# Cupid's Invisible Hand:

## Social Surplus and Identification in Matching Models

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

February 18, 2015[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, the editor and referees and many seminar participants for very useful comments and discussions. Part of the research underlying this paper was done when Galichon was visiting the University of Chicago Booth School of Business and Columbia University, and when Salanié was visiting the Toulouse School of Economics. Galichon thanks the Alliance program for its support, and Salanié thanks the Georges Meyer endowment. Galichon's research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement 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 generalizes Choo and Siow (2006). We first show that the stable matching maximizes a social gain function that trades off exploiting complementarities in observable characteristic sand matching on unobserved characteristics. We use this result to derive simple closed-form formulæ that identify the joint surplus in every possible match and the equilibrium utilities of all participants, given any known distribution of unobserved heterogeneity. If transfers are observed, then the pre-transfer utilities of both partners are also identified. We discuss computational issues and provide an algorithm that is extremely efficient in important instances. Finally, we present two estimators of the joint surplus and we revisit Choo and Siow's empirical application to illustrate the potential of our more general approach.

**Keywords**: matching, marriage, assignment, hedonic prices.

**JEL codes**: C78, D61, C13.

# 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. Assuming utility is transferable, the distributions of tastes and of desirable characteristics determine equilibrium shadow prices, which in turn explain how partners share the marital surplus in any realized match. This insight is not specific to the marriage market: it characterizes the "assignment game" of Shapley and Shubik (1972), i.e. models of matching with transferable utilities. These models have also been applied to competitive equilibrium in good markets with hedonic pricing (Chiappori, McCann and Nesheim, 2010), to trade (e.g. Costinot-Vogel 2014) and to the labour market (Terviö, 2008 and Gabaix and Landier, 2008.) Our results can be used in all of these 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 a simple sorting procedure. This sorting result is both simple and powerful; but its implications are also at variance with the data, in which matches are observed between partners with quite different characteristics. To account for a wider variety of matching patterns, one solution consists in allowing the matching surplus to incorporate latent characteristics—heterogeneity that is unobserved by the analyst. Choo and Siow (2006) have shown how it can be done in a way that yields a highly tractable model in large populations, provided that the unobserved heterogeneities enter the marital surplus quasi-additively and that they are distributed as standard type I extreme value terms. They used their model to evaluate the effect of the legalization of abortion on gains to marriage; and 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), to

compare the preference for marriage versus cohabitation (Mourifié and Siow, 2014) and, in a heteroskedastic version, to estimate the changes in the returns to education on the US marriage market (Chiappori, Salanié and Weiss, 2015.)

We revisit here the theory of matching with transferable utilities in the light of Choo and Siow's insights. Three assumptions underlie their contribution: latent variables do not mutually interact in producing matching surplus, they are distributed as iid type I extreme values, and populations are large. We maintain the first assumption, which we call "separability", and the last one which is innocuous in many applications. Choo and Siow's distributional assumption, on the other hand, is very special; it generates a multinomial logit model that has unappealing restrictions on cross-elasticities. We first show that this distributional assumption can be completely dispensed with. We prove that the optimal matching in our generalized setting maximizes the sum of a term that describes matching on the observables and a generalized entropic term that describes matching on the unobservables. While the first term tends to match partners with complementary observed characteristics, the second one pulls towards randomly assigning partners to each other. The social gain from any matching pattern trades off between these two terms. In particular, when unobserved heterogeneity is distributed as in Choo and Siow (2006), the generalized entropy is simply the usual entropy measure.

The maximization of this social surplus function has very straightforward consequences in terms of identification, both when equilibrium transfers are observed and when they are not. In fact, joint surplus and expected utilities can be obtained from derivatives of the terms that constitute generalized entropy; and that in turn is a function of observed matching probabilities. Moreover, if equilibrium transfers are observed, then we also identify the pre-transfer utilities on both sides of the market. In independent work, Decker et al. (2012) proved the uniqueness of the equilibrium and analyzed its properties in the Choo and Siow multinomial logit framework. We show that most of their comparative statics results in fact hold beyond the logit framework, and can be used as a testable prediction of all separable models. We also show how to derive testable predictions that are specific to a

given specification.

Our first conclusion thus is that the most important structural implications of the Choo-Siow model 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. 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. 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 second estimator also provides a very simple semi-parametric specification test.

Our third contribution is to revisit the original Choo and Siow 2006 study of the effect of the legalization of abortion on marriage outcomes in the US, making use of the new possibilities allowed by our extended framework. We estimate three different specifications; for each of them, we evaluate the impact of Roe vs Wade on expected gains from marriage; and we also test the performance of each estimated model out of sample. We find that Choo and Siow's "problematic finding" of more negative effects for men than for women is sensitive to the specification. One of our test models gives much more reasonable effects; it also behaves better out of sample. While these findings are of course specific to this

particular application, they illustrate that our approach is both practical and fruitful.

There are other approaches to estimating matching models with unobserved heterogeneity; see the handbook chapter by Graham (2011, 2013). For markets with transferable utility, Fox (2010) has proposed pooling data across many similar markets and relying on a "rank-order property": this assumes that given the characteristics of the populations of men and women, a given matching is more likely than another when it produces a higher expected surplus. Fox (2011) and Bajari and Fox (2013) applied this approach to the car industry and to spectrum auctions. More recently, Fox and Yang (2012) focus on identifying the complementarity between unobservable characteristics. A recent contribution by Menzel (2014) investigates large markets with non-transferable utility, and Aggarwal 2014 estimates matching in the US medical resident program.

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. Section 6 applies several instances of separable models to Choo and Siow's data. Finally, we present in Section 5 our IPFP algorithm, which greatly accelerates computations in important cases.

Our arguments use tools from convex analysis as well as optimal transportation. We have tried to keep the exposition intuitive in the body of the paper; all proofs can be found in Appendix A, to which we direct interested readers.

# 1 The Assignment Problem with Unobserved Heterogeneity

We study in this paper a bipartite, one-to-one matching market with transferable utility. We maintain throughout some of the basic assumptions of Choo and Siow (2006): utility transfers between partners are unconstrained, matching is frictionless, and there is no asym-

metric 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* $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_{ij})$ 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 either 0 or 1. However it is known (see e.g. Shapley and Shubik, 1972) that this integrality 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[1] 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 $\tilde{\boldsymbol{\mu}}$ such that

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

[1] Like most of the matching literature, we assume throughout that partners interact efficiently; del Boca and Flinn (2014) find empirical support for this assumption.

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

Any empirical application of this framework must start by acknowledging that the analyst only observes some of the payoff-relevant characteristics that determine the surplus matrix $\tilde{\mathbf{\Phi}}$. 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 $y_j \in \mathcal{Y}$. There is a finite number of groups; they are defined by the intersection of the characteristics which are observed by all men and women, and also by the analyst. On the other hand, men and women of a given group differ along some dimensions that they all observe, but which do not figure in the analyst's dataset.

Like Choo and Siow, we assume that there is an (uncountably) infinite number of men in any group $x$, and of women in any group $y$. We denote $n_x$ the mass of men in group $x$, and $m_y$ the mass of women in group $y$, 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 $S$ growing to infinity, that is $S = \sum_{x\in\mathcal{X}} N_x +$

$\sum_{y\in\mathcal{Y}} M_y \to +\infty$, while the proportion of each group remains constant, that is, the ratios $n_x = (N_x/S)$ and $m_y = (M_y/S)$ remain constant.

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 – we shall return to this in the conclusion. 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 can only compute quantities that depend on the observed groups of the partners in a match. Hence she cannot observe $\tilde{\boldsymbol{\mu}}$, and she must focus instead on the matrix of matches across groups $\boldsymbol{\mu} = (\mu_{xy})$, where $\mu_{xy} = \sum_{i,j} \mathbb{1}\left(x_i = x, y_j = y\right) \tilde{\mu}_{ij}$.

The feasibility constraints are $\boldsymbol{\mu} \in \mathcal{M}\left(\boldsymbol{n}, \boldsymbol{m}\right)$, where $\mathcal{M}\left(\boldsymbol{n}, \boldsymbol{m}\right)$ (or $\mathcal{M}$ in the absence of ambiguity) is the set of $|\mathcal{X}| \times |\mathcal{Y}|$ non-negative numbers $(\mu_{xy})$ that satisfy the $(|\mathcal{X}| + |\mathcal{Y}|)$ constraints that the number of married individuals in each group cannot be greater than the number of individuals in that group:

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

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

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

## 1.2 Matching Surpluses

Choo and Siow (2006) 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 $\alpha_{xy}$ is the systematic part of the surplus; $\tau$ represents the utility transfer (possibly negative) that the man gets from his partner in equilibrium; and $\varepsilon_{iy}$ is a standard type I extreme value random term. If such a man remains single, he gets utility $\varepsilon_{i0}$; that is to say, the systematic utilities of singles $\alpha_{x0}$ 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 $\tau$ is the utility transfer she leaves to her partner. Again, we normalize $\gamma_{0y} = 0$.

As shown in Chiappori, Salanié and Weiss (2015), 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* $\mathbf{\Phi}$ *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 $\mathbf{\Phi}$ is simply

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

which they call the *total systematic net gains to marriage*; and note that by construction, $\Phi_{x0}$ and $\Phi_{0y}$ 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 $\boldsymbol{\varepsilon}$ as a vector of preference shocks of the husband and $\boldsymbol{\eta}$ as a vector of preference shocks 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 $\tau$; 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[2]. In other words, there is no need to assume that the term $\varepsilon_{iy_j}$ is "generated" by man $i$, nor that the term $\eta_{jx_i}$ was "generated" by woman $j$; it may perfectly be the opposite.

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 $\boldsymbol{\Phi}$ (an array of $|\mathcal{X}| \times |\mathcal{Y}|$ unknown numbers) given the observation of $\boldsymbol{\mu}$ (an array of the same size.) In order to study the relation between these two objects, we need to make assumptions on the distribution of the unobserved heterogeneity terms.

### 1.3 Unobserved Heterogeneity

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

[2] 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 $\tau_i = \max_{y \in \mathcal{Y}} \varepsilon_{iy}$.

**Assumption 3** (Distribution of Unobserved Variation in Surplus)**.** *The unobserved heterogeneity in tastes is such that:*

*a) For any man $i$ such that $x_i = x$, the $|\mathcal{Y}_0|$-dimensional random vector $\boldsymbol{\varepsilon_i} = (\varepsilon_{iy})_y$ is drawn from a distribution $\boldsymbol{P}_x$;*

*b) For any woman $j$ such that $y_j = y$, the $|\mathcal{X}_0|$-dimensional random vector $\boldsymbol{\eta_j} = (\eta_{xj})_x$ is drawn from a distribution $\boldsymbol{Q}_y$.*

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

Assumption 3 constitutes a substantial generalization of Choo and Siow. It 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. The need to go beyond the logit framework has long been recognized in Industrial Organization and in consumer demand theory. This has led to a huge 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 an exposition and applications.) The present assumption is more general, as it does not require that the distribution of the terms $\boldsymbol{\varepsilon_i}$ and $\boldsymbol{\eta_j}$ belong to the GEV class.

## 2 Social Surplus, Utilities, and Identification

We derive most of our results by considering the "optimal" matching, which maximizes the total joint surplus. This is known since Shapley and Shubik (1972) to be identical 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 $\tilde{\boldsymbol{\mu}}$. But in order to find the expression of the value function that $\boldsymbol{\mu}$ 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 $\tilde{u}_i = \max_y \left(\Phi_{xy} + \varepsilon_{iy} - \min_{j:y_j=y} \left(\tilde{v}_j - \eta_{xj}\right)\right)$. Denoting

$$V_{xy} = \min_{j:y_j=y} \left(\tilde{v}_j - \eta_{xj}\right)$$

and $U_{xy} = \Phi_{xy} - V_{xy}$, it follows that:

**Proposition 1.** ***(Splitting the Surplus)***
*Under Assumptions 2 and 3, there exist two vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ 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 (2015), with a different proof. It reduces the two-sided matching problem to a series of one-sided discrete choice problems that are only linked through the adding-up formula $U_{xy} + V_{xy} = \Phi_{xy}$. Men of a given group $x$ match with women of different groups, since each man $i$ has idiosyncratic $\boldsymbol{\varepsilon_i}$ shocks. But as a consequence of the separability assumption, if a man of group $x$ matches with a woman of group $y$, then he would be equally well-off with any other woman of this group[3].

The vectors $\boldsymbol{U}$ and $\boldsymbol{V}$ depend on all of the primitives of the model (the vector $\boldsymbol{\Phi}$, the distributions of the utility shocks $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$, and the number of groups $\boldsymbol{n}$ and $\boldsymbol{m}$.) They are only a useful construct, and they should not be interpreted as utilities.

