Estimating Separable Matching Models*

Alfred Galichon† Bernard Salanié‡

May 22, 2023

Abstract

Most recent empirical applications of matching with transferable utility have imposed a natural restriction: that the joint surplus be *separable* in the sources of unobserved heterogeneity. We propose here two simple methods to estimate models in this class. The first method is a minimum distance estimator that relies on the generalized entropy of matching introduced in Galichon and Salanié (2022). The second applies to the more special but popular Choo and Siow (2006) model, which it reformulates as a generalized linear model with two-way fixed effects. Both methods are easy to apply and perform very well.

Keywords: matching, marriage, assignment, structural estimation.

JEL codes: C78, C13, C15.

Introduction

The estimation of models of two-sided matching has made considerable progress in the past decade. While some of this work has used matching under non-transferable utility, many applications have focused on markets where utility is transferable. The pioneering contribution of Choo and Siow (2006) introduced a simple and highly tractable specification for matching models with perfectly transferable utility. Their specification is a natural extension of the multinomial logit model, and it has become quite popular. They applied it to estimate the effect of the 1973 liberalization of

*The authors are grateful to Clément Montes for superb research assistance, and to Antoine Jacquet for his detailed comments.

†New York University and Sciences Po. Support from ERC grant EQUIPRICE No. 866274 is acknowledged.

‡Columbia University.

abortion in the US on marriage outcomes. In doing so, they used a nonparametric estimator of the matching patterns.

The Choo and Siow specification rests on three main assumptions: separability; a large market approximation; and unobserved shocks to the joint surplus that are distributed as iid standard type I extreme value random variables. In Galichon and Salanié (2022), we showed that the third, distributional assumption is not necessary for most of the analysis: for any (separable) distribution of the errors, the joint surplus is nonparametrically identified. The crucial assumption that underlies much of the literature in this subfield is *separability*. As we will explain in Section 1.1, this rules out interactions between partner-specific unobserved shocks in the generation of the joint surplus. The great value of separability is that it generates a simple relationship between the observed matching patterns and the underlying joint surplus that can readily be used to identify and estimate the surplus.

The application to marriage in Choo and Siow only conditioned on the ages of the partners in a couple; and it specified errors as type I extreme value. These two restrictions, combined with separability, allowed them to construct a nonparametric estimator of the joint surplus. This strategy breaks down, however, when more covariates are considered as matching cells become too small; and by construction, it does not allow for parameterized error distributions. Structural models of household behavior also naturally introduce parameters in the joint surplus. In all of these cases, the analyst must resort to parametric models. We describe here two very simple methods to estimate parametric versions of separable matching models with perfectly transferable utility, with special emphasis on the Choo and Siow model and more generally on "semilinear" models, where the joint surplus is linear in the parameters.

Our first method applies a minimum-distance estimator to the identification equation derived in Galichon and Salanié (2022), which relates the joint surplus to the derivatives of a generalized entropy function evaluated at the observed matching patterns. For any fixed distribution of the error terms, the generalized entropy can be evaluated and differentiated, numerically if needed. The estimator selects parameter values and also provides a simple specification test. In semilinear models, the estimator can be obtained in closed form by quasi-generalized least squares.

The second method we present applies more specifically to the semilinear Choo and Siow model. We show that the moment-matching estimator we described in Galichon and Salanié (2022) can be reframed as a generalized linear model, more specifically as a weighted pseudo-maximum likelihood estimator of a Poisson regression with two-sided fixed effects. This is available as `linear_model` in

the `scikit-learn` library in `Python`, as `fepois` in the `R` package `fixest` and as `ppmlhdfe` in Stata, among other common statistical packages.

Section 1 describes the simplest version of the type of matching market to which our methods apply: the bipartite model with perfectly transferable utility. To make the paper self-contained, we give intuitive proofs of the main results; we refer the reader to Galichon and Salanié, 2022 for fully rigorous arguments. Sections 2 and 3 present our two estimation methods, still only for the bipartite model. Section 4 explains how we deal with some common issues, and Section 5 shows how our two methods extend to more general matching markets. Finally, Section 6 shows the results of a small Monte Carlo simulation. We conclude with a brief discussion of the pros and cons of the two methods.

We coded both methods in a Python package called `cupid_matching` that is available on the standard repositories[1].

1 The Bipartite Model

In a bipartite matching market, each match consists of at most one partner in each of two separate sub-populations. For simplicity, we refer here to the two sub-populations as "men" and "women"; each match consists of one man and one woman, and we allow for singles. We use the same notation as in Galichon and Salanié (2022), from which most of the results in this section are taken. We assume that the analyst can only observe which of a finite set of *types* each individual belongs to. Types could represent a combination of education and age, for instance. Each man $i \in \mathcal{I}$ has a type $x_i \in \mathcal{X}$; and, similarly, each woman $j \in \mathcal{J}$ has a type $y_j \in \mathcal{Y}$. Without loss of generality, we assimilate $\mathcal{X}$ to $\{1, \dots, X\}$ and $\mathcal{Y}$ to $\{1, \dots, Y\}$. We will say that "man i is of type x" if $x_i = x$, and "woman j is of type y" if $y_j = y$; and we will use the shorter notation "$i \in x$" and "$j \in y$". In addition, men and women of a given type differ along some dimensions that they all observe, while the analyst does not.

We denote μ_{ij} the indicator function for a match between man i and woman j: it equals 1 if i and j are matched and 0 otherwise. Similarly, μ_{i0} and μ_{0j} equal 1 if i or j remain unmatched, respectively. A *matching* is the specification of who matches with whom; it is characterized by

[1] See http://bsalanie.github.io/code for more information, and https://bsalanie-cupid-matching-st-main-page-pwvpse.streamlitapp.com for an interactive Streamlit app that demonstrates solving and estimating a Choo and Siow (2006) model.

the collection of numbers $(\mu_{ij}, \mu_{i0}, \mu_{0j})$. It is *feasible* if each individual is matched to at most one partner. Since every man i is either matched once or stays single, $\sum_j \mu_{ij} + \mu_{i0} = 1$; similarly, for every woman j we have $\sum_i \mu_{ij} + \mu_{0j} = 1$. A feasible matching is *stable* if no individual who has a partner would prefer to be single, and if no two individuals would prefer forming a couple over their current situation.

Since the analyst only observes types x_i and y_j, she can only count the number of matches between partners of given types. We denote μ_{xy} the number of couples where the man belongs to type x, and where the woman belongs to type y, which is formally defined as $\mu_{xy} = \sum_{i \in x, j \in y} \mu_{ij}$. We also denote $\mu_{x0} = \sum_{i \in x} \mu_{i0}$ and $\mu_{0y} = \sum_{j \in y} \mu_{0j}$ the numbers of single men of type x and of single women of type y. An *observed matching* $\boldsymbol{\mu}$ thus consists of the collection of numbers $(\mu_{xy}, \mu_{x0}, \mu_{0y})$. We denote $\mathcal{A} = (\mathcal{X} \times \mathcal{Y}) \cup (\mathcal{X} \times \{0\}) \cup (\{0\} \times \mathcal{Y})$ the set of potential observed matchings. As the marital options include singlehood, we denote $\mathcal{X}_0 = \mathcal{X} \cup \{0\}$ and $\mathcal{Y}_0 = \mathcal{Y} \cup \{0\}$ the respective sets of marital options of women and men in an observed matching.

If n_x represents the number of men of type $x \in \mathcal{X}$, feasibility requires that $\sum_{y \in \mathcal{Y}} \mu_{xy} + \mu_{x0} = n_x$. Similarly, $\sum_{x \in \mathcal{X}} \mu_{xy} + \mu_{0y}$ must equal m_y the number of women of type y. We denote $\boldsymbol{q} = (\boldsymbol{n}, \boldsymbol{m})$ the vector that collects the numbers n_x and m_y, which we will sometimes refer to as the *margins*. Finally, *feasible marriage patterns* are represented by an (X, Y) matrix $\boldsymbol{\mu}$ of non-negative numbers such that $\sum_{y \in \mathcal{Y}} \mu_{xy} \leq n_x$ for all $x \in \mathcal{X}$ and $\sum_{x \in \mathcal{X}} \mu_{xy} \leq m_y$ for all $y \in \mathcal{Y}$.

1.1 Separability

This paper assumes that utility is perfectly transferable within each couple, at a constant 1-for-1 rate. Then the utility possibility frontier within an (i, j) match has equation $u_i + v_j = \tilde{\Phi}_{ij}$. In this notation, $\tilde{\Phi}_{ij}$ represents the *joint utility* in the couple. A priori, this joint utility might depend on interactions between both the observed types x_i and y_j and the unobserved components of the identity of the partners. This would make the analysis quite difficult, both analytically and in terms of the amount of data required to identify the parameters[2]. For this reason, most of the literature has adopted the simplifying assumption that the joint utility from a match is *separable*:

Assumption 1 (Separability)**.** *There exist a vector* $\boldsymbol{\Phi}$ *in* $\mathbb{R}^{X \times Y}$ *and random terms* $\boldsymbol{\varepsilon}$ *and* $\boldsymbol{\eta}$ *such that*

[2] See Fox et al., 2018 for a contribution in this direction.

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

$$\tilde{\Phi}_{ij} = \Phi_{xy} + \varepsilon_{iy} + \eta_{xj}, \tag{1.1}$$

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

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

(iv) conditional on $i \in x$, the random vectors $\boldsymbol{\varepsilon}_i = (\varepsilon_{iy})_{y \in \mathcal{Y}_0}$ are independent across i; they have probability distribution $\mathbb{P}_x$,

(v) conditional on $j \in y$, the random vectors $\boldsymbol{\eta}_j = (\eta_{xj})_{x \in \mathcal{X}_0}$ are independent across j; they have probability distribution $\mathbb{Q}_y$,

(vi) the random variables

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

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

The core of the separability assumption is item (i), which rules out interactions between partner characteristics that are unobserved by the analyst. This would be violated if, for instance, man i had an idiosyncratic preference for an unobserved characteristic of woman j. Clearly, separability is a stronger assumption when the data contains few attributes of men and women. Exploratory simulations by Chiappori, Nguyen, et al., 2019 suggest that, even when it only holds approximately, assuming it may only generate small biases.

Chiappori et al. (2017) and Galichon and Salanié (2022) showed that under separability, the characterization of stable matchings becomes much simpler. Consider a woman $j \in y$. She could either stay single and obtain utility η_{0j}, or match with a man i and share the joint utility $\tilde{\Phi}_{ij}$ with him. Let u_i denote the utility that man i expects at a stable matching, so that woman j can only obtain $\tilde{\Phi}_{ij} - u_i$ for herself in their match. To select her best option, this woman will solve

$$\max\left(\eta_{0j}, \max_i \left(\tilde{\Phi}_{ij} - u_i\right)\right);$$

the value of this program represents the utility v_j she can expect at a stable matching. Using Assumption 1.(i), we rewrite

$$\begin{aligned} v_j &= \max\left(\eta_{0j}, \max_{x \in \mathcal{X}} \max_{i \in x} \left(\Phi_{xy} + \varepsilon_{iy} + \eta_{xj} - u_i\right)\right) \\ &= \max\left(\eta_{0j}, \max_{x \in \mathcal{X}} \left(\Phi_{xy} + \eta_{xj} + \max_{i \in x} \left(\varepsilon_{iy} - u_i\right)\right)\right) \end{aligned}$$

Now define $V_{xy} = \min_{i \in x}(u_i - \varepsilon_{iy})$. We obtain

$$\begin{aligned} v_j &= \max\left(\eta_{0j}, \max_{x \in \mathcal{X}} (\Phi_{xy} - V_{xy} + \eta_{xj})\right) \\ &= \max_{x \in \mathcal{X}_0} (\Phi_{xy} - V_{xy} + \eta_{xj}), \end{aligned} \tag{1.2}$$

with the convention that $\Phi_{oy} = V_{0y} = 0$. We would obtain in the same way, for a man $i \in x$,

$$u_i = \max_{y \in \mathcal{Y}_0} (\Phi_{xy} - U_{xy} + \varepsilon_{iy}) \tag{1.3}$$

with $U_{xy} = \min_{j \in y}(v_j - \eta_{xj})$ and $\Phi_{x0} = U_{x0} = 0$.

Now consider $i \in x$ and $j \in y$. If they are a couple, then the sum $u_i + v_j$ of their utilities must equal $\tilde{\Phi}_{ij}$. If they are not partners, then $\tilde{\Phi}_{ij}$ cannot exceed the sum $u_i + v_j$; otherwise they would both be better off by matching together, and splitting the additional joint utility $\tilde{\Phi}_{ij} - u_i - v_j$. This deviation would make the matching unstable. Therefore for all i and j we must have $u_i + v_j \geq \tilde{\Phi}_{ij}$. This translates into

$$(\Phi_{xy} - U_{xy} + \varepsilon_{iy}) + (\Phi_{xy} - V_{xy} + \eta_{xj}) \geq \Phi_{xy} + \varepsilon_{iy} + \eta_{xj}$$

so that $U_{xy} + V_{xy} \leq \Phi_{xy}$. If i and j do match at the stable matching, the utility possibility frontier $u_i + v_j = \tilde{\Phi}_{ij}$ gives $U_{xy} + V_{xy} = \Phi_{xy}$.

