estimating a Marriage Matching Model with Spillover Effects," *Demography*, 43(3), 463–490.