## 2.2 Identification of discrete choice problems

To recover the array $\boldsymbol{\Phi}$, we start by showing how the utilities $\boldsymbol{U}$ can be recovered from the choice probabilities $\mu_{y|x} = \mu_{xy}/n_x$. To do this, we introduce a general methodology based on "generalized entropy," a name which arises from reasons which will soon become clear. In the following, for any $\boldsymbol{A} = (A_{xy})$ we denote $\boldsymbol{A_{x\cdot}} = (A_{x1}, \ldots, A_{x|\mathcal{Y}|})$ and $\boldsymbol{A_{\cdot y}} = (A_{1y}, \ldots, A_{|\mathcal{X}|y})$.

Consider a randomly chosen man in group $x$, and a vector $\boldsymbol{U} = (U_{x0} = 0, U_{x1}, \ldots, U_{x,|\mathcal{Y}|})$. His expected utility (conditional to belonging to this group) is

$$G_x(\boldsymbol{U_{x\cdot}}) = \boldsymbol{E}_{\boldsymbol{P}_x} \max_{y \in \mathcal{Y}_0} (U_{xy} + \varepsilon_y), \tag{2.1}$$

where we set $U_{x0} = 0$ and the expectation is taken over the random vector $(\varepsilon_0, \ldots, \varepsilon_{|\mathcal{Y}|}) \sim \boldsymbol{P}_x$. First note that for any two numbers $a, b$ and random variables $(\varepsilon, \eta)$, 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 the obvious multidimensional generalization to the function $G_x$, we get

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

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

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 $y \in \mathcal{Y}_0$

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

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

$$G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = \max_{\tilde{\boldsymbol{U}}_{\boldsymbol{x}\cdot}=(\tilde{U}_{x1},\ldots,\tilde{U}_{x|\mathcal{Y}|})} \left( \sum_{y\in\mathcal{Y}} \mu_{y|x}\tilde{U}_{xy} - G_x(\tilde{\boldsymbol{U}}_{\boldsymbol{x}\cdot}) \right) \tag{2.3}$$

whenever $\sum_{y\in\mathcal{Y}} \mu_{y|x} \leq 1$, and $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = +\infty$ 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 $G_x^*$ we have

$$\frac{\partial G_x^*}{\partial \mu_{y|x}}(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = U_{xy}. \tag{2.4}$$

As a consequence, for any $y \in \mathcal{Y}$, $U_{xy}$ is identified from $\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}$, the observed matching patterns of men of group $x$. Going back to (2.3), convex duality implies that if $\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}$ and $\boldsymbol{U}_{\boldsymbol{x}\cdot}$ are related by (2.2), then

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \sum_{y\in\mathcal{Y}} \mu_{y|x}U_{xy} - G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}). \tag{2.5}$$

By the first order condition (2.2), the optimal $\boldsymbol{U}$ is such that $\mu_{y|x} = \partial G_x\left(U_{x\cdot}\right)/\partial U_{xy}$: $\boldsymbol{U}_{x\cdot}$ 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

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

and, comparing (2.2) with (2.5),

$$-G_x^*\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}\right) = \boldsymbol{E}\left(\varepsilon_{iY_i^*}\right). \tag{2.6}$$

This allows us to provide a useful characterization of $G_x^*\left(\mu_{\cdot|x}\right)$ and $\boldsymbol{U}$ and to show that they can be evaluated by solving an optimal matching problem:

**Proposition 2.** *(**General identification of the systematic surpluses**) Let* $\mathcal{M}\left(\boldsymbol{\mu}_{\cdot|x}, \mathbf{P}_x\right)$ *the set of probability distributions* $\pi$ *of the random joint vector* $(\boldsymbol{Y}, \boldsymbol{\varepsilon})$*, where* $\boldsymbol{Y} \sim \boldsymbol{\mu}_{\cdot|x}$ *is a random element of* $\mathcal{Y}_0$*, and* $\boldsymbol{\varepsilon} \sim \boldsymbol{P}_x$ *is a random vector of* $\mathbb{R}^{\mathcal{Y}_0}$*. For* $\boldsymbol{e} \in \mathbb{R}^{\mathcal{Y}_0}$ *and* $y \in \mathcal{Y}_0$*, let* $\Psi^h$ *denote the projection*

$$\Psi^h(y, \boldsymbol{e}) := e_y.$$

*Then* $-G_x^*(\boldsymbol{\mu}_{\cdot|x})$ *is the value of the optimal matching problem between the distribution* $\boldsymbol{\mu}_{\cdot|x}$ *of* $Y$ *and the distribution* $\boldsymbol{P}_x$ *of* $\boldsymbol{\varepsilon}$*, when the surplus is given by* $\Psi^h$*. That is,*

$$-G_x^*(\boldsymbol{\mu}_{\cdot|x}) = \max_{\pi \in \mathcal{M}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}, \boldsymbol{P}_x\right)} \boldsymbol{E}_\pi\left(\Psi^h(Y, \boldsymbol{\varepsilon})\right). \tag{2.7}$$

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

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

*is the Lagrange multiplier associated with the constraint* $\int \pi_y(\varepsilon) d\varepsilon = \mu_{y|x}$ *at the optimum.*

The intuition behind Proposition 2 is simply that each observed choice probability $\mu_{y|x}$ must be matched to the values of idiosyncratic preference shocks $\boldsymbol{\varepsilon}_i \sim \boldsymbol{P}_x$ for which $y$ is the most preferred choice. The $\boldsymbol{U}_{\cdot|x}$ are the shadow prices that support this matching.

## 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

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

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

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

But by identity (2.5), we get

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

Sum over the total number of men and women, use $\boldsymbol{U}+\boldsymbol{V}=\boldsymbol{\Phi}$ and define

$$\mathcal{E}(\boldsymbol{\mu}) := \sum_{x\in\mathcal{X}} n_x G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) + \sum_{y\in\mathcal{Y}} m_y H_y^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{y}}); \tag{2.8}$$

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

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

The first equality 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 $u_x = G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot})$, and similarly for women. The second equality shows 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,*

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

$$\mathcal{W}(\boldsymbol{\Phi},\boldsymbol{n},\boldsymbol{m}) = \max_{\boldsymbol{\mu}\in\mathcal{M}(\boldsymbol{n},\boldsymbol{m})} \left( \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} \mu_{xy}\Phi_{xy} - \mathcal{E}(\boldsymbol{\mu}) \right). \tag{2.9}$$

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

$$\begin{aligned} \mathcal{W}(\boldsymbol{\Phi},\boldsymbol{n},\boldsymbol{m}) \quad &= \quad \min_{\boldsymbol{U},\boldsymbol{V}} \left( \sum_{x\in\mathcal{X}} n_x G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) + \sum_{y\in\mathcal{Y}} m_y H_y(\boldsymbol{V}_{\cdot\boldsymbol{y}}) \right) \\ &\qquad s.t.\ U_{xy}+V_{xy}=\Phi_{xy}. \end{aligned} \tag{2.10}$$

*The function $\mathcal{W}(\boldsymbol{\Phi},\boldsymbol{n},\boldsymbol{m})$ is convex in $\boldsymbol{\Phi}$ and concave in $(\boldsymbol{n},\boldsymbol{m})$.*

*(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 $\boldsymbol{U}$ solves (2.10).*

*(iii) The average expected utility of the men of group $x$ is*

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

.

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

The right-hand side of equation (2.9) gives the value of the social surplus when the matching patterns are $\boldsymbol{\mu}$. The first term $\sum_{xy} \mu_{xy}\Phi_{xy}$ reflects "group preferences": if groups $x$ and $y$ generate more surplus when matched, then in the absence of unobserved heterogeneity they should be matched with higher probability. 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, $\boldsymbol{\Phi}$ 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 $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$ 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 $\boldsymbol{\mu}$ must maximize $\sum_{x,y} \mu_{xy}\Phi_{xy}$. If on the other hand data is so poor that unobserved heterogeneity dominates ($\boldsymbol{\Phi} \simeq 0$), 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 $\Phi_{xy}$) and some is carried by the unobserved heterogeneity terms $\varepsilon_{iy}$ and $\eta_{xj}$, the market equilibrium trades off matching on group characteristics against matching on unobserved characteristics, as measured by the generalized entropy terms in $\mathcal{E}(\boldsymbol{\mu})$.

## 2.4 Identification of Matching Surplus

Theorem 1 has two mathematically equivalent readings that have quite different practical purposes. One may use it to obtain the expression of $\boldsymbol{\mu}$ as a function of $\boldsymbol{\Phi}$: this is an *equilibrium characterization*. Conversely, one may use it to obtain the expression of $\boldsymbol{\Phi}$ as a function of $\boldsymbol{\mu}$: this *identifies* the joint surplus from the data. Our next result, Theorem 2, illustrates the mathematical duality between these two points of view.

First note that the constraints associated to $\boldsymbol{\mu} \in \mathcal{M}$ in (2.9) 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, $\boldsymbol{\mu}$ 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 $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ all have full support and are absolutely continuous with respect to the Lebesgue measure.*

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. Since the functions $G_x^*$ and $H_y^*$ are convex, they are differentiable almost everywhere—and under Assumption 4 they actually are differentiable everywhere. This is not essential to our approach; in fact, our Example 2 in section 3 violates Assumption 4. But it allows us to obtain striking formulæ, as stated in the following theorem:

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

*(i) $\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}$ and $\boldsymbol{U}_{\cdot|x}$ are related by the equivalent set of relations*

$$\mu_{y|x} = \frac{\partial G_x}{\partial U_{xy}}\left(\boldsymbol{U}_{\boldsymbol{x}\cdot}\right) \text{ for } y \in \mathcal{Y}, \text{ or equivalently} \tag{2.11}$$

$$U_{xy} = \frac{\partial G_x^*}{\partial \mu_{y|x}}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}\right) \text{ for } y \in \mathcal{Y}. \tag{2.12}$$

*(ii) As a result, the matching surplus $\boldsymbol{\Phi}$ is identified by*

$$\Phi_{xy} = \frac{\partial G_x^*}{\partial \boldsymbol{\mu}_{y|x}}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}\right) + \frac{\partial H_y^*}{\partial \mu_{x|y}}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{y}}\right), \tag{2.13}$$

*that is*

$$\Phi_{xy} = \frac{\partial \mathcal{E}}{\partial \mu_{xy}}(\boldsymbol{\mu}). \tag{2.14}$$

The dual nature of Theorem 2 is clear. The equalities (2.11) allow to express $\boldsymbol{\mu}$ as a function of $\boldsymbol{U}$ and $\boldsymbol{V}$ ("equilibrium characterization" point of view); they invert into relations (2.12) which express $\boldsymbol{U}$ and $\boldsymbol{V}$ (and thus $\boldsymbol{\Phi}$) as a function of $\boldsymbol{\mu}$ ("identification" point of view.)

The previous result does not assume that transfers are observed. When they are, the systematic parts of pre-transfer utilities $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ are also identified:

**Corollary 1.** *Under Assumptions 1, 2, 3 and 4, denote $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ the systematic parts of pre-transfer utilities and $\boldsymbol{\tau}$ the transfers as in Section 1. Then $\alpha_{xy}$ and $\gamma_{xy}$ are identified by*

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

*Therefore if transfers $\tau_{xy}$ are observed, both pre-transfer utilities $\alpha_{xy}$ and $\gamma_{xy}$ are also identified.*

This is unlikely to occur in family economics, where the econometrician typically does not observe transfers between partners. It is a much more reasonable asumption in other settings where matching theory has been successfully applied, as the CEO compensation literature, for instance, where the compensation amount is often available.

Combining Proposition 2 with Theorem 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 and testable predictions

Recall from Theorem 1 that the partial derivative of the social surplus $\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{n}, \boldsymbol{m})$ with respect to $n_x$ is $u_x$. It follows immediately that

$$\frac{\partial u_x}{\partial n_{x'}} = \frac{\partial u_{x'}}{\partial n_x};$$

therefore the "unexpected symmetry" result proven by Decker et al. (2012, Theorem 2) for the multinomial logit Choo and Siow model in fact holds for all separable models. We show in Appendix C that in fact most of the comparative statics results of Decker et al extend to the present framework, and can be used as testable predictions for the separable model.

Another remarkable feature of our results is that all of the functions involved are convex (or concave.) Since utilities are derivatives of these functions, their own Jacobian matrices must not only be symmetric but also satisfy positive- or negative-semi definiteness constraints. In particular, this implies the intuitive fact that,

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

This property against holds for all separable models. On the other hand, each separable model has its own specific predictions. As an illustration, hold $\boldsymbol{\Phi}$ and $\boldsymbol{m}$ fixed in the Choo and Siow model, and let the group sizes of men $\boldsymbol{n}$ vary. It is easy to see from (3.4) below that for all $(x, y)$ and all $z \neq x, y$,

$$2\frac{\partial \log \mu_{xy}}{\partial \log n_z} = \frac{\partial \log \mu_{x0}}{\partial \log n_z} + \frac{\partial \log \mu_{0y}}{\partial \log n_z}.$$

This identity is in principle testable by pooling data across similar markets with different group sizes for men. Other specifications would of course yield different predictions.