To summarize:

Theorem 1 (Discrete Choice Representation)**.** *Normalize utilities by $\Phi_{x0} = U_{x0} = V_{0y} = \Phi_{0y} = 0$ for all $x \in \mathcal{X}$, and $y \in \mathcal{Y}$.*

(i) There exist two (X, Y) matrices of numbers $\boldsymbol{U} = U_{xy}$ and $\boldsymbol{V} \leq \boldsymbol{\Phi} - \boldsymbol{U}$ such that a woman $j \in y$ (resp. a man $i \in x$) remains single or marries a man of a type x (resp. a woman of type y), depending on the maximizer in (1.2) (resp. in (1.3)).

(ii) Moreover, $V_{xy} = \Phi_{xy} - U_{xy}$ for any pair of types such that $\mu_{xy} > 0$.

It is clear from (1.2) and (1.3) that separability implies some form of indifference: if woman $j \in y$ opts for a match with a man of type x, then the other characteristics of this man do not matter to her. It does allow for a restricted form of "matching on unobservables": at a stable matching, this woman is more likely to marry a man i of type x if ε_{iy} is high.

1.2 Generalized Entropy

Now consider the group of all men of type x. By Assumption 1, the value of their average utility at a stable matching equals

$$\tilde{G}_x(\boldsymbol{\Phi}_{x\cdot} - \boldsymbol{V}_{x\cdot}) \equiv \frac{1}{n_x} \sum_{i \in x} \max_{y \in \mathcal{Y}_0} (\Phi_{xy} - V_{xy} + \varepsilon_{iy}).$$

The function $\tilde{G}_x(\boldsymbol{U}) \equiv (1/n_x) \sum_{i \in x} \max_{y \in \mathcal{Y}_0} (U_y + \varepsilon_{iy})$ is a linear function of the maximum of $(Y+1)$ linear functions of its arguments. As such, it is a piecewise linear, convex function of these $(Y+1)$ arguments. Moreover, $\tilde{G}_x(a+t) = \tilde{G}_x(a) + t$ for any vector a and real number t.

Since $\tilde{G}_x$ is convex, it has a subgradient $\partial \tilde{G}_x$, which is defined as follows:

$$\boldsymbol{d} \in \partial \tilde{G}_x(\boldsymbol{U}) \text{ iff for all } \boldsymbol{U}', \ \tilde{G}_x(\boldsymbol{U}') - \tilde{G}_x(\boldsymbol{U}) \geq \boldsymbol{d} \cdot (\boldsymbol{U}' - \boldsymbol{U}').$$

The subgradient is a convex subset of $\mathbb{R}^{Y+1}$; it is almost everywhere a singleton, which is the gradient of $\tilde{G}_x$. Then the partial derivative of $\tilde{G}_x$ with respect to U_y is the proportion of men $i \in x$ who prefer a partner of type y. The points of non-differentiability of $\tilde{G}_x$ correspond to the values of $\boldsymbol{U}$ for which $U_{xy} + \varepsilon_{iy} = U_{xt} + \varepsilon_{it}$ for some $i \in \mathcal{X}$ and some $t \in \mathcal{Y}$.

Recall that the Legendre-Fenchel transform of a real-valued function f defined over a domain D of a finite-dimensional vector space V is the function f^* defined over V by

$$f^*(p) = \max_{x \in D} (p \cdot x - f(x)).$$

Any such function is convex, as the maximum of linear functions of p. It may not take a finite value, however. If $f(x+t) \equiv f(x) + t$ for all x and any real number t, then adding a number t to all elements of x changes the objective by $(\sum_k p_k - 1)t$. It follows that $f^*(p) = +\infty$ if $\sum_k p_k \neq 1$. Since $\tilde{G}_x$ has this property, we circumvent this problem by restricting the domain of $\tilde{G}_x$ to vectors $\boldsymbol{U}$ normalized by $U_0 = 0$. For a vector $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_Y)$ such that $\sum_{y \in \mathcal{Y}} \nu_y \leq 1$, we define

$$\tilde{G}_x^*(\boldsymbol{\nu}) = \max_{\boldsymbol{U} \in \{0\} \times \mathbb{R}^Y} \left(\sum_{y \in \mathcal{Y}} \nu_y U_y - \tilde{G}_x(\boldsymbol{U}) \right). \tag{1.4}$$

Note that we impose $U_0 = 0$ when taking the maximum.

The first-order condition in (1.4) gives us $\boldsymbol{\nu} \in \partial \tilde{G}_x(\boldsymbol{U})$ for all $\boldsymbol{U}$ that achieve the maximum. By the envelope theorem, $\partial \tilde{G}_x^*(\boldsymbol{\nu})$ coincides with the set of these vectors $\boldsymbol{U}$. As a consequence, if $U_0 = 0$ then

$$\boldsymbol{\nu} \in \partial \tilde{G}_x(\boldsymbol{U}) \text{ if and only if } \boldsymbol{U} \in \partial \tilde{G}_x^*(\boldsymbol{\nu}).$$

This reciprocal relationship is at the core of the identification and inference results in Galichon and Salanié (2022). Its economic interpretation is simple. To use the language of discrete choice models, suppose that $\boldsymbol{U}_{x\cdot} = (U_{xy})_{y\in\mathcal{Y}}$ represents the mean surplus that men of type x obtain by matching with women of type y. Then the (generically unique) vector $\boldsymbol{\nu}_{\cdot|x}$ in $\partial\tilde{G}_x(\boldsymbol{U}_{x\cdot})$ represents the proportions of men of this type who marry with the Y types of women (and $\nu_{0|x} = 1 - \sum_{y\in\mathcal{Y}} \nu_{y|x}$ represents the proportion of single men of type x). Reciprocally, the generically unique vector $\boldsymbol{U}_{x\cdot} \in \partial\tilde{G}^*_x(\boldsymbol{\nu}_{\cdot|x})$ *rationalizes* the matching patterns $\boldsymbol{\nu}_{\cdot|x}$ of men of type x.

For any woman type y, we can similarly define

$$\tilde{H}_y(V_{\cdot y}) = \frac{1}{m_y}\sum_{j\in y}\max_{x\in\mathcal{X}_0}(V_{xy} + \eta_{xj})$$

and its Legendre-Fenchel transform $\tilde{H}^*_y$; and we obtain, for $V_{0y} = 0$,

$$\boldsymbol{\nu}_{\cdot|y} \in \partial\tilde{H}_y(\boldsymbol{V}_{\cdot y}) \text{ if and only if } \boldsymbol{V}_{\cdot y} \in \partial\tilde{H}^*_y(\boldsymbol{\nu}_{\cdot|y}).$$

We showed in Section 1.1 that at a stable matching, there exist $\boldsymbol{U}$ and $\boldsymbol{V}$ matrices, with $U_{x0} = V_{0y} = 0$, such that $U_{xy} + V_{xy} \leq \Phi_{xy}$ for all pairs of types (x, y), with equality if some individuals of these types do match ($\mu_{xy} > 0$). Now $\mu_{y|x} = \mu_{xy}/n_x$ is the proportion of men of type x who match with a woman of type y, so that $\boldsymbol{U}_{x\cdot} \in \partial\tilde{G}^*_x(\boldsymbol{\mu}_{\cdot|x})$. Similarly, $\boldsymbol{V}_{\cdot y} \in \partial\tilde{H}^*_y(\boldsymbol{\mu}_{\cdot|y})$, where (with some abuse of notation) $\mu_{x|y} \equiv \mu_{xy}/m_y$. It follows that at any stable matching, the elements of all matrices in the sum of sets

$$\partial\tilde{G}^*_x(\boldsymbol{\mu}_{\cdot|x}) + \partial\tilde{H}^*_y(\boldsymbol{\mu}_{\cdot|y})$$

are at most equal to the corresponding elements of $\boldsymbol{\Phi}$; and that if $\mu_{xy} > 0$, then the (x, y) element is uniquely defined and equals Φ_{xy}.

This somewhat cumbersome statement is much simplified if we assume that the population constitutes a "large market" and if the unobserved heterogeneity has full support.

Assumption 2 (Large market)**.** *The population contains a continuum of individuals of each type.*

Assumption 2 allows us to replace all averages with expectations, and all proportions with probabilities in our statements. To denote this change, we now take out the tildes in the notation: we write G_x and H_y instead of $\tilde{G}_x$ and $\tilde{H}_y$, and G^*_x and H^*_y instead of $\tilde{G}^*_x$ and $\tilde{H}^*_y$.

At this point, the functions G_x and H_y may still have points of non-differentiability if some unobserved heterogeneity terms have atoms; and some matching patterns may be zero if they are

dominated for all values of the heterogeneity terms. The continuous, full support assumption rules out these two cases.

Assumption 3 (Continuous full support)**.** *All distributions $\mathbb{P}_x$ and $\mathbb{Q}_y$ have full support and no mass points.*

Given Assumptions 2 and 3, all matching patterns μ_{x0}, μ_{0y}, and μ_{xy} must be positive, so that $U_{xy} + V_{xy} = \Phi_{xy}$ for all (x, y); moreover, all subgradients are gradients. This allows us to write

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

for all (x, y).

Finally, we define the *generalized entropy*: for any (X, Y) matrix $\boldsymbol{\mu}$ of feasible marriage patterns,

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{q}) = -\sum_{x \in \mathcal{X}} n_x G_x^* \left(\frac{\boldsymbol{\mu}_{x\cdot}}{n_x} \right) - \sum_{y \in \mathcal{Y}} m_y H_y^* \left(\frac{\boldsymbol{\mu}_{\cdot y}}{m_y} \right). \tag{1.5}$$

The function $\mathcal{E}$ only depends on the marriage patterns $\boldsymbol{\mu}$ and the margins $\boldsymbol{q} = (\boldsymbol{n}, \boldsymbol{m})$. It is concave; its shape depends on the distributions $(\mathbb{P}_x)$ and $(\mathbb{Q}_y)$ of the unobserved heterogeneity terms $\boldsymbol{\varepsilon}$ and $\boldsymbol{\eta}$. Note that the element μ_{xy} only comes in via the x term of the first sum and the y term of the second sum, so that

$$\frac{\partial \mathcal{E}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}) = -\frac{\partial G_x^*}{\partial \mu_{y|x}} \left(\frac{\mu_{x\cdot}}{n_x} \right) - \frac{\partial H_y^*}{\partial \mu_{x|y}} \left(\frac{\mu_{\cdot y}}{m_y} \right). \tag{1.6}$$

This suggests, and Galichon and Salanié (2022) proves, that:

Theorem 2 (Identifying the Joint Surplus)**.** *Under Assumptions 1, 2, and 3, the stable matching $\boldsymbol{\mu}$ is unique; it solves the following globally convex program:*

$$\max_{\boldsymbol{\mu}} \left(\sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} \mu_{xy} \Phi_{xy} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{q}) \right). \tag{1.7}$$

The solution is characterized by the first-order conditions

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

The objective function in (1.7) is the sum of two terms. The first one represents the social surplus from all matches when all unobserved heterogeneity terms are zero. The second one adds the contribution to the social surplus of matching on unobservables, through the generalized entropy term. The relative contributions of these two terms thus are directly related to the size of unobserved heterogeneity.

Theorem 2 shows that if a large population is governed by a separable matching model with *known* distributions of unobserved heterogeneity that are continuous and have full support, then the joint surplus can be recovered from the matching patterns. In many common specifications, the derivatives of the generalized entropy can be computed in closed form. One key caveat applies: if the distributions $\mathbb{P}_x$ and $\mathbb{Q}_y$ are only known up to say r parameters, then we can only identify $(XY - r)$ parameters of the joint surplus matrix $\mathbf{\Phi}$. This is unavoidable, as we only observe the matching patterns $\boldsymbol{\mu}$. They consist of $(XY + X + Y)$ numbers $(\mu_{xy}, \mu_{x0}, \mu_{0y})$; but only the XY conditional matching patterns $(\mu_{xy}/n_x, \mu_{xy}/m_y)$ enter the identification formula (1.6). To put it differently, all separable models are homogeneous of degree 1: rescaling the market by multiplying all n_x and m_y by the same positive number rescales all stable matching patterns by the same number.

1.3 The Data

We assume that the analyst observes a random sample of size N from a large population of households. By simple counting, she obtains estimators of the matching patterns $\hat{\mu}_{xy}, \hat{\mu}_{x0}$, and $\hat{\mu}_{0y}$ and a consistent estimator $\Sigma_{\hat{\boldsymbol{\mu}}}$ of their asymptotic variance-covariance matrix, given by the standard formula[3]:

$$\Sigma_{\hat{\boldsymbol{\mu}}} = \operatorname{diag} \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}'.$$

This also yields estimators of the margins $\boldsymbol{q}$:

$$\hat{n}_x = \hat{\mu}_{x0} + \sum_{y \in \mathcal{Y}} \hat{\mu}_{xy}$$
$$\hat{m}_y = \hat{\mu}_{0y} + \sum_{x \in \mathcal{X}} \hat{\mu}_{xy}$$

and finally, an estimator $\hat{\mathbf{\Sigma}}$ of the variance-covariance matrix of $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$.

In the following two sections, we will assume throughout that Assumptions 1, 2 and 3 hold in the population from which the data is drawn. Section 4 discusses how our estimators can be adapted to more general settings.

[3] This would be easy to adapt if sampling weights were used.