## 3 Examples

### 3.1 Models of heterogeneity

While Proposition 2 and Theorem 1 provide a general way of computing surplus and utilities, they can be derived in closed form in important special cases. In all formulæ below, the

proportions and 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 $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ distributions are iid standard type I extreme value:[4]

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

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

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

$$\mathcal{E}(\boldsymbol{\mu}) = \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}_0}} \mu_{xy}\log\mu_{y|x} + \sum_{\substack{y \in \mathcal{Y} \\ x \in \mathcal{X}_0}} \mu_{xy}\log\mu_{x|y}, \tag{3.3}$$

*which is a standard entropy[5]. Surplus and matching patterns are linked by*

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

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

The multinomial logit Choo and Siow model is the simplest example which fits into McFadden's Generalized Extreme Value (GEV) framework. This includes most specifications used in classical discrete choice models. A very simple instance is the heteroskedastic model

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

[5]The connection between the logit model and the classical entropy function is well known; see e.g. Anderson et al. (1988).

considered by Chiappori, Salanié and Weiss (2015); it allows the scale parameters of the type I extreme value distributions to vary across genders or groups. Then $\boldsymbol{P}_x$ has a scale parameter $\sigma_x$ and $\boldsymbol{Q}_y$ has a scale parameter $\tau_y$; 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

$$\Phi_{xy} = (\sigma_x + \tau_y) \log \mu_{xy} - \sigma_x \log \mu_{x0} - \tau_y \log \mu_{0y}. \tag{3.5}$$

An extension of this model is the well-known *nested logit model*, which we study as Example 3 in Appendix B.

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[6]. Closer to our framework, hedonic models also build on idiosyncratic preferences for observed characteristics, on both sides of a match[7]. Our setup allows for such specifications; we elaborate on the *mixed logit model* in Example 4 in Appendix B.

Assume, more generally, that men of group $x$ care for a vector of observed characteristics of partners $\boldsymbol{\zeta}_x(y)$, but the intensity of the preferences of each man $i$ in the group depends on a vector $\boldsymbol{\varepsilon}_i$ which is drawn from some given distribution. Then we could take $\boldsymbol{P}_x$ to be the joint distribution of $(\boldsymbol{\zeta}_x(y) \cdot \boldsymbol{\varepsilon}_i)_y$. We investigate a particular case of this specification in the next example: the Random Scalar Coefficient (RSC) model, where the dimension of $\boldsymbol{\zeta}_x(y)$ and $\boldsymbol{\varepsilon}_i$ is one, which makes computations very easy. Assuming further that the distribution of $\boldsymbol{\varepsilon}_i$ is uniform, we are 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.

**Example 2** (Random [Uniform] Scalar Coefficient (RSC/RUSC) models)**.** *Assume that for*

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

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

*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 $\varepsilon_i$'s are iid random variables which are assumed to be continuously distributed according to a c.d.f. $F_\varepsilon$ (which could also depend on $x$.) We call this model the Random Scalar Coefficient (RSC) model; and we show in Appendix B.2 that the entropy is*

$$\mathcal{E}(\boldsymbol{\mu}) = \sum_{xy} \mu_{xy} \left( \zeta_x(y) \bar{e}_x(y) + \xi_y(x) \bar{f}_y(x) \right), \tag{3.6}$$

*where $\bar{e}_x(y)$ is the expected value of $\varepsilon$ 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 $\varepsilon_i$ 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 $\boldsymbol{S}^x$ be the square matrix with elements $S^x_{yy'} = \max(\zeta_x(y), \zeta_x(y'))$ for $y, y' \in \mathcal{Y}_0$. Define $\boldsymbol{T}^x$ by $T^x_{yy'} = S^x_{y0} + S^x_{0y'} - S^x_{yy'} - S^x_{00}$, and let $\boldsymbol{sigma}^x_y = S^x_{00} - S^x_{y0}$.*

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

$$G^*_x(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = \frac{1}{2}(\boldsymbol{\mu}'_{\cdot|\boldsymbol{x}} \boldsymbol{T}^x \boldsymbol{\mu}_{\cdot|\boldsymbol{x}} + 2\boldsymbol{\sigma}^x . \boldsymbol{\mu}_{\cdot|\boldsymbol{x}} - S^x_{00}). \tag{3.7}$$

*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 $\boldsymbol{\mu}$:*

$$\mathcal{E}(\boldsymbol{\mu}) = \frac{1}{2}(\boldsymbol{\mu}' \boldsymbol{A} \boldsymbol{\mu} + 2\boldsymbol{B}' \boldsymbol{\mu} + c), \tag{3.8}$$

*where the expressions for $\boldsymbol{A}$, $\boldsymbol{B}$ and $c$ are given by (B.4)–(B.5) in Appendix B.2. As a consequence, the optimal matching solves a simple quadratic problem. See Appendix B.2 for details.*

The structure of heterogeneity in the RUSC/RSC models is related to that investigated in Ekeland et al. (2004) and Heckman et al. (2010), with continuous observed characteristics. In Ekeland et al. (2004) and Heckman et al. (2010), $y$ is a continuous and scalar quality parameter, and the heterogeneity terms is given by $\varepsilon_{iy} = \varepsilon_i y$. However, their identification strategy differs from ours. Ekeland et al. (2004) use an additively separable form of $\alpha(x, y)$ and are thus able to identify the distribution of $\varepsilon$, which is fixed in our setting. In contrast, In contrast, the assumptions in Heckman et al. (2010) assume that the distribution of $\varepsilon_i$ is fixed, and identification is obtained from a quantile transformation approach, which does not extend to the discrete case we presently investigate.

### 3.2 Discussion

The Choo-Siow multinomial logit model has the virtue of simplicity, but it relies on strong assumptions. This calls for caution when basing conclusions on it; we already pointed out some of its strong testable predictions in sectin 2.5. Expected utilities can also be a much richer function of observed matching patterns than in Choo and Siow's multinomial logit model.

Take the identified surplus for instance, whose expression is given in Choo and Siow's model by (3.4), which rewrites as

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

This shows that if we write $\boldsymbol{\Phi}$ as a function of $(\boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{m})$, then $\Phi_{xy}$ only depends on $\boldsymbol{\mu}$ via $\mu_{xy}$, $\sum_{y' \neq y} \mu_{xy'}$ and $\sum_{x' \neq x} \mu_{x'y}$. Therefore if $y' \neq y'' \neq y$,

$$\frac{\partial \Phi_{xy}}{\partial \mu_{xy'}} = \frac{\partial \Phi_{xy}}{\partial \mu_{xy''}}. \tag{3.9}$$

To interpret this, start from a given matching $\boldsymbol{\mu}$ which is rationalized by some surplus $\boldsymbol{\Phi}$, and suppose that a single man of group $x$ marries a single woman of group $y' \neq y$. Then (3.9) tells us that our estimator of the surplus $\Phi_{xy}$ should change by exactly the same amount as if that woman had belonged to any other group $y'' \neq y$. Derivations in Appendix B.2

show that in the RUSC model, the effect of changes in observed matching patterns on the estimated surplus $\partial\boldsymbol{\Phi}/\partial\boldsymbol{\mu}$ allows for much richer effects than (3.9).

# 4 Parametric Inference

First assume that all observations concern a single matching market. While the formula in Theorem 1(i) gives a straightforward nonparametric estimator of the systematic surplus function $\boldsymbol{\Phi}$, with multiple payoff-relevant observed characteristics $x$ and $y$ it is bound to be very unreliable. In addition, we do not know the distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$. Both of these remarks point to the need for a parametric model in most applications. Such a model would be described by a family of joint surplus functions $\Phi^{\boldsymbol{\lambda}}_{xy}$ and distributions $\boldsymbol{P}^{\boldsymbol{\lambda}}_x$ and $\boldsymbol{Q}^{\boldsymbol{\lambda}}_y$ for $\boldsymbol{\lambda}$ in some finite-dimensional parameter space $\Lambda$.

A dataset consists of a sample of $\hat{S} = \sum_x \hat{N}_x + \sum_y \hat{M}_y$ individuals, where $\hat{N}_x$ (resp. $\hat{M}_y$) denotes the number of men of group $x$ (resp. women of group $y$) in the sample, and $\hat{n}_x = \hat{N}_x/\hat{S}$ and $\hat{m}_x = \hat{M}_x/\hat{S}$ their respective empirical frequencies. We assume that the data was generated by the parametric model above, with true parameter vector $\boldsymbol{\lambda}_0$.

Recall the expression of the social surplus:

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

Let $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ be the optimal matching. We will show in Section 5 how it can be computed, in some cases very efficiently. For now we focus on statistical inference on $\boldsymbol{\lambda}$. We propose two methods: a very general Maximum Likelihood method, and a simpler moment-based method that is only available for some important subclasses of models.

## 4.1 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 probability of being matched with a woman of type $y$ is $\mu^{\boldsymbol{\lambda}}_{xy}/\hat{n}_x$; 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:

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

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

As an illustration, the MLE procedure is described for the nested logit model in Example 3 of Appendix B.3. Sometimes the expression of the likelihood $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ can be obtained in closed form. This is the case in the Random Uniform Scalar Coefficient model:

**Example 2 continued.** *Assume that the data generating process is the RUSC model of Example 2. We parameterize $\boldsymbol{\Phi}, \boldsymbol{\zeta}_x$ and $\boldsymbol{\xi}_y$ by a parameter vector $\boldsymbol{\lambda}\in\mathbb{R}^K$, hence parameterizing $\boldsymbol{S}$ and $\boldsymbol{T}$ and thus $\boldsymbol{A}$ and $\boldsymbol{B}$. If the optimal matching $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ is interior, then it is given by $\boldsymbol{\mu}^{\boldsymbol{\lambda}}=(\boldsymbol{A}^{\boldsymbol{\lambda}})^{-1}\left(\boldsymbol{\Phi}^{\boldsymbol{\lambda}}-\boldsymbol{B}^{\boldsymbol{\lambda}}\right)$ and its log-likelihood can be deduced by (4.1).*

Maximum likelihood estimation has the usual statistical advantages; and it allows for joint parametric estimation of the surplus function and of the unobserved heterogeneity. However, the log-likelihood may have several local extrema and be hard to maximize. In such cases, an alternative, moment-based method may be more appealing.

## 4.2 Moment-based estimation: The Linear Model

We now introduce a method based on moments which is computationally very efficient but can only be used under two additional assumptions.

We first make the restrictive assumption that the distribution of the unobservable heterogeneity is parameter-free—as it is in Choo and Siow 2006 for instance; or at least we conduct the analysis for fixed values of its parameters. Then we take a linear parameterization of the $\boldsymbol{\Phi}$ matrix:

$$\Phi^{\boldsymbol{\lambda}}_{xy} = \sum_{k=1}^{K} \lambda_k \phi^k_{xy} \tag{4.2}$$

where the parameter $\boldsymbol{\lambda} \in \mathbb{R}^K$ and $\boldsymbol{\phi^1}, \ldots, \boldsymbol{\phi^K}$ are $K$ known linearly independent *basis surplus vectors*. If the number of basis surplus vectors expands, this specification can be seen as a sieve approximation that can converge to any surplus function. We do not pursue this, and we assume that the true $\boldsymbol{\Phi}$ takes the form (4.2) for finite $K$.

For any feasible matching $\boldsymbol{\mu}$, we define the associated *comoments* as the average values of the basis surplus vectors:

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

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

The *Moment Matching estimator* of $\boldsymbol{\lambda}$ we propose in this section simply matches the comoments predicted by the model with the empirical comoments; that is, it solves the system

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

It is in fact more useful to define it as

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

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

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

*(i) The Moment Matching estimator yields the equality of predicted comoments and observed comoments: the equality $C^k(\hat{\boldsymbol{\mu}}) = C^k(\mu^{\boldsymbol{\lambda}})$ holds for all $k$ for $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}^{MM}$.*

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

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

Both (4.3) and (4.4) are globally convex programs, and hence very easy to solve numerically. The intuition for the result is simple: we know from (2.9) that if $\boldsymbol{\mu}$ is optimal for $\boldsymbol{\Phi}$, then

$$\frac{\partial \mathcal{W}}{\partial \boldsymbol{\Phi}} = \boldsymbol{\mu}.$$

Hence the first-order condition in (4.3) is simply $(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{\boldsymbol{\lambda}}) \cdot \boldsymbol{\phi} = 0$.

While the Moment Matching estimator is not asymptotically efficient in general, it coincides with the MLE in the Choo and Siow model (see Appendix B.1):

**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* $\boldsymbol{\Phi}^{\boldsymbol{\lambda}}$ *is linearly parameterized as in (4.2). Then the log-likelihood can be written as*

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

*Therefore in this setting the Conditional Maximum Likelihood estimator and the Moment Matching estimator coincide:* $\hat{\boldsymbol{\lambda}}^{MM} = \hat{\boldsymbol{\lambda}}^{MLE}$.

This equivalence is particular to the Choo and Siow multinomial logit setting. It does not obtain in the RUSC model for instance; but the latter is interesting as one can obtain an explicit expression of $\hat{\boldsymbol{\lambda}}^{MM}$ in the "interior" case in which no cell is empty ($\mu_{xy} > 0$ for all $(x, y)$)—see Appendix B.2.

Finally, Part (ii) of Theorem 3 generates a very simple specification test. Compare the *actual* value $\mathcal{E}(\hat{\boldsymbol{\mu}})$ of the generalized entropy associated to the empirical distribution to the value $\mathcal{E}_{\min}(\hat{\boldsymbol{\mu}})$ of the program (4.4). By definition, $\mathcal{E}(\hat{\boldsymbol{\mu}}) \geq \mathcal{E}_{\min}(\hat{\boldsymbol{\mu}})$; moreover, these two

values coincide if and only if the model is well-specified. We state this in the following proposition:[8]

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

### 4.3 Parameterization, Testing, and Multimarket Data

Theorem 1 shows that, given a specification of the distribution of the unobserved heterogeneities $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$, there is a one-to-one correspondence between $\boldsymbol{\mu}$ amd $\boldsymbol{\Phi}$. Any matching on a single market can be rationalized by exactly one model that satisfies assumptions 1, 2, and 3. The downside of this result is that it is impossible to test separability using only data on one market, even assuming perfect knowledge of the distributions of unobserved heterogeneity.

Of course, assuming the distribution of the unobserved heterogeneity terms is a strong assumption, while on the other hand, we may be content with a parametric specification of $\boldsymbol{\Phi}$. Typically, we will use much fewer than the $|\mathcal{X}| \times |\mathcal{Y}|$ degrees of freedom of a nonparanetric $\boldsymbol{\Phi}$. Any unused degrees of freedom can be used for inference on the distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$, appropriately parameterized, or in order to test the assumptions of the model. For example, if $\mathcal{X}$ and $\mathcal{Y}$ are finite subsets of $\mathbb{R}^d$, we could resort to a semiparametric specification in the spirit of Ekeland et al. (2004): $\Phi_{xy} = \phi_1(y) + y'\phi_2(x)$. This would restrict the number of degrees of freedom in $\boldsymbol{\Phi}$ to $|\mathcal{Y}| + d \times |\mathcal{X}|$.

An alternative empirical strategy is to use multiple markets with restricted parametric variation in the joint surplus $\boldsymbol{\Phi}$ and the distributions of unobserved heterogeneity $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$. The variations in the groups sizes $\boldsymbol{n}$ and $\boldsymbol{m}$ across markets then generate variation in optimal matchings that can be used to overidentify the model and generate testable

[8] The critical values of the test can be obtained by parametric bootstrap for instance. One could also run the test for different specifications of the distributions of heterogeneities and invert it to obtain confidence intervals for the parameters of $\mathbf{P}_x$ and $mathbfQ_y$.

restrictions. Chiappori, Salanié and Weiss (2015) pursue a simple variant of this strategy.

## 5 Computation

Maximizing the conditional likelihood requires computing the optimal matching $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ for a large number of values of $\boldsymbol{\lambda}$. But the optimal matching is a large object in realistic applications; and obtaining it by maximizing $\mathcal{W}$ in (2.9) is not a very practical option. We develop here an algorithm based on the Iterative Projection Fitting Procedure (IPFP) that often provides a much more efficient solution.

Take the multinomial logit Choo-Siow model of Example 1 for instance. Fix a value of $\boldsymbol{\lambda}$ and drop it from the notation: let the joint surplus function be $\boldsymbol{\Phi}$, with optimal matching $\boldsymbol{\mu}$. Formula (3.4) can be rewritten as

$$\mu_{xy} = \exp\left(\frac{\Phi_{xy}}{2}\right)\sqrt{\mu_{x0}\mu_{0y}}. \tag{5.1}$$

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 $\mu_{x0}$ and $\mu_{0y}$. Unfortunately, this results in a system of $|X| + |Y|$ quadratic equations in as many unknowns that is quite unwieldy as soon as we consider more than a few groups.

To find a feasible solution of (3.4), an alternative approach is to start from an infeasible solution and to project it on the set of feasible matchings $\mathcal{M}(\hat{n}, \hat{m})$. The difficulty is to do this and still satisfy (5.1); we now show how it can be done using IPFP. As its name indicates, the Iterative Projection Fitting Procedure was designed to find projections on intersecting sets of constraints, by projecting iteratively on each constraint[9].

Assume that there exists a convex function $E$ that extends the generalized entropy $\mathcal{E}$ in the following manner: $E(\bar{\boldsymbol{\mu}})$ is defined for all non-negative $\bar{\boldsymbol{\mu}} = \big(\boldsymbol{\mu} = (\mu_{xy})_{x,y}, (\mu_{x0})_x, (\mu_{0y})_y\big)$; and it coincides with $\mathcal{E}(\boldsymbol{\mu})$ whenever

$$\sum_y \mu_{xy} + \mu_{x0} = n_x \;\text{ and }\; \sum_x \mu_{xy} + \mu_{0y} = m_y \;\text{ for all } x, y. \tag{5.2}$$

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

Problem (2.9) can be rewritten as the maximization of $\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}\Phi_{xy}-E\left(\bar{\boldsymbol{\mu}}\right)$ over the set of vectors $\bar{\boldsymbol{\mu}}\geq 0$ that satisfy the constraints in (5.2). Denoting $u_x$ and $v_y$ the Lagrange multipliers of the constraints, and introducing the Lagrangian

$$\mathcal{L}(\bar{\boldsymbol{\mu}},\boldsymbol{u},\boldsymbol{v})=\sum_{x\in\mathcal{X}}u_x\left(n_x-\sum_{y\in\mathcal{Y}_0}\mu_{xy}\right)+\sum_{y\in\mathcal{Y}}v_y\left(m_y-\sum_{x\in\mathcal{X}_0}\mu_{xy}\right)+\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\mu_{xy}\Phi_{xy}-E\left(\bar{\boldsymbol{\mu}}\right), \tag{5.3}$$

the value of the matching problem is

$$\min_{\boldsymbol{u},\boldsymbol{v}}\max_{\bar{\boldsymbol{\mu}}\geq 0}\mathcal{L}(\bar{\boldsymbol{\mu}},\boldsymbol{u},\boldsymbol{v}) \tag{5.4}$$

and its first order conditions are

$$\frac{\partial E}{\partial\mu_{xy}}=\Phi_{xy}-u_x-v_y\ ;\ \frac{\partial E}{\partial\mu_{x0}}=-u_x\ ;\ \frac{\partial E}{\partial\mu_{0y}}=-v_y.$$

However, instead of solving the full problem (5.4), we solve it iteratively. Start with any initial choice of $\left(\bar{\boldsymbol{\mu}}^{(0)}\boldsymbol{u}^{(0)},\boldsymbol{v}^{(0)}\right)$ and set $k=0$. After completing step $2k$ with values $(\bar{\boldsymbol{\mu}}^{(2k)},\boldsymbol{u}^{(2k)},\boldsymbol{v}^{(2k)})$, at step $(2k+1)$ we keep $\boldsymbol{v}=\boldsymbol{v}^{(2k)}$ and we solve the minmax problem over $\boldsymbol{u}$ and $\bar{\boldsymbol{\mu}}$:

$$\min_{\boldsymbol{u}}\max_{\bar{\boldsymbol{\mu}}\geq 0}\mathcal{L}(\bar{\boldsymbol{\mu}},\boldsymbol{u},\boldsymbol{v}^{(2k)}). \tag{5.5}$$

Using the first order conditions, this is equivalent to solving

$$\frac{\partial E}{\partial\mu_{xy}}(\bar{\boldsymbol{\mu}})=\Phi_{xy}-u_x-v_y^{(2k)}\ ;\ \frac{\partial E}{\partial\mu_{x0}}(\bar{\boldsymbol{\mu}})=-u_x\ ;\ \frac{\partial E}{\partial\mu_{0y}}(\bar{\boldsymbol{\mu}})=-v_y^{(2k)} \tag{5.6}$$

along with $\sum_{y\in\mathcal{Y}_0}\mu_{xy}=n_x$. Call $\boldsymbol{u}^{(2k+1)}$ and $\bar{\boldsymbol{\mu}}^{(2k+1)}$ the solutions to this problem..

At step $(2k+2)$, we keep $\boldsymbol{u}=\boldsymbol{u}^{(2k+1)}$ and we solve the minmax problem over $\boldsymbol{v}$ and $\bar{\boldsymbol{\mu}}$:

$$\min_{\boldsymbol{v}}\max_{\bar{\boldsymbol{\mu}}\geq 0}\mathcal{L}(\bar{\boldsymbol{\mu}},\boldsymbol{u}^{(2k+1)},\boldsymbol{v}); \tag{5.7}$$

which amounts to solving

$$\frac{\partial E}{\partial\mu_{xy}}(\bar{\boldsymbol{\mu}})=\Phi_{xy}-u_x^{(2k+1)}-v_y\ ;\ \frac{\partial E}{\partial\mu_{x0}}(\bar{\boldsymbol{\mu}})=-u_x^{2k+1}\ ;\frac{\partial E(\boldsymbol{\mu})}{\partial\mu_{0y}}=-v_y \tag{5.8}$$

along with $\sum_{x\in\mathcal{X}_0}\mu_{xy}=m_y$. Call $\boldsymbol{v}^{(2k+2)}$ and $\bar{\boldsymbol{\mu}}^{(2k+2)}$ the solutions.

If $\bar{\boldsymbol{\mu}}^{(2k+2)}$ is close enough to $\bar{\boldsymbol{\mu}}^{(2k)}$, then we take $\bar{\boldsymbol{\mu}} = \bar{\boldsymbol{\mu}}^{2k+2}$ to be the optimal matching and we stop; otherwise we add one to $k$ and we move to step $(2k+3)$.

Note that the algorithm can be interpreted as a Walrasian tâtonnement process where the prices of the observed characteristics $x$ 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 $\bar{\boldsymbol{\mu}}$ of (2.9) and to the corresponding expected utilities.*

Note that there are many possible ways of extending $\mathcal{E}$ (which is defined only on $\mathcal{M}$) to the entire space of $\bar{\boldsymbol{\mu}} \geq 0$. In practice, good judgement should be exercised, as the choice of an extension $E$ that makes it easy to solve (5.6) and (5.8) is crucial for the performance of the algorithm.

To illustrate, take the multinomial logit Choo and Siow model from Example 1. Here we extend $\mathcal{E}(\boldsymbol{\mu})$ to

$$E\left(\bar{\boldsymbol{\mu}}\right) = \sum_{x\in\mathcal{X}} \sum_{x\in\mathcal{Y}_0} \mu_{xy} \log \mu_{xy} + \sum_{x\in\mathcal{X}_0} \sum_{x\in\mathcal{Y}} \mu_{xy} \log \mu_{xy},$$

and we have $\partial E/\partial \mu_{xy} = 2 + 2\log \mu_{xy}$, $\partial E/\partial \mu_{x0} = 1 + \log \mu_{x0}$, and $\partial E/\partial \mu_{0y} = 1 + \log \mu_{0y}$.

Therefore solving (5.6) at step $(2k+1)$ boils down to the simple task of finding the root $\mu_{x0}$ of the univariate quadratic equation

$$\mu_{x0} + \sqrt{\mu_{x0}\mu_{0y}^{(2k)}} \exp\left(\frac{\Phi_{xy}}{2}\right) = n_x. \tag{5.9}$$

We illustrate the algorithm for the nested logit model in Example 3 in Appendix B.3. We tested the performance of our proposed algorithm on an instance of Choo and Siow problem; we report the results in Appendix E. Our IPFP algorithm is extremely fast compared to standard optimization or equation-solving methods.

## 6 Empirical Application

Choo and Siow 2006 illustrated the potential of separable matching models by taking the specification of Example 1 to US Census data before and after the 1973 *Roe vs Wade* Supreme Court ruling that barred state restrictions on abortion before the third semester of pregnancy. As some states (which CS call the "non-reform states") had already relaxed restrictions on abortion, the ruling had fewer effects there than in the "reform states". This naturally suggests a difference-of -difference approach in which the expected utilities of groups of women and men on the marriage markets of reform and non-reform states are estimated before and after the ruling. To do this, Choo and Siow pooled data from the 1970 Census and the marriage records of 1971 and 1972 to describe the marriage markets before the 1973 ruling, and the 1980 Census and the marriage records of 1981 and 1982.

Our aim in this section is to compare three different approaches to this data[10]:

1. using the same specification as CS (nonparametric Choo and Siow, or "CS NP" in the figures)
2. using a parametric specification of their model (parametric Choo and Siow, "CS P")
3. and finally, using the RUSC specification of Example 2 ("RUSC".)