2 Minimum-distance Estimation

Under our assumptions, Theorem 2 shows that at the stable matching $\boldsymbol{\mu}$, the joint surplus matrix $\boldsymbol{\Phi}$ can be obtained by the following simple formula:

$$\Phi_{xy} = -\frac{\partial \mathcal{E}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}). \tag{2.1}$$

Suppose that the distributions $\mathbb{P}_x$ and $\mathbb{Q}_y$ are specified up to a parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^{d_{\boldsymbol{\alpha}}}$, while the joint surplus matrix $\boldsymbol{\Phi}$ is specified up to a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{d_{\boldsymbol{\beta}}}$. We denote the generalized entropy function by $\mathcal{E}^{\boldsymbol{\alpha}}$ and the parameterized surplus vector by $\boldsymbol{\Phi}^{\boldsymbol{\beta}}$. We assume that the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are point-identified[4]. Then we can use (2.1) as the basis for a minimum distance estimator, see Newey and McFadden (1994). That is, we write a mixed hypothesis as

$$\exists \boldsymbol{\lambda} = (\boldsymbol{\alpha}, \boldsymbol{\beta}), \quad \boldsymbol{D}^{\boldsymbol{\lambda}}(\boldsymbol{\mu}, \boldsymbol{q}) \equiv \boldsymbol{\Phi}^{\boldsymbol{\beta}} + \frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}, \boldsymbol{q}) = \boldsymbol{0},$$

stacking all $X \times Y$ conditions in (2.1) in a vector $\boldsymbol{D}^{\boldsymbol{\lambda}}$.

2.1 The Estimation Procedure

To estimate $\boldsymbol{\lambda} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, we will replace the population values of $\boldsymbol{\mu}$ and $\boldsymbol{q}$ with the observed values $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{q}}$. Now the sample will obviously be smaller than the population. Fortunately, we saw in Section 1.2 that all separable matching models exhibit constant returns to scale. More precisely, the derivatives of the entropy are homogeneous of degree 0 in $(\boldsymbol{\mu}, \boldsymbol{q})$, so that the mixed hypothesis is scale-invariant.

The smaller size of the sample raises a more serious issue. Even though under Assumptions 1, 2 and 3 all μ_{xy}, μ_{x0} and μ_{0y} are positive, some of them may be quite small. This may result in zero values in the observed sample. If, as in many common specifications, the derivatives of the entropy are infinite in zero, this will create difficulties. We will discuss this further in Section 4. For now, we rule out such "zero cells".

Assumption 4 (No zero cells)**.** *All values of* $\hat{\boldsymbol{\mu}}$ *are positive.*

Under Assumption 4, we can both test our mixed hypothesis and estimate $\boldsymbol{\lambda}$ by minimizing $\|\boldsymbol{D}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})\|^2_{\boldsymbol{S}}$ for some positive definite $(X \times Y, X \times Y)$ matrix $\boldsymbol{S}$. By the general theory of minimum

[4] Note that in general, this requires that $d_{\boldsymbol{\alpha}} + d_{\boldsymbol{\beta}} \leq X \times Y$.

distance estimators, we know that this yields a consistent estimator of $\boldsymbol{\lambda}$ if the model is well-specified and $\boldsymbol{\lambda}$ is point-identified. Moreover, if we choose $\boldsymbol{S} = \hat{\boldsymbol{\Omega}}^{-1}$ where $\hat{\boldsymbol{\Omega}}$ consistently estimates $V\boldsymbol{D}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$, the minimum distance estimator will reach its efficiency bound[5]. Finally, under this choice of $\boldsymbol{S}$ the minimized value of the squared norm follows a χ^2 of degree $X \times Y - d_{\boldsymbol{\alpha}} - d_{\boldsymbol{\beta}}$ under the mixed hypothesis. This gives us a straightforward specification test.

This procedure is summarized in Box 1.

Box 1: minimum-distance estimation (general case)

1. Choose any positive definite matrix $\boldsymbol{S}$ and minimize over $\boldsymbol{\lambda} \in \mathbb{R}^{d_\alpha + d_\beta}$

$$\|\boldsymbol{D}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})\|^2_{\boldsymbol{S}} = \sum_{x,y,z,t} S_{xy,zt} \left(\Phi^{\boldsymbol{\beta}}_{xy} + \frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}}) \right) \left(\Phi^{\boldsymbol{\beta}}_{zt} + \frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{zt}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}}) \right).$$

   This gives a consistent estimator $\boldsymbol{\lambda}^*$.

2. Use the delta method to estimate the variance $\boldsymbol{\Omega}^*$ of $\boldsymbol{D}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$; let $\boldsymbol{S}^* = (\boldsymbol{\Omega}^*)^{-1}$.

3. Minimize as in 1. with $\boldsymbol{S} = \boldsymbol{S}^*$, to obtain another consistent estimator $\hat{\boldsymbol{\lambda}}$.

4. The variance-covariance matrix of $\hat{\boldsymbol{\lambda}}$ is consistently estimated by

$$\left(\hat{\boldsymbol{F}}' \boldsymbol{S}^* \hat{\boldsymbol{F}} \right)^{-1}$$

   where $\hat{\boldsymbol{F}}$ is the Jacobian of $\boldsymbol{D}^{\boldsymbol{\lambda}}$ with respect to $\boldsymbol{\lambda}$ at $\hat{\boldsymbol{\lambda}}$.

5. Under the null of correct specification, the statistic

$$\hat{\boldsymbol{T}} = \left(\hat{\boldsymbol{D}}^{\hat{\boldsymbol{\lambda}}} \right)' \boldsymbol{S}^* \hat{\boldsymbol{D}}^{\hat{\boldsymbol{\lambda}}}$$

   converges to a $\chi^2(X \times Y - d_{\boldsymbol{\alpha}} - d_{\boldsymbol{\beta}})$ distribution.

2.2 The Linear Case

In general, the optimization problem in Box 1 is *not* globally convex. There is an important special case in which it is in fact quadratic, and minimum-distance estimation is a particularly appealing

[5] Given an initial consistent estimate of $\boldsymbol{\lambda}$ and the estimator $\hat{\boldsymbol{\Sigma}}$ of the variance-covariance matrix, such a consistent estimate $\hat{\boldsymbol{\Omega}}$ can be obtained by applying the delta method.

strategy.

2.2.1 Linearity in the Parameters

Suppose that both the derivatives of the generalized entropy function $\mathcal{E}^{\boldsymbol{\alpha}}$ and the surplus matrix $\boldsymbol{\Phi}^{\boldsymbol{\beta}}$ are linear in the parameters:

$$\frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}) = e^0_{xy}(\boldsymbol{\mu}, \boldsymbol{q}) + e_{xy}(\boldsymbol{\mu}, \boldsymbol{q}) \cdot \boldsymbol{\alpha} \tag{2.2}$$

where for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $e^0_{xy}(\boldsymbol{\mu}, \boldsymbol{q})$ is a scalar and $e_{xy}(\boldsymbol{\mu}, \boldsymbol{q})$ is a vector of size $d_{\boldsymbol{\alpha}}$; and

$$\Phi^{\boldsymbol{\beta}}_{xy} = \boldsymbol{\phi}_{xy} \cdot \boldsymbol{\beta} \tag{2.3}$$

where $\boldsymbol{\phi}_{xy}$ is of size $d_{\boldsymbol{\beta}}$.

The basis functions $\boldsymbol{e}_0(\boldsymbol{\mu}, \boldsymbol{q})$, $\boldsymbol{e}(\boldsymbol{\mu}, \boldsymbol{q})$ and $\boldsymbol{\phi}$ are imposed by the modeler. Then

$$D^{\boldsymbol{\lambda}}_{xy}(\boldsymbol{\mu}, \boldsymbol{q}) = \boldsymbol{\phi}_{xy} \cdot \boldsymbol{\beta} + e^0_{xy}(\boldsymbol{\mu}, \boldsymbol{q}) + e_{xy}(\boldsymbol{\mu}, \boldsymbol{q}) \cdot \boldsymbol{\alpha}$$

is linear in the parameters $\boldsymbol{\lambda}$. Note that the parameter-free part $\boldsymbol{e}^0$ is necessary in order to identify $\boldsymbol{\lambda}$: since we are minimizing the norm of $\boldsymbol{D}^{\lambda}$, making $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ equal $\mathbf{0}$ would give a trivial solution otherwise. This is simply due to the fact that the scale of the error terms is not identified in discrete-choice models, as will become clearer in our examples.

These linearity assumptions call for two remarks. Condition (2.2) trivially holds in models where the $\mathbb{P}_x$ and $\mathbb{Q}_y$ are parameter-free, like the ubiquitous Choo and Siow (2006) specification. More generally, it holds in families of models where the parameters $\boldsymbol{\alpha}$ change the location of the unobserved heterogeneity terms and/or rescale them identically across partner types.

Proposition 3. *Linear derivatives in identical-scale models*

Suppose that there exist parameter-free distributions $(\mathbb{P}^0_x)$ *and* $(\mathbb{Q}^0_y)$ *such that*

- *for every* x*, the unobserved heterogeneity terms*

$$\varepsilon_{iy} = \boldsymbol{a}_{xy} \cdot \boldsymbol{\alpha} + \boldsymbol{b}_x \cdot \boldsymbol{\alpha}\, \varepsilon^0_{iy}$$

  where the vector (ε^0_{iy}) *is distributed as* $\mathbb{P}^0_x$*; the vectors* $\boldsymbol{a}_{xy}$ *and* $\boldsymbol{b}_x$ *are known and non-random; and* $\boldsymbol{b}_x \cdot \boldsymbol{\alpha}$ *is positive for all* $\boldsymbol{\alpha}$*;*

- *for every* y*, the unobserved heterogeneity terms*

$$\eta_{xj} = \boldsymbol{c}_{xy} \cdot \boldsymbol{\alpha} + \boldsymbol{d}_y \cdot \boldsymbol{\alpha}\, \eta^0_{xj}$$

where the vector (η^0_{xj}) *is distributed as* $\mathbb{Q}^0_y$*; the vectors* $\mathbf{c}_{xy}$ *and* $\mathbf{d}_y$ *are known and non-random; and* $\mathbf{d}_y \cdot \boldsymbol{\alpha}$ *is positive for all* $\boldsymbol{\alpha}$*.*

Then the derivatives of the generalized entropy $\mathcal{E}$ *with respect to* $\boldsymbol{\mu}$ *are linear in* $\boldsymbol{\alpha}$*.*

While the statements in Proposition 3 may look cumbersome, they give a simple recipe to generate a market with linear derivatives of the generalized entropy:

1. pick any families of distributions $(\mathbb{P}^0_x)$ and $(\mathbb{Q}^0_y)$ and draw random vectors $(\boldsymbol{\varepsilon}^0_{i\cdot})$ and $(\boldsymbol{\eta}^0_{\cdot j})$ from them
2. shift the location of the $\boldsymbol{\varepsilon}^0$ and $\boldsymbol{\eta}$ terms arbitrarily
3. and allow for heteroskedasticity that can depend in an arbitrary manner on x (resp. y) *only* for $\boldsymbol{\varepsilon}^0$ (resp. $\boldsymbol{\eta}$).

As we will see in Section 2.3, Proposition 3 applies to several leading examples. It does rule out, for instance, models in which the preference shock of men of type x for women of type y has a variance that depends on y.

The joint surplus is the sum of the pre-transfer utilities of both partners, maximized over household decisions. As such, it is a function of prices, income, taxes, and other environmental variables. The linearity of the surplus function (Condition (2.3))is clearly a restrictive condition. We prefer to view (2.3) as a "seminonparametric" or flexible expansion over a set of basis functions. If the set of basis functions is large enough, the estimated joint surplus can then be projected on a class of structural models that are nonlinear in the parameters.

2.2.2 Estimation in the Linear Case

Under conditions (2.2) and (2.3), the minimum distance estimator can be implemented by linear least-squares. Let $\hat{\boldsymbol{F}}$ denote the $(X \times Y, d_{\boldsymbol{\alpha}} + d_{\boldsymbol{\beta}})$ matrix that stacks $\hat{\boldsymbol{e}} = \boldsymbol{e}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ and $\boldsymbol{\phi}$ vertically, so that $\boldsymbol{D}^\lambda(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{r}}) = \hat{\boldsymbol{e}}^0 + \hat{\boldsymbol{F}}\boldsymbol{\lambda}$, where $\hat{\boldsymbol{e}}^0 = \boldsymbol{e}^0(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$. Then for any choice of $\boldsymbol{S}$, the minimum distance estimator $\hat{\boldsymbol{\lambda}}$ solves the linear system

$$\left(\hat{\boldsymbol{F}}'\boldsymbol{S}\hat{\boldsymbol{F}}\right)\ \hat{\boldsymbol{\lambda}} = -\hat{\boldsymbol{F}}'\boldsymbol{S}\hat{\boldsymbol{e}}^0; \tag{2.4}$$

that is, $\hat{\boldsymbol{\lambda}}$ is simply the GLS estimator in the linear regression

$$-\hat{\boldsymbol{e}}^0 = \hat{\boldsymbol{F}}\boldsymbol{\lambda} + \boldsymbol{u} \tag{2.5}$$

when $V\boldsymbol{u} = \boldsymbol{S}^{-1}$.

The resulting QGLS procedure is summarized in Box 2.