We concentrate throughout on men and women aged 16 to 40 (this covers more than 90% of marriages in this period.) In each case, we estimate four models of the marriage market—two for the 1970s data (in reform and in non-reform states) and two for the 1980s. In cases 2 and 3, we use model selection criteria to choose a specification. Then we take the diff-of-diff to estimate the effect of Roe vs Wade on the expected utilities from marriage of men and women of different ages. Finally, we estimate two models on marriage data from 1981 alone, and we use the estimated model to predict marriages in 1982, when the numbers of available

[10]We are very grateful to Eugene Choo and Aloysius Siow for giving us access to their data and code. We corrected a small mistake in the construction of the data—CS did not update the ages of the subjects between Census year+1 and Census year+2. This does not affect their conclusions.

men and women for each age have changed. This last exercise allows us to evaluate the comparative statics of the models, as well as their performance out-of-sample.

## 6.1 Implementing the Specifications

### 6.1.1 Nonparametric Choo and Siow

Expected utilities in Example 1 are simply minus the logarithms of the marriage rates and are therefore extimated straightforwardly. We also estimated the joint surplus matrix $\mathbf{\Phi}$ from marriages in 1981; and we used these estimates to project marriages in 1982, given the numbers of available men and women at each age in 1982[11].

### 6.1.2 Parametric Choo and Siow

We also estimated parameterized versions of the Choo and Siow models, writing

$$\Phi_{xy} = X_{xy}\boldsymbol{\lambda}$$

where the $X$ are polynomials of the ages of the two partners. For each value $\boldsymbol{\lambda}$, we computed the optimal matching using the IPFP algorithm of section 5; and we maximized the resulting log-likelihood function. To choose the polynomial terms to be included in the covariates $X$, we turned to standard model selection techniques. Selecting models by the Akaike or the Bayesian information criteria only works for unweighted maximum likelihood estimation. Since the surveys we use have fairly large variation in sampling weights, we needed to resort to a more general version of AIC: the Takeuchi Information Criterion. TIC applies even if the models may all be misspecified and for other estimation methods than MLE. Let $\bar{\theta}_k$ be the pseudo-value for model $k$, and $I_k$ and $J_k$ be the usual matrices. Then AIC would minimize

$$-2\log L_k + 2p_k;$$

[11] This required imputing numbers for the youngest men and women in 1982, since by then the availables of 1980 are aged at least 18; we used the same numbers that we observed in 1980, and we assigned to them the same marriage rates as in 1981.

we should instead minimize

$$-2\log L_k + 2\mathrm{Tr}(I_k J_k^{-1}).$$

With variable sampling weights, these two criteria give quite different results. We applied TIC to select from a set of regressors in ages of men and women of the form $A_m^k A_w^l$. The selected models allow for terms of total degrees up to $k + l = 5$. Table 1 shows the main characteristics of the selected models. The "maximum $\log L$" column of the table shows the value of the log-likelihood for the nonparametric specification of Example 1. The parametric models achieve a very good fit on the whole[12].

| Wave | Reform | Zero cells | maximum $\log L$ | Parametric CS | | RUSC | |
|---|---|---|---|---|---|---|---|
| | | | | Parameters | $\log L$ | Parameters | $\log L$ |
| 1970 | N | 1.9% | -1.04 | 14 | -1.05 | 21 | -1.07 |
| 1970 | Y | 1.8% | -1.00 | 14 | -1.02 | 21 | -1.03 |
| 1980 | N | 0.4% | -0.84 | 14 | -0.85 | 21 | -0.86 |
| 1980 | Y | 0.2% | -0.72 | 19 | -0.73 | 21 | -0.82 |
| 1981 | N | 3.4% | -0.48 | 13 | -0.49 | 21 | -0.49 |
| 1981 | Y | 2.6% | -0.45 | 13 | -0.46 | 21 | -0.48 |

Table 1: Comparing the Models

### 6.1.3 RUSC

Finally, we turn to the RUSC specification of Example 2. One of the distinguishing features of this specification is that the error terms have compact support. As such and unlike the Choo and Siow models, it allows for zero matching probabilities. This is easily seen by referring to the formula for the entropy in Example 2: if the solution $\boldsymbol{\mu}$ has all of its components positive, then it is given by the first-order condition

$$\boldsymbol{\mu} = \boldsymbol{A}^{-1}(\boldsymbol{\Phi} - \boldsymbol{B}); \tag{6.1}$$

[12] In the Choo and Siow model, it follows from the example on page 27 that the entropy-based specification test of Proposition 3 is just a likelihood-ratio test, and we did not implement it.

and if a component of the right-hand side of this equality is negative, this is impossible. Suppose for instance that $(\boldsymbol{\Phi} - \boldsymbol{B})$ has at least one negative component; then since all elements of $\boldsymbol{A}$ and $\boldsymbol{\mu}$ are non-negative, the corresponding first-order condition cannot hold. Note also that this is by no means a non-generic case: it holds on open sets of parameter values.

Allowing for zero matching probabilities is both an asset and a complication. It is an asset because much economic data in fact contains zero-probability cells. In the Choo and Siow data, several of the subsamples we use have some zero cells[13], as shown in the third column of Table 1. But finite-support errors, which make zero probabilities possible, also beget numerical difficulties. To start with, the log-likelihood element for cell $(x, y)$ is $\hat{\boldsymbol{\mu}}_{xy} \log \mu_{xy}(\boldsymbol{\lambda})$; parameter values that predict $\mu_{xy}(\boldsymbol{\lambda}) = 0$ when $\hat{\mu}_{xy} > 0$ will give (minus) infinite log-likelihood. Moreover, the finite support generates min and max operators in the definitions of $A$ and $B$ (see Appendix B.2.) For instance,

$$A_{xy,xy'}(\boldsymbol{\lambda}) = \min(\zeta_x(y; \boldsymbol{\lambda}), \zeta_x(y'; \boldsymbol{\lambda}))$$

if the $\zeta$'s are positive. At any parameter values where two values of $\zeta_x$ coincide, the resulting $\boldsymbol{\mu}$ will be a non-differentiable function of $\boldsymbol{\lambda}$; and the log-likelihood function will have a kink. There are many such kinks, and this creates serious difficulties with the usual gradient-based maximization algorithms. It is relatively easy to compute derivatives analytically where they exist, but this only gives misleading information to the algorithm. We experimented with algorithms that do not rely on derivatives; but they were very slow and got stuck at clearly non-optimal points. Not having good initial parameter values for the $\zeta_x(y; \boldsymbol{\lambda})$ and $\xi_y(x; \boldsymbol{\lambda})$ only compounded the problem.

We resorted to a hybrid strategy to obtain good estimates of the parameters. We used separate parameter vectors $\boldsymbol{\lambda}_\zeta, \boldsymbol{\lambda}_\xi$ and $\boldsymbol{\lambda}_\Phi$ for the $\zeta, \xi$ and $\boldsymbol{\Phi}$ components of the RUSC model. For fixed parameter values $(\boldsymbol{\lambda}_\zeta, \boldsymbol{\lambda}_\xi)$, we optimized the log-likelihood[14] over $\boldsymbol{\lambda}_\Phi$,

[13] Trade is another area where matching methods have become popular in recent years (see Costinot–Vogel 2014); and trade data also has typically many zero cells.

[14] Since the parameter values sometimes predicted zero probabilities for cell where marriages were actually

starting from the parameters that best fit (6.1) and using a derivative-free algorithm. To find the optimal $\boldsymbol{\mu}$ for any $\boldsymbol{\lambda}$, we used the Gurobi quadratic optimizer[15].

Getting a set of estimates using this process is too slow to replicate the kind of extensive model selection we used for the parametric CS model. We settled on a quadratic specification of the $\zeta_x(y)$ and $\xi_y(x)$ in the ages of both partners, with polynomials of total degree three of less for $\Phi_{xy}$. The estimated models have a total of 21 parameters.

To compute expected utilities $u_x$ and $v_y$ in the RUSC model, we used the differential equality

$$u_x = \frac{\partial \mathcal{W}}{\partial n_x}(\boldsymbol{\Phi}, \boldsymbol{n}, \boldsymbol{m})$$

of Theorem 1(ii) and we combined it with (3.8).

### 6.2 Diff-of-diff Estimates

Figure 1 computes the impact of Roe vs Wade on expected utilities for men and women aged 16 to 40, along with 95% confidence bars for the CS models[16]. CS did not report standard errors for their estimates; correct standard errors must take into account the sampling errors and weights, and the complex correlation structure of sampling errors as they go through logarithmic transforms and double differences. Even with such large samples (the sampling rates vary from 1/2 to 1/50), many of the diff-of-diff estimates in the first row of Figure 1 are quite imprecise. In particular, the effect of Roe vs Wade on women's expected utilities from marriage is small and insignificant at most ages. On the other hand, the fall in expected utilities of young men is large and is highly significant at their peak marriage ages. As noted by Choo and Siow (page 193), this is a "problematic" result; changes in abortion laws should have a larger impact on women than on men.

As one would expect, the diff-of-diff estimates from the parametric specification of the

observed, we replaced the logarithm function by a $C^2$ extension for $x < \delta$, so that it only took finite values. Then we gradually took $\delta$ to 0.

[15] See Gurobi (2015).

[16] Given the non-differentiable log-likelihood function for the RUSC model, we did not attempt to compute standard errors.

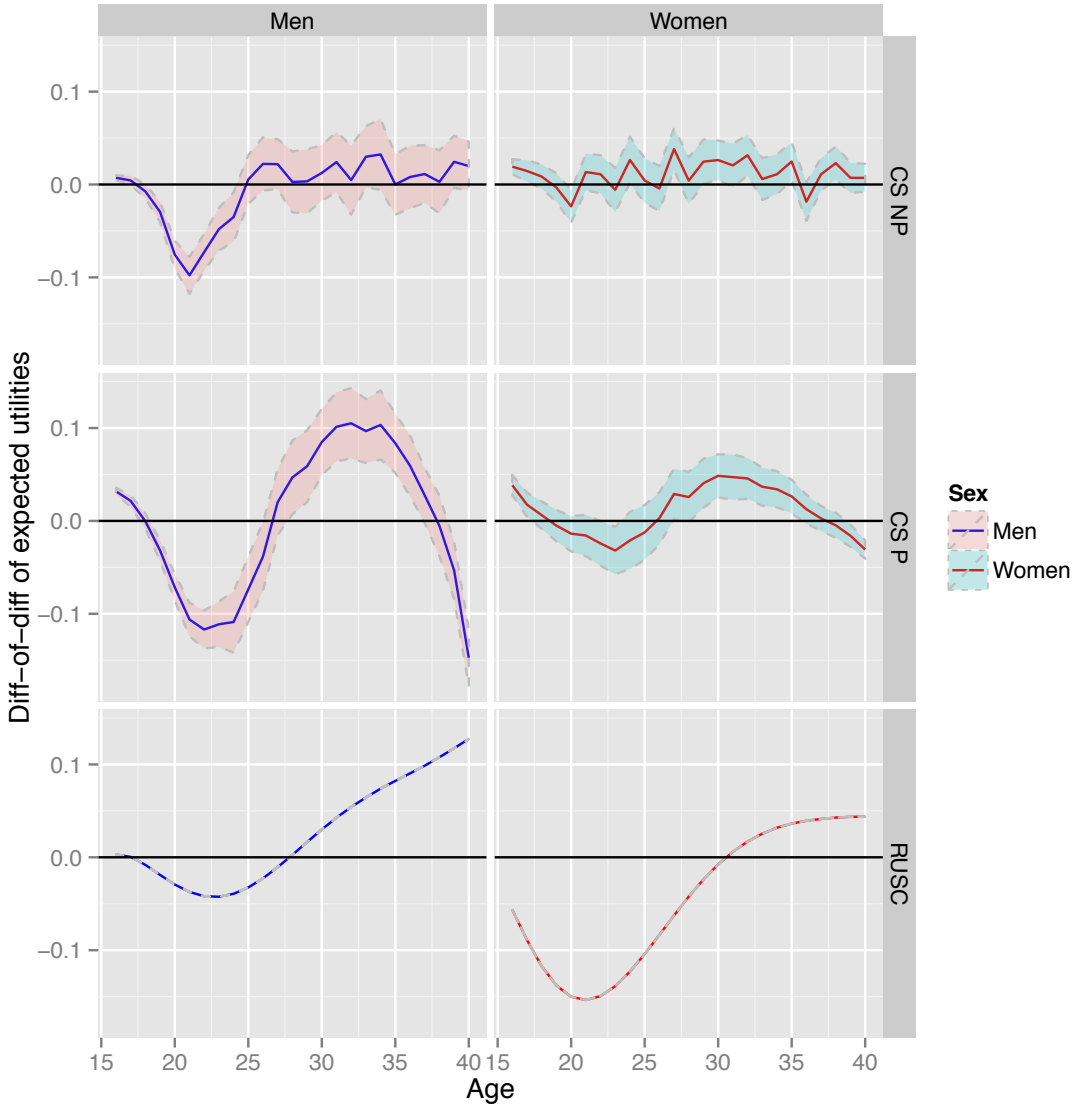


Figure 1: Effect of Roe vs Wade on Expected Utilities from Marriage

CS model are very similar, but they are both smoother and more precisely estimated—see the second row of Figure 1.

Interestingly, the effect of Roe vs Wade estimated from the RUSC model (third row) has a pattern that is much closer to what intuition suggests: the liberalization of abortion had a more negative effect on the utility from marriage of young women than on that of young men. The left-hand panel (for men) is very similar to those on the first two rows, but the right-hand panel shows a much larger negative effect for young women than for young men[17].

### 6.3 Predicting Marriages in 1982

For each of our three models, the simplest way of predicting marriages in 1982 is to simply project the marriage rates at each age simulated from the 1981 estimates on the observed margins, that is the men and women who are not married at the beginning of 1981. We call this the *naive* predictor, as it assumes away any change in marriage behavior between 1981 and 1982.

Remember that in each model the optimal matching $\boldsymbol{\mu}$ is a function of both the joint surplus $\boldsymbol{\Phi}$ and the margins $\boldsymbol{n}$ and $\boldsymbol{m}$. Changes in the margins act both by simply scaling up or down the supply of available partners, and by changing the allocation of the surplus and therefore the equilibrium marriage patterns. The naïve predictor neglects the second, "price" effect. Our *model* predictor allows for it, by simulating the marriage rates for the margins observed in 1982.

We implemented these two predictors for each model and in both categories of states; and we focus here on the predicted numbers of marriages at each age for both genders. Figure 2 plots the prediction errors (predicted minus observed) for the numbers of marriages in 1982.

In each panel, the difference between the dashed curve and the solid curve corresponds to the "price effects" induced by the change in margins betwen 1981 and 1982. These price

[17] Unlike the CS models, we need to normalize the scale of utilities in the RUSC model. We did it so that the range of diff-of-diff effects is similar; but this is arbitrary.

effects are much smaller in non-reform states than in reform states. A look at the data explains why: while the total number of available individuals did not change much in non-reform states, it decreased markedly in non-reform states; and this decrease was mostly due to women. With fewer women, the expected utilities of men fall; and the price effect increases the number of predicted marriages for women and reduces it for men. This is exactly what the second and fourth column show consistently. The changes in margins in non-reform states turn out to be too small for the price effects to show.

The second row of Figure 2 shows that the parametric Choo and Siow models behave very badly out of sample: they underpredict marriages by about 5% at the peak ages of 20-25. The RUSC model performs much better in this respect, even if their performance at the youngest ages (18-20) is not very good.

## 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 distributions $\boldsymbol{P}_x$ and $\boldsymbol{Q}_y$ and a parametrization $\boldsymbol{\lambda}$ of $\boldsymbol{U}$ and $\boldsymbol{V}$, and fitting the multinomial choice model where for instance men maximize $U_{xy}(\boldsymbol{\lambda})+\varepsilon_{iy}$ over their marital options $y \in \mathcal{Y}_0$. The downside is that unlike the joint surplus, the utilities $\boldsymbol{U}$ and $\boldsymbol{V}$ are not primitive objects; and they cannot even be interpreted as utilities, and it is hard to argue for a particular specification.

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; given the high dimensionality of unobservable shocks, their method, while very ingenious, has yet to be tested on real data.

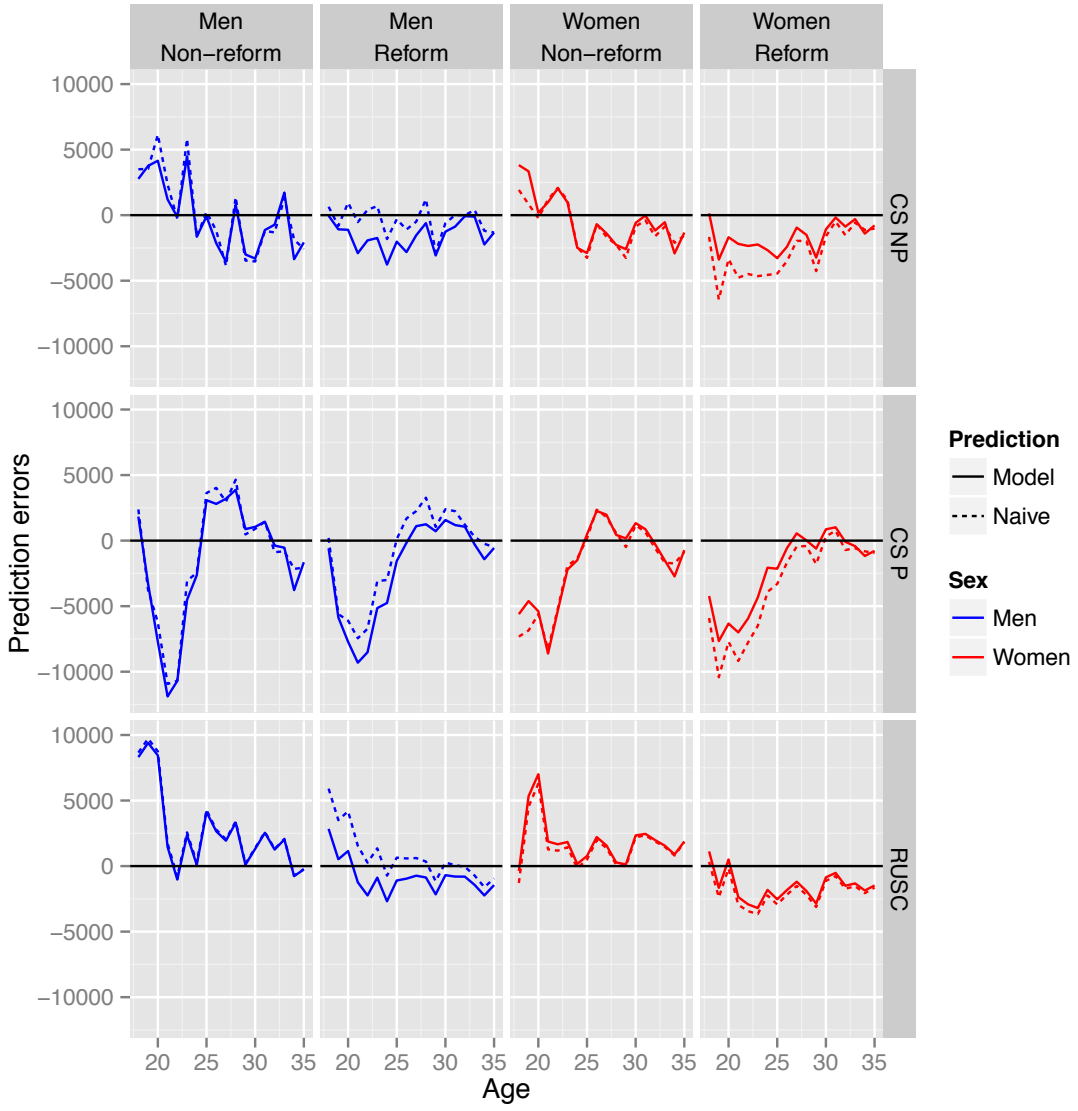


Figure 2: Prediction errors on marriages in 1982

We have left several interesting theoretical issues for future research. One such issue is the behavior of the finite population approximation of the model. We have worked in an idealized model with an infinite number of agents within each observable group; 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. It is easy to see from the proof in Appendix A that Theorem 1 still holds with $G_x$ and $H_y$ replaced by

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

The law of large numbers implies the pointwise convergences $\hat{G}_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) \to G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot})$ and $\hat{H}_y(\boldsymbol{V}_{\cdot\boldsymbol{y}}) \to H_y(\boldsymbol{V}_{\cdot\boldsymbol{y}})$ as the number of individuals gets large. It is natural to expect that the solutions $\hat{\boldsymbol{\mu}}$ and $(\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}})$ of the finitely sampled primal and dual problems converge to their large population analogs[18]. This goes beyond the scope of the present paper and is left as a conjecture. Likewise, we leave the exploration of the rate of convergence for future research.

To conclude, let us emphasize the wide applicability of the methods introduced in the present paper, and the potential for extensions. On the methodological front, one challenge is to extend the logit setting of Choo and Siow to the case where the observable characteristics of the partners may be continuous. This issue is addressed by Dupuy and Galichon (2014) using the theory of extreme value processes; they also propose a test of the number of relevant dimensions for the matching problem. While the framework we used here is bipartite, one-to-one matching, our results open the way to possible extensions to other matching problems. For instance, it may 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). Also, while the present paper operates under the maintained assumption that utility is fully transferable without frictions, this assumption can be relaxed; Galichon, Kominers, and Weber (2014) study models with imperfectly

[18] 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.

transferable utility and separable logit heterogeneity, while Hsieh (2012) looks at models with nontransferable utility and a similar form of heterogeneity.

## 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 & J.F. Thisse (1988): "A Representative Consumer Theory of the Logit Model," *International Economic Review* 29, 461–466.

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

[4] Agarwal, N. (2015): "An Empirical Model of the Medical Match," *American Economic Review*, forthcoming.

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

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

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

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

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

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

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

[12] 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.

[13] Chiappori, P.-A., S. Oreffice, and C. Quintana-Domeque (2012): ".Fatter Attraction: Anthropometric and Socioeconomic Characteristics in the Marriage Market," *Journal of Political Economy* 120, 659–695.

[14] Chiappori, P.-A., B. Salanié (2015): "The Econometrics of Matching Models," *Journal of Economic Literature*, forthcoming.

[15] Chiappori, P.-A., B. Salanié, and Y. Weiss (2015): "Partner Choice and the Marital College Premium" mimeo.

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

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

[18] Costinot, A. and J. Vogel (2014): "Beyond Ricardo: Assignment Models in International Trade", forthcoming in the *Annual Review of Economics*.

[19] Gurobi Optimization, Inc. (2015): "Gurobi Optimizer Reference Manual", http://www.gurobi.com.

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

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

[22] 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.

[23] Del Boca, D., and C. Flinn (2014): "Household Behavior and the Marriage Market," *Journal of Economic Theory* 150, 515–550.

[24] Dupuy, A. and A. Galichon (2014): "Personality traits and the marriage market," *Journal of Political Economy* 122 (6), 1271–1319.

[25] Echenique, F., S. Lee, M. Shum, and B. Yenmez (2013): "The Revealed Preference Theory of Stable and Extremal Stable Matchings," *Econometrica* 81, 153–171.

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

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

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

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

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

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

[32] Galichon, A., S. Kominers, and S. Weber (2014): "An Empirical Framework for Matching with Imperfectly Transferable Utility," mimeo.

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

[34] 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.

[35] Graham, B. (2013): "Errata in 'Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers' ". Mimeo, UC Berkeley.

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

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

[38] Hagedorn, M., T.-H. Law, and I. Manovskii (2014): "Identifying Equilibrium Models of Labor Market Sorting," mimeo University of Pennsylvania.

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

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

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

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

[43] Hsieh, Y-W. (2012) "Understanding Mate Preferences from Two-Sided Matching Markets: Identification, Estimation and Policy Analysis," mimeo, USC.

[44] Jacquemet, N., and J.-M. Robin (2012): "Assortative Matching and Search with Labor Supply and Home Production," mimeo.

[45] 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.

[46] Menzel, K. (2014): "Large Matching Markets as Two-Sided Demand Systems," *Econometrica*, forthcoming.

[47] Mourifié, I., and A. Siow (2014): "Cohabitation vs. Marriage: Marriage matching with peer effects," mimeo, University of Toronto.

[48] 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.

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

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

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

[52] Sinha, S. (2014): "Identification and Estimation in One-to-One Matching Models with Nonparametric Unobservables," mimeo, Northwestern University.

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

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

# Appendix

## A Proofs

### A.1 Proof of Proposition 2

Replace the expression of $G_x$ (2.1) in the formula for $G_x^*$ (2.3) to obtain

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

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 $\bar{\boldsymbol{U}}_{\boldsymbol{x}\cdot}$ under the distribution $\mu_{Y|x=x}$; therefore this can be rewritten as

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

where the minimum is taken over all pairs of functions $(\bar{\boldsymbol{U}}_{\boldsymbol{x}\cdot},\bar{W}_x(\boldsymbol{\varepsilon}_{\boldsymbol{i}}))$ 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 $\boldsymbol{P}_x$ and the surplus is $\varepsilon_{iy}$. By the equivalence of the primal and the dual, this yields Expression (2.7).