Box 2: min-distance estimation (linear case)

1. Evaluate $\hat{\boldsymbol{e}}^0 \equiv \boldsymbol{e}^0(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ and $\hat{\boldsymbol{e}} \equiv \boldsymbol{e}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$; stack $\boldsymbol{e}$ and $\boldsymbol{\phi}$ vertically in a matrix $\hat{\boldsymbol{F}}$.
2. Choose any positive definite matrix $\boldsymbol{S}$ and run the GLS regression (2.5) to get a consistent estimator $\boldsymbol{\lambda}^*$.
3. Use the delta method to estimate the variance $\boldsymbol{\Omega}^*$ of $\boldsymbol{D}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$; let $\boldsymbol{S}^* = (\boldsymbol{\Omega}^*)^{-1}$.
4. Take $\boldsymbol{S} = \boldsymbol{S}^*$ and run the GLS regression (2.5) to obtain $\hat{\boldsymbol{\lambda}}$. Denote $\hat{\boldsymbol{u}}$ the residuals of the regression.
5. The variance-covariance matrix of $\hat{\boldsymbol{\lambda}}$ is consistently estimated by
$$\left(\hat{\boldsymbol{F}}'\boldsymbol{S}^*\hat{\boldsymbol{F}}\right)^{-1}.$$
6. Under the null of correct specification, the statistic
$$\hat{\boldsymbol{T}} = \hat{\boldsymbol{u}}'\boldsymbol{S}^*\hat{\boldsymbol{u}}$$
converges to a $\chi^2(X \times Y - d_{\boldsymbol{\alpha}} - d_{\boldsymbol{\beta}})$ distribution.

If moreover all distributions $(\mathbb{P}_x)$ and $(\mathbb{Q}_y)$ are parameter-free, then $\boldsymbol{\lambda}$ only consists of $\boldsymbol{\beta}$; the generalized entropy is also parameter-free and $\hat{\boldsymbol{e}}$ is zero. Then the matrix $\boldsymbol{\Omega}^*$ is simply the variance of $\boldsymbol{e}^0(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ since $\boldsymbol{\phi\beta}$ does not depend on the observed matching. Moreover, $\hat{\boldsymbol{F}}$ reduces to the matrix $\boldsymbol{\phi}$. The estimators of $\boldsymbol{\lambda} = \boldsymbol{\beta}$ can be obtained following the procedure described in Box 3.

Box 3: minimum-distance estimator (linear case with parameter-free entropy)

1. Evaluate $\boldsymbol{\Omega}^* = V\hat{\boldsymbol{e}}^0$ and $\boldsymbol{S}^* = (\boldsymbol{\Omega}^*)^{-1}$.
2. Solve the linear system $(\boldsymbol{\phi}'\boldsymbol{S}^*\boldsymbol{\phi})\ \boldsymbol{\beta} = -\boldsymbol{\phi}'\boldsymbol{S}^*\hat{\boldsymbol{e}}^0$.
3. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is consistently estimated by
$$\left(\boldsymbol{\phi}'\boldsymbol{S}^*\boldsymbol{\phi}\right)^{-1}.$$
4. Under the null of correct specification, the statistic
$$\hat{\boldsymbol{T}} = \left(\boldsymbol{\phi}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}^0\right)' \boldsymbol{S}^* \left(\boldsymbol{\phi}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}^0\right)$$
converges to a $\chi^2(X \times Y - d_{\boldsymbol{\beta}})$ distribution.

The one remaining difficulty in the procedures of Boxes 2 and 3 is the evaluation of $\boldsymbol{\Omega}^*$ by the delta method. Now
$$V\hat{\boldsymbol{D}}^{\boldsymbol{\lambda}} = \begin{pmatrix} \mathcal{E}'_{\boldsymbol{\mu\mu}} & \mathcal{E}'_{\boldsymbol{\mu q}} \end{pmatrix} V \begin{pmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{q}} \end{pmatrix} \begin{pmatrix} \mathcal{E}_{\boldsymbol{\mu\mu}} \\ \mathcal{E}_{\boldsymbol{\mu q}} \end{pmatrix}.$$
where $\mathcal{E}_{\boldsymbol{\mu}} = \boldsymbol{e}^0 + \boldsymbol{e} \cdot \boldsymbol{\alpha}$. Computing $\boldsymbol{\Omega}^*$ therefore requires evaluating the derivatives in $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ of $\hat{\boldsymbol{e}}^0$ and (unless the entropy is parameter-free) $\hat{\boldsymbol{e}}$, which constitute the elements of $\mathcal{E}_{\boldsymbol{\mu\mu}}$ and $\mathcal{E}_{\boldsymbol{\mu q}}$. While this can be done in closed form for several commonly used specifications, it may require numerical approximation in general. It is easy to see from the definition in (1.5) that the first derivative of $\mathcal{E}$ with respect to μ_{xy} only depends on the conditional matching patterns $\mu_{\cdot|x} = (\mu_{x1}/n_x, \ldots, \mu_{xY}/n_x)$ of men of type x, and on those of women of type y. As a consequence, the Hessians of $\mathcal{E}$ are very sparse and are often easy to evaluate.

2.2.3 Estimation under Linear Derivatives

Proposition 3 shows that many useful classes of models can be parameterized in such a way that Condition (2.2) holds:
$$\frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}) = e^0_{xy}(\boldsymbol{\mu}, \boldsymbol{q}) + e_{xy}(\boldsymbol{\mu}, \boldsymbol{q}) \cdot \boldsymbol{\alpha}.$$
On the other hand, the analyst may favor a structural model of joint surplus $\boldsymbol{\Phi}$ that cannot be written as a linear function of its parameters $\boldsymbol{\beta}$. The linearity of the derivatives of the generalized

entropy still simplifies minimum distance estimation here.

Given values for the parameters $\boldsymbol{\beta}$, the equation

$$\boldsymbol{\Phi}^{\boldsymbol{\beta}} = -\hat{\boldsymbol{e}}_{\mathbf{0}} - \hat{\boldsymbol{e}} \cdot \boldsymbol{\alpha}$$

implies that for any matrix $\boldsymbol{M}$ such that $\boldsymbol{M}\boldsymbol{e} = \mathbf{0}$,

$$\boldsymbol{M}(\hat{\boldsymbol{e}}_0 + \boldsymbol{\Phi}^{\boldsymbol{\beta}}) = \mathbf{0}.$$

This gives a family of moment conditions that yield a consistent estimator of $\hat{\boldsymbol{\beta}}$. In a second step, a least squares estimator $\hat{\boldsymbol{\alpha}}$ can be obtained by regressing $(\hat{\boldsymbol{e}}_0 + \boldsymbol{\Phi}^{\hat{\boldsymbol{\beta}}})$ on $-\hat{\boldsymbol{e}}$.

The matrix M used in the first step can simply be the projector on $\hat{\boldsymbol{e}}^{\perp}$

$$\boldsymbol{M} = \boldsymbol{I} - \hat{\boldsymbol{e}} \left(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}\right)^{-1} \hat{\boldsymbol{e}}';$$

in the second step, generalized least squares can be used as in Box 1 so as to maximize efficiency.

2.3 Examples

We start with two examples for which the generalized entropy and its derivatives are available in closed form; in both cases, the derivatives are linear in the parameters $\boldsymbol{\alpha}$. It is easy to see that both cases are covered by Proposition 3; we give self-contained derivations below.

In our third example, the calculation requires finding the fixed point of a contraction, in a way that is similar to the "Berry inversion" of empirical industrial organization (Berry, 1994).

2.3.1 The Choo and Siow Model with Heteroskedasticity

Let us start with an easy extension of the Choo and Siow (2006) logit model: the distributions $\mathbb{P}_x$ and $\mathbb{Q}_y$ are type I-EV iid vectors with unknown scale factors σ_x and τ_y respectively. Collecting the scale factors in vectors $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$, the parameters of the generalized entropy are $\boldsymbol{\alpha} = (\boldsymbol{\sigma}, \boldsymbol{\tau})$. The derivatives of $\mathcal{E}^{\boldsymbol{\alpha}}$ with respect to $\boldsymbol{\mu}$ are linear in $\boldsymbol{\alpha}$:

$$\frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}) = -\sigma_x \log \frac{\mu_{xy}}{\mu_{x0}} - \tau_y \log \frac{\mu_{xy}}{\mu_{0y}}$$

where $\mu_{x0} = n_x - \sum_{y \in \mathcal{Y}} \mu_{xy}$ and $\mu_{0y} = m_y - \sum_{x \in \mathcal{X}} \mu_{xy}$. As explained earlier, we need to normalize the scale of the vector $\boldsymbol{\lambda}$. The most natural way to do it is to fix the value of one of the parameters in $\boldsymbol{\alpha}$. If for instance we fix $\sigma_1 = 1$, then

- $e^0_{1y} = -\log(\mu_{1y}/\mu_{10})$ and $\boldsymbol{e}_{1y} = \mathbf{0}$ for all $y \in \mathcal{Y}$;
- for $x = 2, \ldots, X$, the $(x-1)$-th element of $\boldsymbol{e}_{xy}$ is $-\log(\mu_{xy}/\mu_{x0})$ and its $(X-1+y)$-th element is $-\log(\mu_{xy}/\mu_{0y})$;
- all other elements of $\boldsymbol{e}^0$ and $\boldsymbol{e}$ are zero.

The Choo and Siow homoskedastic model obtains when all σ_x and τ_y equal one; a gender-heteroskedastic model would have all σ_x equal to one and all τ_y equal to an unknown τ. Chiappori et al. (2017) applied a minimum distance estimator to the homoskedastic and heteroskedastic logit models.

The second derivatives of the generalized entropy take a very simple form in this class of models:

$$\frac{\partial^2 \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy} \partial \mu_{zt}}(\boldsymbol{\mu}, \boldsymbol{q}) = -\frac{\sigma_x}{\mu_{x0}} \mathbb{1}(z = x) - \frac{\tau_y}{\mu_{0y}} \mathbb{1}(t = y) - \frac{\sigma_x + \tau_y}{\mu_{xy}} \mathbb{1}(z = x, t = y) \tag{2.6}$$

and

$$\frac{\partial^2 \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy} \partial n_z}(\boldsymbol{\mu}, \boldsymbol{q}) = \frac{\sigma_x}{\mu_{x0}} \mathbb{1}(z = x); \qquad \frac{\partial^2 \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy} \partial m_t}(\boldsymbol{\mu}, \boldsymbol{q}) = \frac{\tau_y}{\mu_{0y}} \mathbb{1}(t = y). \tag{2.7}$$

2.3.2 Nested Logit

Consider a two-layer nested logit model. Take men of type x first. Alternative 0 (singlehood) is obviously special; we put it alone in its nest. Each other nest $n \in \mathcal{N}_x$ contains alternatives $y \in \mathcal{Y}_n$. The correlation of alternatives within nest n is proxied by $1 - (\rho^x_n)^2$ (with $\rho^x_0 = 1$ for the nest made of alternative 0). Similarly, for women of type y, alternative 0 is in a nest by itself with parameter $\delta^y_0 = 1$ and alternatives $x \in \mathcal{X}_{n'}$ are in a nest $n \in \mathcal{N}'_y$ with parameter $\delta^y_{n'}$. We collect the parameters $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ into $\boldsymbol{\alpha}$.

The formulæ in Example 2.1 of Galichon and Salanié (2022) imply that if y is in nest $n \in \mathcal{N}_x$ and x is in nest $n' \in \mathcal{N}_y$, then

$$\begin{aligned}\frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}) = &-\rho^x_n \log \frac{\mu_{xy}}{\mu_{x0}} - (1 - \rho^x_n) \log \frac{\mu_{xn}}{\mu_{x0}} \\ &- \delta^y_{n'} \log \frac{\mu_{xy}}{\mu_{0y}} - (1 - \delta^y_{n'}) \log \frac{\mu_{n'y}}{\mu_{0y}},\end{aligned} \tag{2.8}$$

where we defined $\mu_{xn} = \sum_{t \in \mathcal{Y}_n} \mu_{xt}$ and $\mu_{n'y} = \sum_{z \in \mathcal{X}_{n'}} \mu_{zy}$. Once again, this is linear in the parameters $\boldsymbol{\alpha}$; it remains linear if we impose constraints on the nests (for instance, that $\mathcal{N}_x$ is the same for all types x) and/or linear constraints on the $\boldsymbol{\rho}$ parameters (for instance, that ρ^x_n only depends on n). Here too the second derivatives, while more complicated than in the multinomial logit, can be computed in closed form.

2.3.3 Mixed Logit

Let us now describe a random coefficient logit model. Consider a man i of type x, endowed with preferences $\boldsymbol{e}_i$ over a set of d observable characteristics $\boldsymbol{Z}$ of potential partners. We add an idiosyncratic shock $\boldsymbol{\zeta}_i$ that is distributed as a standard iid type I extreme value vector over $\mathbb{R}^{Y+1}$, independently of $\boldsymbol{e}_i$, and a scale factor $s > 0$:

$$\varepsilon_{iy} = \sum_{k=1}^{d} Z_{yk} e_{ik} + s\zeta_{iy}$$

or in matrix form: $\boldsymbol{\varepsilon} = \boldsymbol{Z}\boldsymbol{e} + s\boldsymbol{\zeta}$. This specification is standard in empirical IO[6].

Let individual preferences $\boldsymbol{e}$ of men of type x have distribution $\mathbb{P}_x^e$. We will seek to estimate the parameters $\boldsymbol{\beta}$ of the joint surplus, the scale factor s, and the parameters of the distributions $\mathbb{P}_x^e$. We collect s and the parameters of $\mathbb{P}_x^e$ in a vector $\boldsymbol{\alpha}$. Note that except in trivial cases, this model does not have the "identical-scale" property that underlies Proposition 3. In fact, the derivatives of the generalized entropy generally fail to be linear in $\boldsymbol{\alpha}$.

To compute the derivative of the generalized entropy function, we recall from Galichon and Salanié (2022) that we only need to replace the max operator in the definition of the function G_x with a regularized "softmax" that accounts for the integration over the shocks $\boldsymbol{e}$:

$$G_x(\boldsymbol{U};\boldsymbol{\alpha}) = \int s \log \sum_{y=0,1,\ldots,Y} \exp\left(\frac{U_y + (\boldsymbol{Z}\boldsymbol{e})_y}{s}\right) d\mathbb{P}_x^e(\boldsymbol{e}).$$

This gives

$$G_x^*(\boldsymbol{\nu};\boldsymbol{\alpha}) = \max_{\boldsymbol{U}\in\mathbb{R}^{Y+1}} \left[\sum_{y\in\mathcal{Y}_0} \nu_y U_y - \int s \log \sum_{y=0,1,\ldots,Y} \exp\left(\frac{U_y + (\boldsymbol{Z}\boldsymbol{e})_y}{s}\right) d\mathbb{P}_x^e(\boldsymbol{e})\right].$$

By the envelope theorem, the derivative of $G_x^*(\boldsymbol{\nu};\boldsymbol{\alpha})$ with respect to $\boldsymbol{\nu}$ is the vector $\boldsymbol{U}$ that solves the system

$$\nu_y = \int \frac{\exp((U_y + (\boldsymbol{Z}\boldsymbol{e})_y)/s)}{\sum_{t=0,1,\ldots,Y} \exp((U_t + (\boldsymbol{Z}\boldsymbol{e})_t)/s)} d\mathbb{P}_x^e(\boldsymbol{e}) \quad \forall y = 1,\ldots,Y.$$

This is exactly isomorphic to the inversion problem in Berry et al. (1995), with the unknown $\boldsymbol{U}$ standing for the product effects and $\boldsymbol{\nu}$ playing the role of the product market shares[7]. After replacing $\boldsymbol{\nu}$ with the observed $\boldsymbol{\mu}_{x\cdot}/n_x$, the system can be solved by any of the algorithms that are standard

[6] In Berry et al. (1995), the covariates in $\boldsymbol{Z}$ stand for the observed characteristics of the products; the $\boldsymbol{e}$ are individual valuations of these characteristics; and the $\boldsymbol{\zeta}$ are idiosyncratic shocks.

[7] The limit case $s = 0$ yields the pure characteristics model of Berry and Pakes (2007).

in this literature. The solution gives row x of the matrix $\boldsymbol{U}$. Proceeding in the same way for other types of men, and solving for $\boldsymbol{V}$ for women, gives the derivatives of the generalized entropy function:

$$\frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}) = -\frac{\partial G_x^*}{\partial \nu_{xy}}\left(\frac{\boldsymbol{\mu}_{x\cdot}}{n_x}\right) - \frac{\partial H_y^*}{\partial \nu_{xy}}\left(\frac{\boldsymbol{\mu}_{\cdot y}}{m_y}\right) = -U_{xy} - V_{xy} = -\Phi_{xy}.$$

3 Moment-based Estimation by Poisson Regression

Let us now return to the linear case with a parameter-free generalized entropy: $\boldsymbol{\Phi}^{\boldsymbol{\beta}} = \boldsymbol{\phi}\boldsymbol{\beta}$ and the function $\mathcal{E}$ is known (or assumed).

Galichon and Salanié (2022) introduced a moment-matching procedure that gives a consistent estimator of the parameter vector $\boldsymbol{\beta}$ if the model is well-specified. The *moment matching estimator* equalizes the observed and simulated *comoments*, that is the expectations of the basis functions $\boldsymbol{\phi}$ under the observed and simulated matching patterns:

$$\sum_{x,y} \hat{\mu}_{xy}\boldsymbol{\phi}_{xy} = \sum_{x,y} \mu_{xy}^{\boldsymbol{\beta}}\boldsymbol{\phi}_{xy},$$

where $\boldsymbol{\mu}^{\boldsymbol{\beta}}$ denotes the stable matching patterns for the parameter vector $\boldsymbol{\beta}$. As explained in Galichon and Salanié (2022), these are the first-order conditions of the following maximization problem:

$$\max_{\boldsymbol{\beta}} \left(\hat{\boldsymbol{\mu}}\boldsymbol{\Phi}^{\boldsymbol{\beta}} - \mathcal{W}(\boldsymbol{\beta}, \boldsymbol{q})\right) \tag{3.1}$$

where

$$\mathcal{W}(\boldsymbol{\beta}, \boldsymbol{q}) = \max_{\boldsymbol{\mu}} \left(\boldsymbol{\mu}\boldsymbol{\Phi}^{\boldsymbol{\beta}} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{q})\right) \tag{3.2}$$

is the value of the total joint surplus. To see this, first apply the envelope theorem to (3.2) to obtain

$$\frac{\partial \mathcal{W}}{\partial \boldsymbol{\beta}} = \boldsymbol{\mu}^{\boldsymbol{\beta}}\frac{\partial \boldsymbol{\Phi}^{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} = \boldsymbol{\mu}^{\boldsymbol{\beta}}\boldsymbol{\phi};$$

then use the first-order condition in (3.1):

$$\hat{\boldsymbol{\mu}}\boldsymbol{\phi} = \boldsymbol{\mu}^{\boldsymbol{\beta}}\boldsymbol{\phi}.$$

Moreover, both of these maximization problems are globally convex. We now show that in the specific (but popular) case of the Choo and Siow (2006) model, moment matching can be reformulated as a generalized linear model, and estimated by a Poisson regression with two-sided fixed effects.

Define $\mathcal{A} = \mathcal{X} \times \mathcal{Y} \cup \mathcal{X} \times \{0\} \cup \{0\} \times \mathcal{Y}$ the set of possible marital arrangements. For $(x, y) \in A$, we denote by w_{xy} the size of household (x, y): it is 2 if the household is a couple and 1 if it is a single. This defines a vector $w \in \mathbb{R}^{\mathcal{A}}$. The following theorem summarizes our results.

Theorem 4 (Estimating the logit model with a Poisson regression)**.** *In the Choo and Siow model, the moment-matching estimator $\hat{\boldsymbol{\beta}}$ is the solution to a Poisson regression of $(\hat{\mu}_{xy})_{xy\in\mathcal{A}}$ on $\left(\Phi^{\boldsymbol{\beta}}_{xy}/w_{xy}\right)_{xy\in\mathcal{A}}$, with x- and y- fixed effects and with weights w_{xy} defined above, and where we take by convention $\Phi^{\beta}_{x0}=0$ and $\Phi^{\beta}_{0y}=0$ and $a_0=0$ and $b_0=0$. In other words, β is the solution to*

$$\max_{\beta_k,a_x,b_y}\sum_{xy\in\mathcal{A}} w_{xy}\hat{\mu}_{xy}\left(\frac{\Phi^{\beta}_{xy}-a_x-b_y}{w_{xy}}\right)-\sum_{xy\in\mathcal{A}} w_{xy}\exp\left(\frac{\Phi^{\beta}_{xy}-a_x-b_y}{w_{xy}}\right).$$

The proof of Theorem 4 is given in Appendix C. The result is very useful in that it allows for inference on $\boldsymbol{\beta},\boldsymbol{u}$ and $\boldsymbol{v}$ in semilinear logit models with standard statistical packages such as `glm` in R, or `scikit-learn` in Python. Note that like Santos Silva and Tenreyro (2006) in the international trade literature, we end up fitting a Poisson regression to a model that is definitely not generated by a Poisson count process. The motivation is different, however. They start from a semiparametric model of the gravity equation and use the robustness of the Poisson pseudo-maximum likelihood estimator. We start from a more complex, fully specified structural model and we show that a semiparametric estimator (moment-matching) is numerically equivalent to the maximum likelihood estimator of a Poisson model[8].

In the sequel we will denote $\boldsymbol{I}_m$ the (m,m) identity matrix; $\boldsymbol{p}_{(m,n)}$ the (m,n) matrix whose elements all equal p; and $\boldsymbol{p}_m\equiv\boldsymbol{p}_{(m,1)}$. Also, we say that we stack an (X,Y) matrix in "row-major order" when we create a vector of $X\times Y$ elements whose first Y elements are the first row of the matrix, etc. If $A\in\mathbb{R}^{m\times n}$ and $B\in\mathbb{R}^{p\times q}$, then the Kronecker product $A\otimes B$ is a matrix of size $(mp)\times(nq)$ defined by a blockwise expression as:

$$A\otimes B=\begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B\\ a_{21}B & a_{22}B & \cdots & a_{2n}B\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B\end{bmatrix}$$

where each $a_{ij}B$ denotes the matrix B scaled by the scalar a_{ij}.

[8]The formal analogy between the (one-sided) multinomial logit model and a Poisson process is known in the statistics literature—see e.g. Palmgren (1981).

Box 4: GLM estimator (linear case with logit heterogeneity)

1. Flatten the observed matching patterns $\hat{\boldsymbol{\mu}}$ into a vector of size $|\mathcal{A}|$, by first stacking the elements $xy \in \mathcal{X} \times \mathcal{Y}$ in row-major order, then adding the elements $x0 \in \mathcal{X} \times \{0\}$, and finally adding the elements $0y \in \{0\} \times \mathcal{Y}$.

2. For each basis function $k = 1, \ldots, K$, use the same ordering to create a vector $\boldsymbol{\phi}^k \equiv (\phi^k_{xy})_{xy \in \mathcal{A}}$. Then combine these K column vectors of size $|\mathcal{A}|$ into an $|\mathcal{A}| \times K$ matrix.

3. Using the same ordering again, define the vector $\boldsymbol{w}$ in $\mathbb{R}^{\mathcal{A}}$:

$$\boldsymbol{w} = (\mathbf{2}'_{X \times Y}, \mathbf{1}'_X, \mathbf{1}'_Y)'.$$

4. Finally, define the $|\mathcal{A}| \times (K + X + Y)$ matrix $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{\phi}/2 & -\frac{1}{2}\boldsymbol{I}_X \otimes \mathbf{1}_{(Y,1)} & -\frac{1}{2}\mathbf{1}_{(X,1)} \otimes \boldsymbol{I}_Y \\ \mathbf{0}_{(X,K)} & -\boldsymbol{I}_X & \mathbf{0}_{(X,Y)} \\ \mathbf{0}_{(Y,K)} & \mathbf{0}_{(Y,X)} & -\boldsymbol{I}_Y \end{pmatrix}.$$

5. Run a Poisson regression of $\hat{\mu}$ on $\boldsymbol{Z}$ with weights $\boldsymbol{w}$. Do not add any fixed effect, as these have already been included in the design of $\boldsymbol{Z}$. Let $\hat{\boldsymbol{\gamma}}$ be the vector of coefficients obtained this way; it solves

$$\max_{\gamma \in \mathbb{R}^{K+X+Y}} \left(\sum_{a \in \mathcal{A}} w_a \hat{\mu}_a \left(Z\gamma\right)_a - \sum_{a \in \mathcal{A}} w_a \exp\left(\left(Z\gamma\right)_a\right) \right).$$

6. Decompose $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{a}}', \hat{\boldsymbol{b}}')' \in \mathbb{R}^{K+X+Y}$. Then $\hat{\boldsymbol{\beta}}$ is the moment-matching estimator, and a_x and b_y are the x- and y- fixed effects.

As a result, we get that:

Theorem 5. *The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ can be estimated with*

$$\hat{V}\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{A}}^{-1}\, \hat{\boldsymbol{B}}\, \hat{\boldsymbol{A}}^{-1}$$

where, letting $\boldsymbol{W} = \operatorname{diag} w$, we have

$$\begin{aligned} \hat{\boldsymbol{A}} &= \left(\boldsymbol{Z}^\top \boldsymbol{W} \operatorname{diag}\left(\exp\left(\boldsymbol{Z}\gamma\right)\right) \boldsymbol{Z}\right) \\ &= \sum_{a \in \mathcal{A}} w_a \exp(\boldsymbol{Z}_a \hat{\boldsymbol{\gamma}}) \boldsymbol{Z}_a^\top \boldsymbol{Z}_a \end{aligned}$$

and

$$\begin{aligned}\hat{\boldsymbol{B}} &= \boldsymbol{Z}^\top \boldsymbol{W}(\text{diag}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top)\boldsymbol{W}\boldsymbol{Z} \\ &= \sum_{a\in\mathcal{A}} w_a \hat{\mu}_a \boldsymbol{Z}_a^\top \boldsymbol{Z}_a - \sum_{a,a'\in\mathcal{A}} w_a w_{a'}\ \hat{\mu}_a \hat{\mu}_{a'}\ \boldsymbol{Z}_a^\top \boldsymbol{Z}_{a'}.\end{aligned}$$

4 Dealing with Common Issues

It is often the case in applications that some (x, y) cells contain no match: $\hat{\mu}_{xy} = 0$. This generally creates no particular difficulty for our GLM estimator. On the other hand, it may require adapting our minimum-distance estimator. We discuss several solutions in Section 4.1.