### A.2 Proof of Theorem 1

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

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

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

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

Now, $\int\max_{y\in\mathcal{Y}}\{U_{xy}+\varepsilon_y,\varepsilon_0\}\,d\tilde{n}(x,\varepsilon)=\sum_x n_x\boldsymbol{E}_{\boldsymbol{P}_x}[\max_{y\in\mathcal{Y}}U_{xy}+\varepsilon_y,\varepsilon_0]=n_xG_x(\boldsymbol{U}_{\boldsymbol{x}\cdot})$, where $\boldsymbol{E}_{\boldsymbol{P}_x}$ denotes the expectation over the population of men in group $x$, and where we have used Assumption 1. Adding the similar expression for women, we get that $\mathcal{W}$ is the maximum over $\boldsymbol{\mu}\geq0$ of $\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}\Phi_{xy}-A(\boldsymbol{\mu})-B(\boldsymbol{\mu})$, where $A(\boldsymbol{\mu})=\max_{(U_{xy})}\{\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\mu_{xy}U_{xy}-\sum_{x\in\mathcal{X}}n_xG_x(\boldsymbol{U}_{\boldsymbol{x}\cdot})\}$, and $B$ has a similar expression involving $H$ and $m$ instead of $G$ and $n$.

Now consider the term with first subscript $x$ in $A(\boldsymbol{\mu})$. They sum up to $n_x(\sum_{y\in\mathcal{Y}}\mu_{y|x}U_{xy}-G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}))$, which is $n_x$ times the Legendre transform of $G$ evaluated at $\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}$. We can therefore rewrite $A(\boldsymbol{\mu})$ and $B(\boldsymbol{\mu})$ in terms of the Legendre-Fenchel transforms:

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

Expression (2.9) follows. Convexity of $\mathcal{W}$ w.r.t. $\Phi$ is a an immediate consequence of the latter expression, while concavity of $\mathcal{W}$ w.r.t. $(\boldsymbol{n},\boldsymbol{m})$ is immediately deduced from Expression (2.10). Points (ii), (iii) and (iv) are then deduced immediately.

## A.3 Proof of Theorem 2

If Assumption 4 holds for $\boldsymbol{P}_x$, 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. Moreover, $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}})$ 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 (2.9). 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 (2.12). The envelope theorem in the same program gives us (2.11), 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 (2.13), 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 $\tau_{xy}$ is observed, then $\alpha$ and $\gamma$ 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 $\hat{\boldsymbol{\lambda}}$ solves (4.3). Hence, by its FOC $\hat{\boldsymbol{\lambda}}$ satisfies $\sum_{x,y} \hat{\mu}_{xy}\Phi_{xy}^k = \partial\mathcal{W}/\partial\lambda_k(\boldsymbol{\Phi}^{\hat{\boldsymbol{\lambda}}}, \boldsymbol{n}, \boldsymbol{m})$; but by the Envelope Theorem, $\partial\mathcal{W}/\partial\lambda_k(\boldsymbol{\Phi}^{\hat{\boldsymbol{\lambda}}}, \boldsymbol{n}, \boldsymbol{m}) = \sum_{x,y} \mu_{xy}^{\hat{\boldsymbol{\lambda}}}\Phi_{xy}^k$.

(ii) Program (4.3) can be rewritten as

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

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

## A.6 Proof of Proposition 3

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

$\sum_{x,y} \hat{\mu}_{xy} \Phi_{xy}^{\hat{\boldsymbol{\lambda}}} = \sum_{x,y} \mu_{xy}^{\hat{\boldsymbol{\lambda}}} \Phi_{xy}^{\hat{\boldsymbol{\lambda}}}$ holds by construction; hence $\mathcal{E}(\hat{\boldsymbol{\mu}}) \geq \mathcal{E}(\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}})$ with equality if and only if $\boldsymbol{\mu}^{\hat{\boldsymbol{\lambda}}} = \hat{\boldsymbol{\mu}}$.

### A.7 Proof of Theorem 4

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

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

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

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

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

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

By Theorem 8.4 of Bauschke and Borwein, the iterated projection algorithm converges[19] to the projection $\boldsymbol{\mu}$ of $\boldsymbol{\mu}^{(0)}$ on $\mathcal{M}(\boldsymbol{n}, \boldsymbol{m})$, which is also the maximizer $\boldsymbol{\mu}$ of (2.9).

## 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

[19] 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 $\phi$.

functions defined by $F(w_0, ..., w_J) = \exp(-g_x(e^{-w_0}, ..., e^{-w_J}))$ are multivariate cumulative distribution functions, associated to a distribution which we denote $\boldsymbol{P}_x$. Then introducing utility shocks $\varepsilon_x \sim \boldsymbol{P}_x$, we have by a theorem of McFadden (1978):

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

where $\gamma$ 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 $\mathcal{E}(\boldsymbol{\mu})$ arising from the heterogeneity on the men side is

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

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

## B.1 Ex. 1: The Choo-Siow model

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

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

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

**Claims of Section 4.2**. We can rewrite $L$ as

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

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

## B.2 Ex. 2: The RUSC model

**Claims of Section 3.1**. From Proposition 2, $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = -\max_{\pi\in\mathcal{M}_x} \boldsymbol{E}_\pi\left[\zeta_x(Y)\varepsilon\right]$, where $\pi$ has margins $F_\varepsilon$ and $\mu(Y|x=x)$. Since the function $(\varepsilon,\zeta) \longrightarrow \varepsilon\zeta$ is supermodular, the optimal matching must be positively assortative: larger $\varepsilon$'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 $\varepsilon$ 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 $F_\varepsilon\left(\varepsilon_{(k)}\right) = \sum_{\zeta_x(y)\le\zeta_{(k)}} \mu_{y|x}$; and as a result, $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = -\sum_{1\le k\le|\mathcal{Y}|+1} \zeta_{(k)} e_k$, 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 $\varepsilon$ in interval $\left[\varepsilon_{(k-1)}, \varepsilon_{(k)}\right]$; then $-n_x G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = n_x \sum_{1\le k\le|\mathcal{Y}|+1} \zeta_{(k)} \sum_{\zeta_x(y)=\zeta_{(k)}} \mu_{y|x}\bar{e}_k = \sum_y \mu_{xy}\bar{e}_{K(y)}$, 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 $\varepsilon$ is distributed uniformly over $[0,1]$, (B.3) becomes $\varepsilon_{(k)} = \sum_{\zeta_x(y)\le\zeta_{(k)}} \mu_{y|x}$, and $\boldsymbol{E}\left[\varepsilon\mathbf{1}(\varepsilon \in [\varepsilon_{(k-1)}, \varepsilon_{(k)}])\right] = \left(\varepsilon_{(k)} - \varepsilon_{(k-1)}\right)\left(\varepsilon_{(k)} + \varepsilon_{(k-1)}\right)/2$, we obtain

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

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

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

Now define a matrix $T^x$ and a vector $\sigma^x$ 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 $G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = \frac{1}{2}\left(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}' T^x \boldsymbol{\mu}_{\cdot|\boldsymbol{x}} + 2\sigma^x.\boldsymbol{\mu}_{\cdot|\boldsymbol{x}} - S_{00}^x\right)$.

Introducing

$$A_{xy,x'y'} = \frac{1}{n_x} 1\left\{x = x'\right\} T_{yy'}^x + \frac{1}{m_y} 1\left\{y = y'\right\} T_{xx'}^y \tag{B.4}$$

$$B_{xy} = \sigma_y^x + \sigma_x^y \text{ and } c = -\sum_{x\in\mathcal{X}} n_x S_{00}^x - \sum_{y\in\mathcal{Y}} m_y S_{00}^y \tag{B.5}$$

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

**Claims of Section 3.2**. Note that $\boldsymbol{\mu}$ is determined by

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

where $\boldsymbol{\Phi}.\boldsymbol{\mu}$ is the vector product $\sum_{xy}\mu_{xy}\Phi_{xy}$. Hence, if $\mu$ is interior, i.e. if there are no empty cells, the solution is given by $\boldsymbol{\mu} = A^{-1}(\boldsymbol{\Phi} - B)$ and $\mathcal{W} = \frac{1}{2}((\boldsymbol{\Phi} - B)' A^{-1}(\boldsymbol{\Phi} - B) - c)$, 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

$$M_{x'y,x''y'}^x = 1\left\{x = x' = x''\right\} T_{yy'}^x \tag{B.6}$$

hence, one gets $u_x = \partial\mathcal{W}/\partial n_x = (\boldsymbol{\Phi} - B)' A^{-1} M^x A^{-1}(\boldsymbol{\Phi} - B)/(2n_x^2) + S_{00}^x/2$, 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, (C.4) yields

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

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 (C.5) yields

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

**Claims of Section 4.2**. Assume that the data generating process is the RUSC model of Example 2, where we fix $\boldsymbol{\zeta}_x$and $\boldsymbol{\xi}_y$, and where $\boldsymbol{\Phi}^{\boldsymbol{\lambda}}$ is linearly parameterized as in (4.2). Assume further that all $\mu$'s are positive. Then

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

where $\boldsymbol{\phi} = \left(\phi^k_{xy}\right)_{xy,k}$ is reshaped as a matrix. As a consequence, the Moment Matching estimator is a simple affine function of the observed comoments:

$$\hat{\boldsymbol{\lambda}}^{MM} = \left(\boldsymbol{\phi}' \boldsymbol{A}^{-1} \boldsymbol{\phi}\right)^{-1} \left(C(\hat{\boldsymbol{\mu}}) + \boldsymbol{\phi}' \boldsymbol{A}^{-1} \boldsymbol{B}\right).$$

## B.3 Ex. 3: The nested logit model

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

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

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

*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 $s$*

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

$$G(\boldsymbol{U}.) = \log(1 + \sum_s (\sum_e e^{U_{se}/\sigma_s})^{\sigma_s}), \tag{B.12}$$

hence, by (2.2), it follows that $\mu_{se}/\mu_0 = (\sum_e e^{U_{se}/\sigma_s})^{\sigma_s - 1} e^{U_{se}/\sigma_s}$, where $\mu_0$ is again defined in (3.1). Thus $\log(\mu_s/\mu_0) = \sigma_s \log(\sum_e \exp(U_{se}/\sigma_s))$, and therefore $U_{se} = \log(\mu_s/\mu_0) + \sigma_s \log(\mu_{se}/\mu_s)$. Now, by (2.5),

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

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

If the heterogeneity structure is the same for all men and all women (with possibly different dispersion parameters $\sigma$ for men and $\tau$ for women), then the expressions of $\mathcal{E}(\boldsymbol{\mu})$ and $\mathcal{W}(\boldsymbol{\mu})$ can easily be obtained. Now if the nested logit applies for men of group $x$ with parameters $(\sigma^x_{s'})$ and for women of group $y$ with parameters $(\tau^y_s)$, 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 the surplus from a match between a man of group $x = (s, e)$ and a woman of group $y = (s', e')$

$$\Phi_{xy} = \log \frac{\mu_{xy}^{\sigma^x_{s'} + \tau^y_s} \mu_{x,s'}^{1-\sigma^x_{s'}} \mu_{s,y}^{1-\tau^y_s}}{\mu_{x0}\mu_{0y}}. \tag{B.13}$$

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

**Maximum Likelihood Estimation**. In the Nested Logit model of Example 3, 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 $\mathbf{\Phi}^\theta$ by a parameter $\theta$ of dimension less than or equal to $N_s^2 \times N_e^2 - 2N_s^2 \times N_e$. Letting $\boldsymbol{\lambda} = \left(\sigma_{s_y}^{s_x e_x}, \sigma_{s_x}^{s_y, e_y}, \theta\right)$, $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ is the solution in $M$ to the system of equations

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

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

**Computing the Equilibrium**. Consider the Nested Logit model of Example 3, and assume for simplicity that there is only one social group, so the model boils down to a heteroskedastic logit model with scale parameters $\sigma^x$ and $\tau^y$. Recall the equilibrium formula which comes from (3.5)

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

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

$$n_x = \mu_{x0} + \sum_{y \in \mathcal{Y}} \mu_{x0}^{\frac{\sigma_x}{\sigma_x + \tau_y}} \mu_{0y}^{\frac{\tau_y}{\sigma_x + \tau_y}} \exp \frac{\Phi_{xy}}{\sigma_x + \tau_y} \tag{B.14}$$

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

$$m_y = \mu_{0y} + \sum_{x \in \S} \mu_{0y}^{\frac{\tau_y}{\sigma_x + \tau_y}} \mu_{x0}^{\frac{\sigma_x}{\sigma_x + \tau_y}} \exp \frac{\Phi_{xy}}{\sigma_x + \tau_y}. \tag{B.15}$$

Note that steps (B.14) and (B.15) 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.