Data aggregation occurs when we only observe matching patterns at a more aggregated level than the (x, y) cell, where we imposed separability. Assumption 1 rules out interactions between the unobserved characteristics of the two partners, conditionally on (x, y). Suppose for instance that we collected data on age and education for women, but only on education for men. Separability at the age $\times$ education level requires that if men i and i' have the same age and education, and women j and j' have the same age and education, then

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

If now men k and k' have the same education and different ages, it is quite possible that

$$\tilde{\Phi}_{kj} + \tilde{\Phi}_{k'j'} \neq \tilde{\Phi}_{k'j} + \tilde{\Phi}_{kj'}\ :$$

separability is not preserved by aggregation. This affects both of our proposed estimators; Section 4.2 proposes a general strategy to deal with it.

4.1 Zero Cells

The minimum-distance estimator relies on the equation

$$\forall (x, y),\ \Phi_{xy}^\beta + \frac{\partial \mathcal{E}^\alpha}{\partial \mu_{xy}}(\mu, \boldsymbol{n}, \boldsymbol{m}) = 0,$$

which is the first-order condition of the surplus maximization in the population under Assumptions 1, 2, and 3. Assumption 3 (continuous full support) may fail for two reasons: some distributions $\mathbb{P}_x$ or $\mathbb{Q}_y$ may have bounded support, and/or they may have mass points. In both cases, some matching

patterns may be zero in the population, and a fortiori in the sample. This requires allowing for inequalities in the identifying equations, and replacing derivatives with subgradients at mass points.

The most relevant case for applications arises because of sampling variation. Even if Assumptions 1, 2, and 3 hold and all population matching patterns are positive, some may be so small that the corresponding cells are empty in the sample. In such cases we would have $\mu_{xy} > 0$ but $\hat{\mu}_{xy} = 0$.

First note that it may not be a problem at all: if there exists $\boldsymbol{\lambda} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ that solves

$$\Phi_{xy}^{\beta} + \frac{\partial \mathcal{E}^{\alpha}}{\partial \mu_{xy}}(\hat{\boldsymbol{\mu}}; \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}) = 0 \text{ for all } x, y$$

then this $\boldsymbol{\lambda}$ is a perfectly reasonable estimator. Still, for many specifications, like those we studied in Section 2.3, the partial derivatives of the generalized entropy are infinite in zero. Then such a $\boldsymbol{\lambda}$ cannot exist.

Perhaps the simplest solution then is to add a small number $\delta > 0$ to each component of $\hat{\boldsymbol{\mu}}$. This can be done without modifying the number of households in the sample, using:

$$\hat{\mu}_a^{\delta} = (\hat{\mu}_a + \delta) \frac{N}{N + \delta |A|}$$

where $a \in A$ goes over all components. If we fix δ and the sample size N grows to infinity, the number of households in any cell grows like N. As a consequence, the estimators $(\boldsymbol{\alpha}_N^{\delta}, \boldsymbol{\beta}_N^{\delta})$ that result from applying minimum-distance estimation to the pseudo-data $\hat{\boldsymbol{\mu}}^{\delta}$ will be consistent.

More generally, we could replace $\hat{\mu}_a$ with

$$\tilde{\mu}_a = k_N \hat{\mu}_a + l_N$$

for positive numbers k_N, l_N such that $k_N \to 1$ and $l_N = o(N)$. One such formula is

$$\tilde{\mu}_a = N \frac{\hat{\mu}_{xy} + 1/2}{N + 1/2},$$

which in principle helps correct finite-sample bias (see Appendix B).

Note that the Poisson-GLM estimator of the Choo and Siow linear model that we presented in Section 3 maximizes a function that is *linear* in the observed matching patterns $\hat{\boldsymbol{\mu}}$; as such, it applies without modification when some cells are empty.

4.2 Aggregation

Now suppose that we have a data-generating process for which all elements of both μ and $\hat{\mu}$ are positive at the (x, y) level. If we had the data, we could use our minimum-distance estimator, or,

for the linear Choo and Siow model, the Poisson-GLM estimator. Unfortunately, we only observe aggregate (a, b) categories. Let us denote "$x \in a$" and "$y \in b$" to mean that x belongs to the aggregate a type, and y to the aggregate b type. The unobserved disaggregated matching patterns must add up to the observed aggregates: for all a and b,

$$\begin{aligned} \hat{\mu}_{ab} &= \sum_{\substack{x \in a \\ y \in b}} \tilde{\mu}_{xy} \\ \hat{\mu}_{a0} &= \sum_{x \in a} \tilde{\mu}_{x0} \\ \hat{\mu}_{0b} &= \sum_{y \in b} \tilde{\mu}_{0y}. \end{aligned} \tag{4.1}$$

On the other hand, the minimum distance estimator readily extends to such incomplete data. To see this, recall that with complete data we would use

$$\forall (x, y),\ \Phi^{\boldsymbol{\beta}}_{xy} + \frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{m}}) = 0$$

to estimate $\boldsymbol{\lambda} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$. Our proposal here is to augment the set of parameters $\boldsymbol{\lambda}$ with new positive parameters $\tilde{\mu}_{xy}, \tilde{\mu}_{x0}, \tilde{\mu}_{0y}$. That is, we replace $\boldsymbol{\lambda}$ with

$$\boldsymbol{\Lambda} = (\boldsymbol{\lambda}, (\tilde{\mu}_{xy}), (\tilde{\mu}_{x0}, (\tilde{\mu}_{0y}))$$

and we choose $\boldsymbol{\Lambda}$ to minimize the norm of the vector $\boldsymbol{D}^{\boldsymbol{\Lambda}}$ under the constraints that the $\tilde{\mu}_{xy}, \tilde{\mu}_{x0}, \tilde{\mu}_{0y}$ (i) be positive and (ii) satisfy the system of equalities (4.1). This can be seen as a minimum-distance estimator for the augmented mixed hypothesis

$$\exists \boldsymbol{\Lambda} = (\boldsymbol{\lambda}, \tilde{\boldsymbol{\mu}} > 0) \ \text{ s.t. } \boldsymbol{D}^{\boldsymbol{\Lambda}} = \mathbf{0} \ \text{ and (4.1)}.$$

While the constraints in (4.1) are linear in the new parameters, the partial derivatives of the generalized entropy now depend in a non-linear way on the additional parameters $\tilde{\boldsymbol{\mu}}$. This breaks the linearity of the procedure, even if $\boldsymbol{\Phi}$ is linear in $\boldsymbol{\beta}$ and the generalizes entropy is linear in $\boldsymbol{\alpha}$. Still, it is a simple constrained optimization problem that can be solved using off-the-shelf software.

A similar remark applies to the Poisson estimator of the linear Choo and Siow model: the optimization now runs over both $\boldsymbol{\gamma}$ and the $\tilde{\boldsymbol{\mu}}$ auxiliary variables under the constraints in (4.1).

5 Extensions

While we described our two estimation methods in the setting of a bipartite model, they can be used in a broader class of matching models with perfectly transferable utility.

5.1 Multipartite Matching

Suppose that each match must consist of at most p partners; the partner in position $k = 1, \dots, p$ (if any) must be drawn from sub-population k. The separability assumption extends naturally: for a match of individuals $i_1 \in x_1, \dots, i_p \in x_p$, we assume that

$$\tilde{\Phi}_{i_1 \dots i_p} = \Phi_{x_1 \dots x_p} + \sum_{l=1}^{p} \varepsilon_{i_l ; x_{-l}}$$

where we denote x_{-l} the characteristics of all partners except i_l. Then we can define e.g.

$$G_{x_l}(U_{x_l \cdot}) = E_{\mathbb{P}_{x_l}} \max_{x_{-l}} \left(U(x_l, x_{-l}) + \varepsilon_{i_l ; x_{-l}} \right),$$

the associated Legendre-Fenchel transforms, and the generalized entropy

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{q}) = - \sum_{l=1}^{K} \sum_{x_l \in \mathcal{X}_l} n_{x_l} G^*_{x_l} \left(\frac{\boldsymbol{\mu}_{x_l \cdot}}{n_{x_l}} \right)$$

given sub-population numbers $\boldsymbol{q} = (n_{x_1}, \dots, n_{x_p})$ and matching patterns $\mu_{x_1 \dots x_p}$.

A minimum-distance estimator obtains easily from these definitions; the algorithms in Boxes 1, 2 and 3 apply with obvious changes. For the linear Choo and Siow model, the Poisson-GLM estimator also only requires simple changes. In both cases, the main change is that with $p > 2$, other configurations than complete partner sets and singles may exist[9]. We give detailed formulae for the case $p = 3$ in Appendix D.

5.2 Many-to-one Matching

The pioneering contribution of Kelso and Crawford (1982) had firms matching with workers. While a given worker may only match with one firm, a firm may hire many workers. We now turn to a separable, transferable utility version of their model.

First assume that all workers occupy interchangeable functions within the firm. Each firm has an observed type x and each worker has an observed type y. Consider a firm i of type x that employs a set of workers S (a *team*). Denote n_y^S the number of workers in S who have observed type y, and $\boldsymbol{n}^S$ the vector $(n_1^S, \dots, n_Y^S)$. We call $\boldsymbol{n}^S$ the *type profile* of the team S. Then we define the joint surplus as

$$\tilde{\Phi}(i, S) = \Phi(x, \boldsymbol{n}^S) + \varepsilon_i(\boldsymbol{n}^S) + \sum_{j \in S} \eta_j(x).$$

[9] E.g. with $p = 3$, positions 1 and 3 may be filled in a match but not position 2. Such a partial match may still generate a positive joint surplus.

Here $\varepsilon_i(\boldsymbol{n}^S)$ could represent the unobserved shock to the production of this team in firm i; and $-\eta_j(x)$ could be the unobserved shock to the disutility of work of worker j in a firm of characteristics x[10].

Denote v_j the equilibrium payoff of a worker j. A firm $i \in x$ solves

$$u_i = \max_S \left(\Phi(x, \boldsymbol{n}_S) + \varepsilon_i(\boldsymbol{n}_S) + \sum_{j \in S} \eta_j(x) - \sum_{j \in S} v_j \right).$$

Using similar reasoning as in Section 1, this can be rewritten as

$$\begin{aligned} u_i &= \max_{\boldsymbol{n}_S} \left(\Phi(x, \boldsymbol{n}^S) + \varepsilon_i(\boldsymbol{n}^S) - \min_{T | \boldsymbol{n}^T = \boldsymbol{n}^S} \sum_{j \in T} (v_j - \eta_j(x)) \right) \\ &= \max_{\boldsymbol{n}_S} \left(\Phi(x, \boldsymbol{n}^S) + \varepsilon_i(\boldsymbol{n}^S) - \sum_{y=1}^{Y} n_y^S \min_{T_y} \frac{1}{n_y^S} \sum_{j \in T_y} (v_j - \eta_j(x)) \right) \end{aligned}$$

where for each y, the "minimum over T_y" ranges over all sets of n_y^S workers with type y. The value of this minimum is a function of x, y, and $\boldsymbol{n}^S$. In our large market approximation, it does not depend on n_y^S; we denote it by $V(x, y)$. It will represent the mean utility of type y of workers in firms of type x. We obtain

$$u_i = \max_{\boldsymbol{n} \in \mathbb{N}^Y} \left(\Phi(x, \boldsymbol{n}) - \sum_{y=1}^{Y} n_y V(x, y) + \varepsilon_i(\boldsymbol{n}) \right)$$

where the maximum ranges over all integer-valued vectors $\boldsymbol{n}$ in $\mathbb{N}^Y$.

We generate in the same way the utility of a worker j of observed type y who may join a firm i with a set S of other workers. Let us denote $\boldsymbol{n} + \mathbf{1}_y$ the type profile obtained by adding a worker of type y to an existing type profile $\boldsymbol{n}$. We have

$$\begin{aligned} v_j &= \max_{i,S} \left(\Phi(x, \boldsymbol{n}^S + \mathbf{1}_y) + \varepsilon_i(\boldsymbol{n}^S + \mathbf{1}_y) + \sum_{k \in S} \eta_k(x) + \eta_j(x) - u_i - \sum_{k \in S} v_k \right) \\ &= \max_{x, \boldsymbol{n} \in \mathbb{N}^Y} \left(\Phi(x, \boldsymbol{n} + \mathbf{1}_y) + \eta_j(x) - \min_{i \in x} (u_i - \varepsilon_i(\boldsymbol{n} + \mathbf{1}_y)) - \min_{T | \boldsymbol{n}^T = \boldsymbol{n}} \sum_{k \in T} (v_k - \eta_k(x)) \right) \\ &= \max_{\substack{x, \boldsymbol{n} \in \mathbb{N}^Y \\ n_y \geq 1}} \left(\Phi(x, \boldsymbol{n}) + \eta_j(x) - U(x, \boldsymbol{n}) - \sum_{t=1}^{Y} n_t V(x, t) \right), \end{aligned}$$

after defining $U(x, \boldsymbol{n}) = \min_{i \in x} (u_i - \varepsilon_i(\boldsymbol{n}))$.