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

$$G_x(\boldsymbol{U}_{\boldsymbol{x}\cdot}) = \sum_{k=1}^K \beta_k^x \sigma_k^x \log\left(1 + \sum_{y\in\mathcal{Y}} e^{U_{xy}/\sigma_k^x}\right). \tag{B.16}$$

*Still from the results of Example 1, $G_{xk}^*(\boldsymbol{\mu}) = \sigma_k^x \sum_{y\in\mathcal{Y}_0} \mu_y \log \mu_y$. 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 $\boldsymbol{U}_{x.} \to \beta_k^x G_{xk}(\boldsymbol{U}_{x.})$ is $f^*(\boldsymbol{\mu}_\cdot^k) = \beta_k^x G_{xk}^*\left(\frac{\boldsymbol{\mu}^k}{\beta_k^x}\right)$; thus it follows that*

$$G_x^*(\boldsymbol{\mu}_{\cdot|\boldsymbol{x}}) = \min_{\sum_{k=1}^K \mu_y^k = \mu_{y|x}} \sum_{k=1}^K \sigma_k^x \left(\mu_0^k \log \frac{\mu_0^k}{\beta_k^x} + \sum_{y\in\mathcal{Y}} \mu_y^k \log \frac{\mu_y^k}{\beta_k^x}\right). \tag{B.17}$$

*Then $U_{xy}$ is given by $U_{xy} = \sigma_k^x \log(\mu_y^k/\mu_0^k)$, where $(\mu_y^k)$ is the minimizer of (B.17).*

# C Comparative statics

The results of Theorem 1 can be used 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 $\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{n}, \boldsymbol{m})$ is given by the dual expressions

$$\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{n}, \boldsymbol{m}) = \max_{\mu\in\mathcal{M}(\boldsymbol{n},\boldsymbol{m})} \left(\sum_{xy} \mu_{xy}\Phi_{xy} - \mathcal{E}(\mu)\right), \text{ and} \tag{C.1}$$

$$\mathcal{W}(\boldsymbol{\Phi}, \boldsymbol{n}, \boldsymbol{m}) = \min_{U_{xy}+V_{xy}=\Phi_{xy}} \left(\sum n_x G_x(U_{xy}) + \sum m_y H_y(V_{xy})\right) \tag{C.2}$$

As a result, note that by (C.1), $\mathcal{W}$ is a convex function of $\boldsymbol{\Phi}$, and by (C.2) it is a concave function of $(\boldsymbol{n}, \boldsymbol{m})$. By the envelope theorem in (C.1) and in (C.2), we get respectively

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

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

$$\frac{\partial u_x}{\partial n_{x'}} = \frac{\partial^2 \mathcal{W}}{\partial n_x \partial n_{x'}} = \frac{\partial u_{x'}}{\partial n_x} \tag{C.3}$$

(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 (C.3) 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 $(\boldsymbol{n}, \boldsymbol{m})$ 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 $\Phi_{xy}$ yields

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

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 $\Phi_{xy}$ and $\Phi_{x'y'}$ yields

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

The interpretation is the following: if increasing the matching surplus between groups $x$ and $y$ has a positive effect on marriages between groups $x'$ and $y'$, then increasing the matching surplus between groups $x'$ and $y'$ has a positive 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.

# D Geometric interpretation

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

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

This is a convex polyhedron, which we call the *covariogram*; and if the model is well-specified the covariogram must contain the observed matching $\hat{\boldsymbol{\mu}}$. For any value of the parameter vector $\boldsymbol{\lambda}$, the optimal matching $\boldsymbol{\mu}^{\boldsymbol{\lambda}}$ generates a vector of comoments $C^{\boldsymbol{\lambda}}$ that belongs to the covariogram; and it also has an entropy $\mathcal{E}^{\boldsymbol{\lambda}} \equiv \mathcal{E}(\boldsymbol{\mu}^{\boldsymbol{\lambda}})$. We already know that this model is just-identified from the comoments: the mapping $\boldsymbol{\lambda} \longrightarrow C^{\boldsymbol{\lambda}}$ is invertible on the covariogram. Denote $\boldsymbol{\lambda}(C)$ its inverse. The corresponding optimal matching has entropy $\mathcal{E}_r(C) = \mathcal{E}^{\boldsymbol{\lambda}(C)}$. The level sets of $\mathcal{E}_r(.)$ are the isoentropy curves in the covariogram; they are represented on Figure 3. The figure assumes $K = 2$ dimensions; then $\boldsymbol{\lambda}$ can be represented in polar coordinates as $\boldsymbol{\lambda} = r \exp(it)$. 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 $\partial F$ of the covariogram corresponds to $r = \infty$. Then there is no unobserved heterogeneity and generically over $t$, the comoments generated by $\boldsymbol{\lambda}$ must belong to a finite set of vertices, so that $\boldsymbol{\lambda}$ is only set-identified. As $r$ decreases for a given $t$,

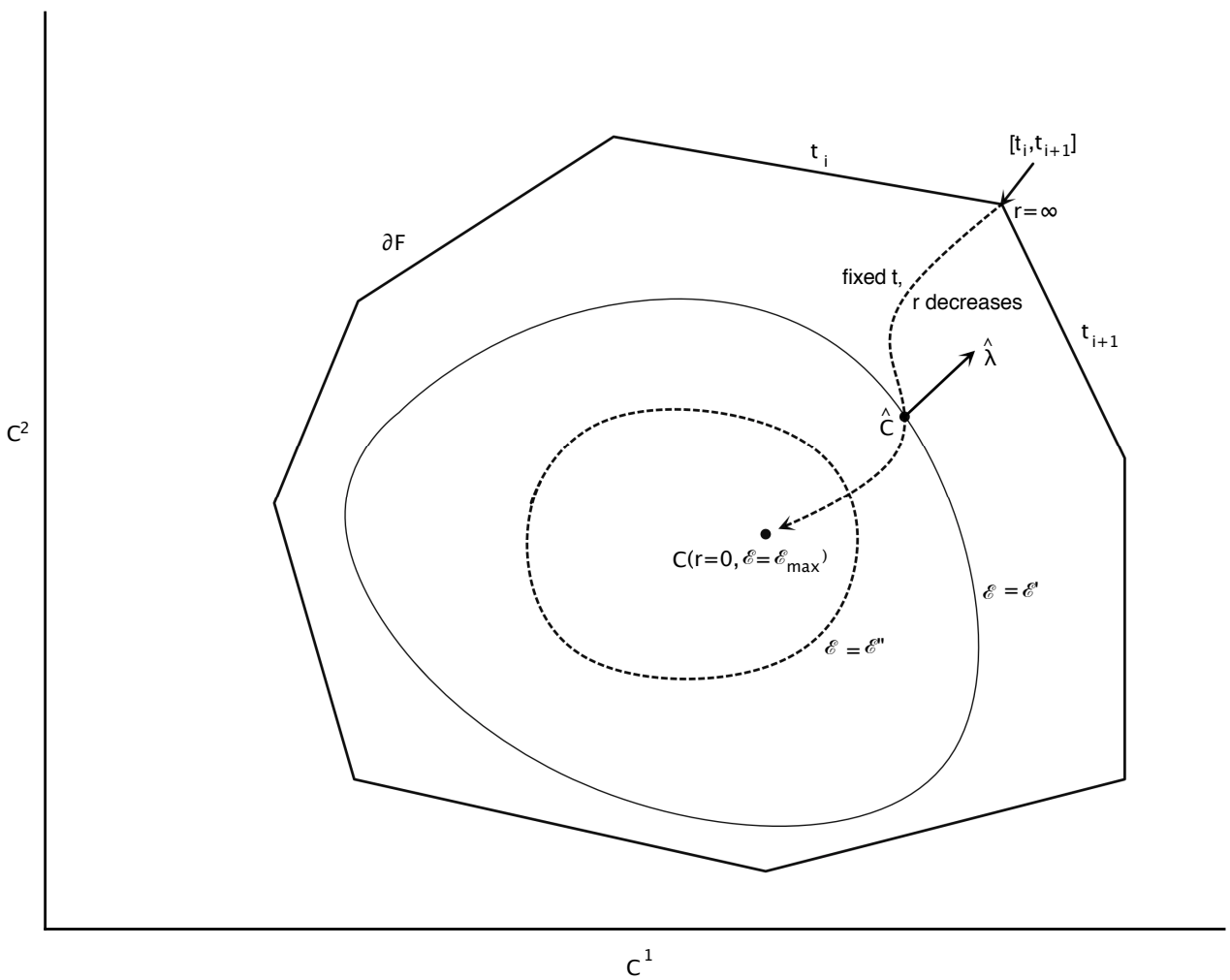


Figure 3: The covariogram and related objects

the corresponding comoments follow a trajectory indicated by the dashed line on Figure 3, from the boundary $\partial F$ to the point $C_0$. At the same time, the entropy $\mathcal{E}^{\boldsymbol{\lambda}}$ increases, and the trajectory crosses contours of higher entropy ($\mathcal{E}^{'}$ then $\mathcal{E}^{''}$ on the figure.) The CML Estimator $\hat{\boldsymbol{\lambda}}$ could also be obtained by taking the normal to the isoentropy contour that goes through the observed comoments $\hat{C}^k = C^k(\hat{\boldsymbol{\mu}})$, as shown on Figure 3. Indeed, the estimator $\hat{\boldsymbol{\lambda}}^{MM}$ of the parameter vector is given by the gradient of $\mathcal{E}_r(.)$ at the point $\hat{C}$, that is $\partial\mathcal{E}_r(\hat{C})/\partial C^k = \hat{\lambda}_k^{MM}$.

# E Computational experiments

Equation (5.9) 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 $\Phi_{xy}$ 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

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

and

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

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 (E.1)–(E.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[20] 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.

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

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 4. 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[21].

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.

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

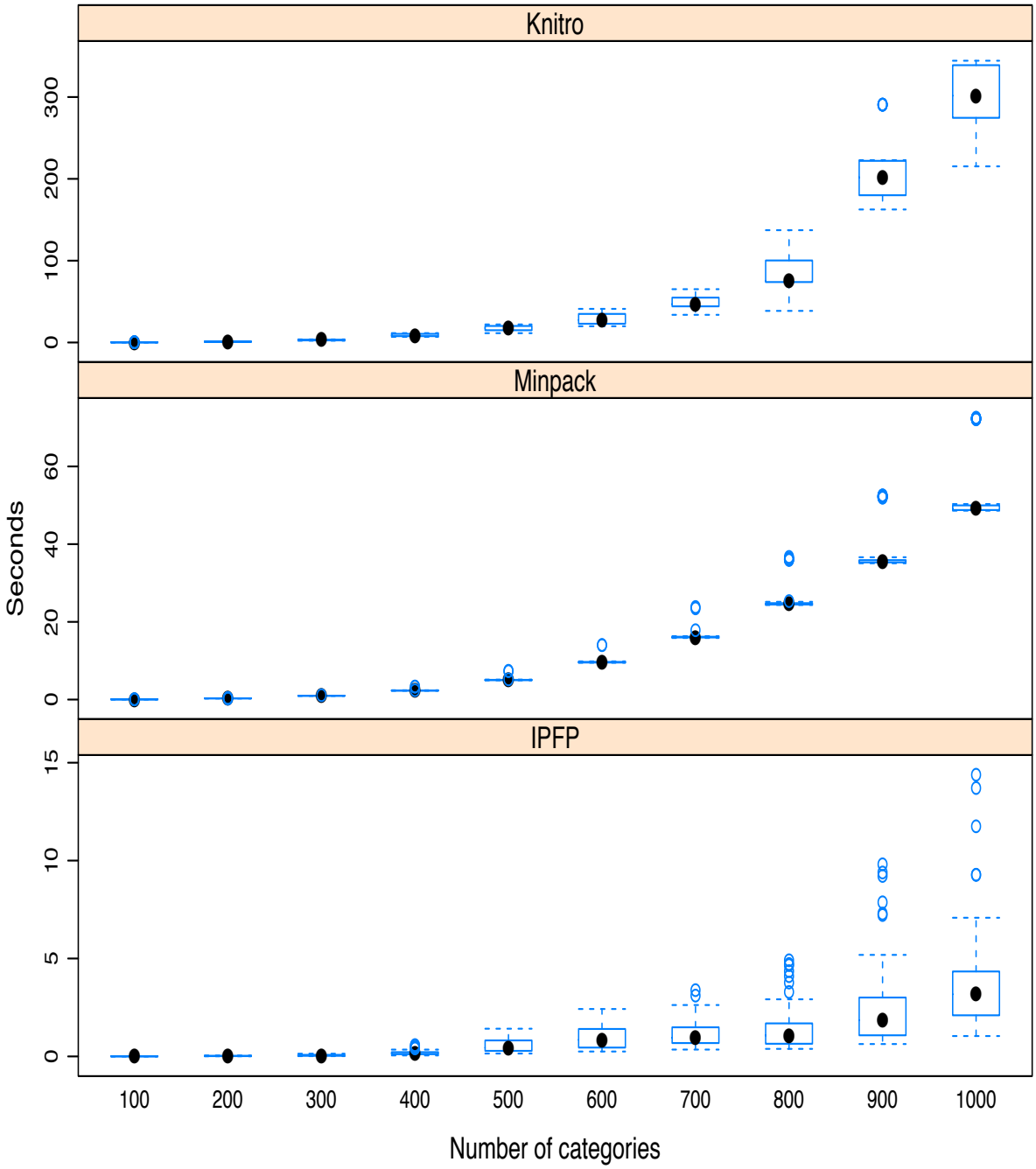


Figure 4: Solving for the optimal matching