To summarize: there exist functions $U(x, \boldsymbol{n})$ and $V(x, y)$ such that

$$U(x, \boldsymbol{n}) + \sum_{y=1}^{Y} n_y V(x, y) \geq \Phi(x, \boldsymbol{n})$$

[10] This disutility could also vary with the type profile of the team.

with equality if some firm of type x employs a team $\boldsymbol{n}$ of workers. In a match of a firm $i \in x$ with a team of type profile $\boldsymbol{n}$, the firm has a payoff

$$u_i = U(x, \boldsymbol{n}) + \varepsilon_i(\boldsymbol{n})$$

and a worker j of type y in this firm obtains

$$v_j = V(x, y) + \eta_j(x).$$

We can define the expected maximum utility of a firm of type x as

$$G_x(\boldsymbol{U}) = E_{\boldsymbol{\varepsilon}} \max_{\boldsymbol{n} \in \mathbb{N}^Y} (U(x, \boldsymbol{n}) + \varepsilon_i(\boldsymbol{n}))$$

and that of a worker of type y as

$$H_y(\boldsymbol{V}) = E_{\boldsymbol{\eta}} \max_{x \in \mathcal{X}} (V(x, y) + \eta_j(x)) .$$

From the functions G_x and H_y we can derive the Legendre-Fenchel transforms G_x^* and H_y^* in the standard manner. Note that the G_x^* is defined over the set of probabilities $\mu(\boldsymbol{n}|x)$, while H_y^* is defined over probabilities $\mu(y|x)$. These probabilities are linked, however. Since a firm with a team of type profile $\boldsymbol{n}$ employs n_y workers, the number of workers of type y employed by all firms of type x must be

$$\mu(x, y) = \sum_{\boldsymbol{n} \in \mathbb{N}^Y} n_y \mu(x, \boldsymbol{n}).$$

This allows us to redefine H_y^* as a function $\bar{H}_y$ of the probabilities $\mu(x, \boldsymbol{n})$.

Finally, given a sample with N_x firms of type x and M_y workers of type y, we define the generalized entropy as

$$\mathcal{E}(\boldsymbol{\mu}; \boldsymbol{N}, \boldsymbol{M}) = -\sum_{x=1}^{X} N_x G_x^*(\mu(\cdot|x)) - \sum_{y=1}^{Y} M_y \bar{H}_y(\mu).$$

for a collection of matching patterns $\mu = \mu(x, \boldsymbol{n})$. It is easy to check that

$$\Phi(x, \boldsymbol{n}) = -\frac{\partial \mathcal{E}}{\partial \mu(x, \boldsymbol{n})}(\boldsymbol{\mu}; \boldsymbol{N}, \boldsymbol{M}). \tag{5.1}$$

The observed matching patterns are $\hat{\mu}(x, \boldsymbol{n})$, the number of firms of type x which hire a team with type profile $\boldsymbol{n}$. Our minimum distance estimator can be directly applied to (5.1). The only difficulty is that with so many possible values of the type profile $\boldsymbol{n}$, many observed matching cells will be zero; then the adjustment suggested in Section 4.1 should be used. Corblet (2022) proposed such a many-to-one variant of the Choo and Siow model to examine how workers in different age and education groups sort across firms in different industries on the Portuguese labor market.

One could go further and distinguish the tasks that different workers can accomplish within a firm. The type profile would then become a type-task profile: n_{yt}^S would represent the number of workers of type y allocated to a task t within a team S. While the notation becomes more cumbersome, the extension follows naturally.

5.3 Unipartite Matching

Unipartite matching (also called *roommate matching*) is the instance of the one-to-one matching problem in which both partners can be taken from the same population. This increases greatly the number of possible blocking coalitions; as a result, a stable unipartite matching may fail to exist. Chiappori, Galichon, et al. (2019) showed that this difficulty vanishes asymptotically in large markets. Moreover, they show how a virtual bipartite matching problem can be associated to any large unipartite matching instance. In the case of continuous observable characteristics, Ciscato et al. (2020) leveraged that connection to build estimators can be used directly in the original unipartite problem.

With discrete observable characteristics, a separable unipartite matching problem is given by a set of types $\mathcal{X}$; margins n_x; and the following joint surplus:

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

where

- the matrix $\mathbf{\Phi}$ is *symmetric*: $\Phi_{xy} = \Phi_{yx}$ for all $x, y \in \mathcal{X}$;
- for each $i \in x$, the vector $(\varepsilon_{iy})_{y \in \mathcal{X}_0}$ is distributed according to $\mathcal{P}_x$.

Note the two differences with Assumption 1: since the partners are drawn from the same population, both the deterministic and the stochastic part of the joint surplus must be symmetric. In addition, the matching patterns $\boldsymbol{\mu}$ must also be symmetric: if we define μ_{xy} to be the number of matches with one partner of type x and one of type y, then

$$\mu_{xy} = \mu_{yx} \quad \text{for all} \quad x, y \in \mathcal{X}.$$

The margin equation expressing that the mass of agents of type x equals n_x then becomes

$$2\mu_{xx} + \sum_{y \neq x} \mu_{xy} \leq n_x$$

where the factor 2 in front of μ_{xx} is due to the fact that in a xx pair, there are two agents of type x.

Theorem 1 takes a slightly different form: there exists a (not necessarily symmetric) matrix $\boldsymbol{U}$ such that an agent $i \in x$ matches with an agent of type y (possibly 0) that maximizes $U_{xy} + \varepsilon_y^i$ over $\mathcal{X}_0$; and

$$U_{xy} + U_{yx} \leq \Phi_{xy},$$

with equality if there exist matches between types x and y.

We can define the Emax functions G_x and their Legendre-Fenchel transforms G_x^* exactly as before:

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

and for $\sum_{y \in \mathcal{X}_0} \mu_{y|x} = 1$,

$$G_x^*(\boldsymbol{\mu}_{\cdot|x}) = \max_{\boldsymbol{U}_{x\cdot}} \left(\sum_{y \in \mathcal{X}_0} \mu_{y|x} U_{xy} - G_x(\boldsymbol{U}_{x\cdot}) \right).$$

By the envelope theorem, this identifies U_{xy} as

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

which can be used as the basis for a minimum-distance estimator.

It is easy to see that the stable matching maximizes the total joint surplus

$$\mathcal{W}(\boldsymbol{\mu}) = \sum_x \mu_{xx} \Phi_{xx} + \frac{1}{2} \sum_{x \neq y} \mu_{xy} \Phi_{xy} + \mathcal{E}(\boldsymbol{\mu}; \boldsymbol{n}), \tag{5.2}$$

where the generalized entropy now takes the following form:

$$\mathcal{E}(\boldsymbol{\mu}; \boldsymbol{n}) = -\sum_{x=1}^{X} n_x G_x^*(\mu_{x\cdot}/n_x).$$

Theorem 2 still holds and (since $\mu_{yx} = \mu_{xy}$)

$$\Phi_{xy} = -\frac{\partial \mathcal{E}}{\partial \mu_{xy}}(\boldsymbol{\mu}; \boldsymbol{n}).$$

Finally, our Poisson estimator adapts readily for the special case of the linear Choo and Siow model. We give the formulæ in Appendix E.

6 Monte Carlo Simulation

We coded these two estimation methods in a `Python` package called `cupid_matching` that is available from the standard repositories[11].

In order to test the quality of the estimators on a realistic example, we generated data from a simplified version of the Choo and Siow model that we estimated in Galichon and Salanié, 2022, Section 7. More precisely, we first estimated an eight-parameter semilinear Choo and Siow model on the same dataset. Our model has $X = Y = 25$ and the following $K = 8$ basis functions[12]:

$$1, x, y, x^2, xy, y^2, \mathbb{1}(x \geq y), \max(x - y, 0).$$

The first six basis functions generate a quadratic expansion, while the last two allow the surplus to differ according to whether the husband or the wife is older.

The original dataset combines two different sources: the 1970 Census for the as-yet-unmarried men and women ("available"), and the 1971-72 Vital Statistics for 216,428 observed marriages. Choo and Siow (2006) and Galichon and Salanié (2022) used sampling weights when estimating their models; the population has about 8 times more availables than marriages. For our simulations, we maintain this proportion and we create a "small sample" of 260,000 observations and a "large sample" of 1.7m observations. The large sample has the original number of marriages and rescaled availables; the small sample has the original number of availables and a only about 30,000 marriages.

We fit the Choo and Siow model with the eight basis functions on both samples; we then use the estimated coefficients to generate $1,000$ new samples. Finally, we compute both the Poisson estimator and the minimum-distance estimator on each generated sample.

As several of the cells are empty in the original data and in our generated samples, we used the adjustment of Section 4.1; we took δ to be the size of the smallest positive cell, which happens to be one. We applied this adjustment both to the original dataset and to the generated samples. We found that the precise value of δ does not matter much for the quality of the estimates[13].

To make the figures more readable, we normalized the basis functions so that the estimates in the original dataset equal 1. Their estimated standard errors are in Table 1.

[11] See https://pypi.org/project/cupid-matching/.

[12] We applied a preliminary quantile transform to the margin vectors $\boldsymbol{n}$ and $\boldsymbol{m}$ and we used Legendre polynomials on $[0, 1]$.

[13] The bias-correction mentioned in Appendix B turned out to be numerically unstable: with many thousands of observations, the derivative of the logarithm becomes very large in empty cells.

Base	Estimate	Standard Error	
		Small sample	Large sample
1	1	0.007	0.003
$\mathbb{1}(x > y)$	1	0.024	0.010
$\max(x - y, 0)$	1	0.018	0.007
x	1	0.024	0.009
y	1	0.015	0.006
x^2	1	0.034	0.013
xy	1	0.070	0.026
y^2	1	0.126	0.044

Table 1: Estimates on the Original Datasets

Figure 1 plots the distribution of the minimum distance and Poisson estimates on the smaller sample. Each of the eight panels corresponds to the coefficient of a base function. The blue "Expected" curve plots the distribution of $1,000$ draws from a normal distribution centered at the true value of 1, with a standard error equal to the estimated standard error in Table 1. Note that the horizontal axes have very different scales.

It is clear from Figure 1 that the minimum distance and Poisson estimators have very similar distributions, with standard deviations that are in the same ballpark as the standard errors estimated on the original dataset. We checked that in most cases, the values of the two estimators differ by less than 0.1. The biases go from a couple of percentage points for the top panels to a more sizable 50% in the bottom panels. Tests for equality to the true value of the coefficient of the quadratic terms, for instance, would reject much too often.

The dominant term in finite-sample bias is quite generally in $1/n$, so that one would expect it to be about six times smaller with the larger sample. This is essentially what Figure 2 shows: the biases are all smaller than 10% and testing procedures would be less misleading. Still, a 5% test that the coefficient of y^2 equals its true value would reject the null close to half of the time.

It is important to emphasize here that by construction, the minimum distance estimator exhausts all of the empirical content of the model; so does the Poisson GLM for the Choo and Siow model. It seems unlikely that alternative estimators would perform much better. In particular, we checked

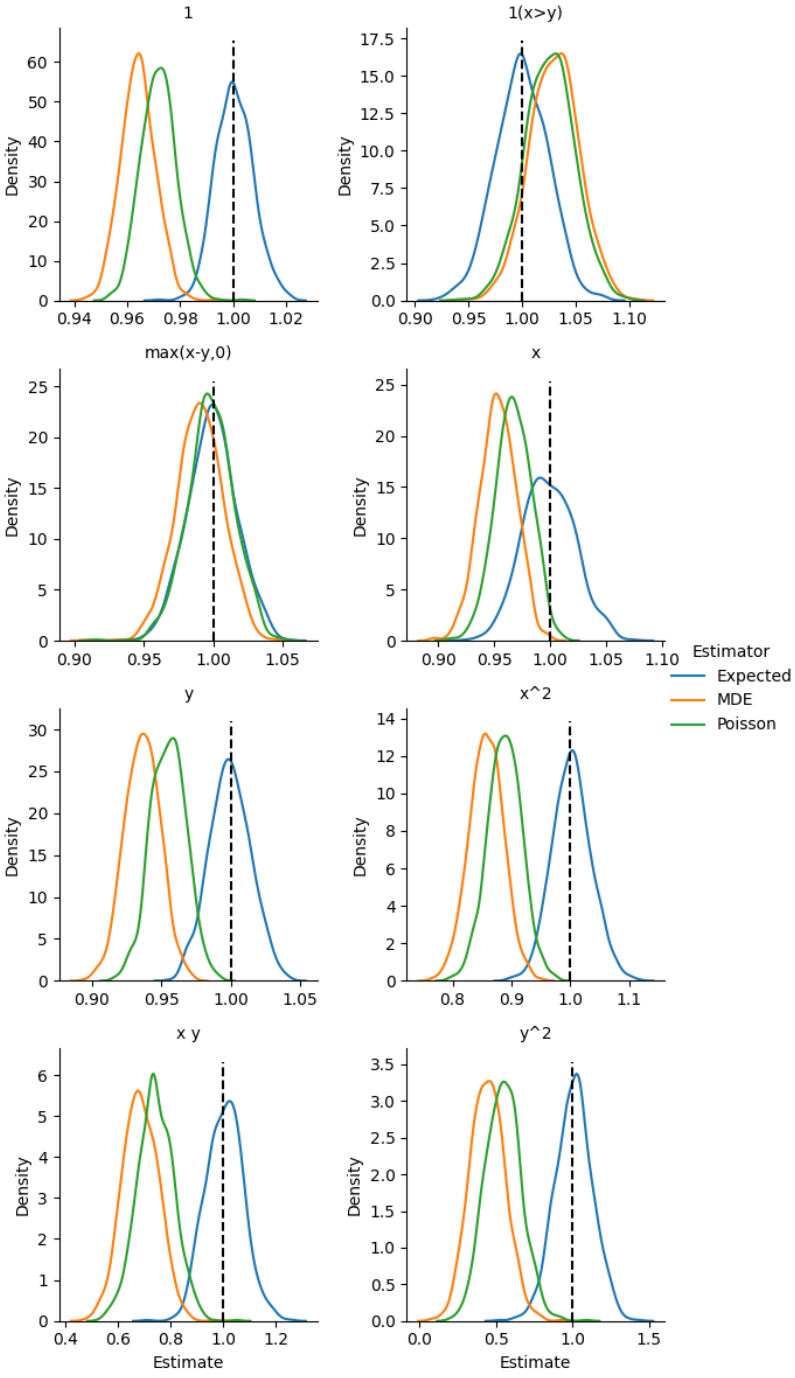

Figure 1: Estimating the Choo and Siow Model (small sample)

that the nonparametric estimator used by Choo and Siow (2006)

$$\hat{\Phi}_{xy} = \log \frac{\hat{\mu}_{xy}^2}{\hat{\mu}_{x0}\hat{\mu}_{0y}}$$

(which coincides with our estimator when the basis functions are the $X \times Y$ cell indicators) has non-negligible bias in moderate-size samples.

Even in relative large samples where cells with no match may be rare, the range of variation of cell counts can be very large. To illustrate, the dataset we used in Galichon and Salanié (2022) has $216,428$ marriages. With $X = Y = 25$, there are 625 marriage cells. Table 2 show some quantiles of the 625 elements of the matrix $(\boldsymbol{\mu}_{xy})$ in this dataset. While only 12 of the 625 cells are empty, the P90-P10 ratio is larger than 100.

Table 2: Quantiles of Marriage Numbers in the Original Dataset

Percentile	Value
1%	0
5%	3
10%	7
25%	27
50%	79
75%	179
90%	704
95%	2,077
99%	5,095

This huge dispersion of cell counts appears to be the reason why it takes so many observations to get reliable estimates of the parameters in the Choo and Siow data. To prove this, we simulated 1,000 samples from a data-generating process that has the same (estimated) parameter values, but with basis functions uniformly divided by 5. This simulates a marriage market where the dispersion in the matching patterns across cells is roughly ten times smaller as in the original dataset. Figure 3 shows that in this "shrunk" marriage market, both of our estimators perform very well: the biases vanished and the dispersion is very close to what one would expect from the asymptotic formula.

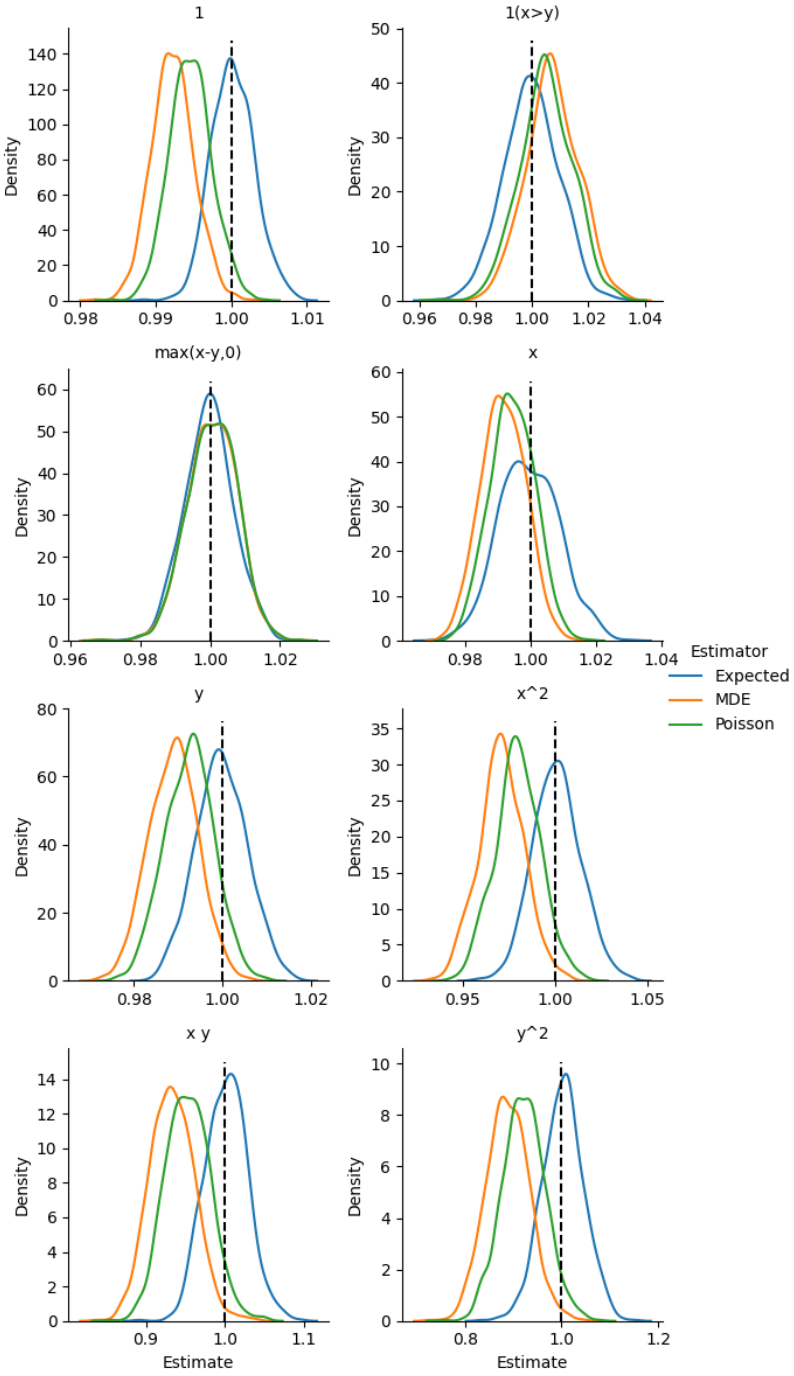

Figure 2: Estimating the Choo and Siow Model (large sample)

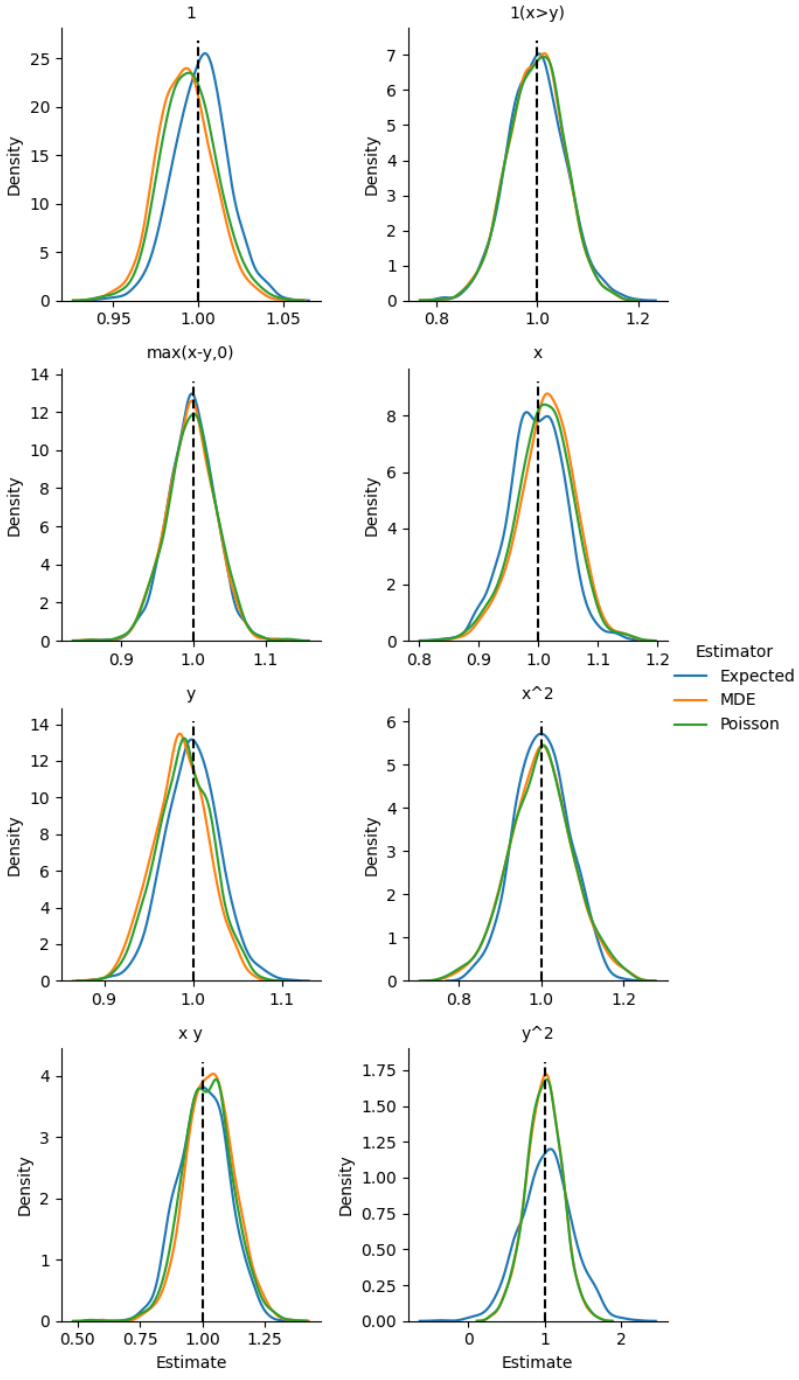

Figure 3: Estimating the "shrunk" Choo and Siow Model (small sample)

Concluding remarks

Given a fully specified model, the maximum likelihood estimator can always be used. It requires numerical optimization and, in an "inside loop", the computation of the stable matching for each current value of the parameter vector. This may be quite costly when fast algorithms (such as IPFP) cannot be used. It also requires much more coding than the procedures we present in this paper.

Each of the two methods we presented has its pros and cons. The minimum-distance estimator applies to all separable models; it is most convenient in semilinear models. To achieve maximum efficiency, and to test the specification, one needs to evaluate the second derivatives of the entropy with respect to the matching patterns. This may be difficult. In addition, when the data contains zero cells it needs to be adjusted as explained in Section 4.1. The Poisson regression estimator only applies to semilinear Choo and Siow (2006) models. It is appealing in its simplicity of use, as one can rely on standard statistical packages. It is also robust to zero cells.

Our simulations suggest that even in "simple" models such as the Choo and Siow (2006), it may take large sample sizes to get reliable estimates. Analysts should be aware that the dispersion in matching patterns seems to be a crucial determinant of the performance of the estimators for a given sample size.

In labor markets or in marriage markets, large samples are readily available. When they are not (as with matching between firms), finite-sample bias may be a concern. The alternative is to develop a bias-correction procedure that is more powerful than the rather elementary one we presented.

References

Berry, S. (1994). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics*, *25*, 242–262.

Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*, 841–890.

Berry, S., & Pakes, A. (2007). The pure characteristics demand model. *International Economic Review*, *48*, 1193–1225.

Chiappori, P.-A., Nguyen, D. L., & Salanié, B. (2019). *Matching with random components: Simulations* [Columbia University mimeo].

Chiappori, P.-A., Salanié, B., & Weiss, Y. (2017). Partner choice, investment in children, and the marital college premium. *American Economic Review*, *107*, 2109–67.

Chiappori, P.-A., Galichon, A., & Salanié, B. (2019). On human capital and team stability. *Journal of Human Capital*, *13*, 236–259.

Choo, E., & Siow, A. (2006). Who marries whom and why. *Journal of Political Economy*, *114*, 175–201.

Ciscato, E., Galichon, A., & Goussé, M. (2020). Like attract like? a structural comparison of homogamy across same-sex and different-sex households. *Journal of Political Economy*, *128*(2), 740–781.

Corblet, P. (2022). *Education expansion, sorting, and the decreasing education wage premium* [Sciences Po Paris mimeo].

Fox, J., Yang, C., & Hsu, D. (2018). Unobserved heterogeneity in matching games with an apppplication to venture capital. *Journal of Political Economy*, *126*, 1339–1373.

Galichon, A., & Salanié, B. (2022). Cupid's invisible hand: Social surplus and identification in matching models. *Review of Economic Studies*, *89*, 2600–2629.

Kelso, A. S., & Crawford, V. P. (1982). Job matching, coalition formation, and gross substitutes. *Econometrica*, *50*, 1483–1504.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Palmgren, J. (1981). Fisher information matrix for log linear models arguing conditionally on observed explanatory variables. *Biometrika*, *68*, 563–6.

Santos Silva, J. M. C., & Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics*, *88*, 641–658.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.

A IPFP for the Nested Logit

Let us consider a nested logit model in which the nests do not depend on the type ($\mathcal{N}_x \equiv \mathcal{N}$ and $\mathcal{N}'_y \equiv \mathcal{N}'$) and their parameters $\boldsymbol{\rho}$ and $\boldsymbol{\delta}$ only depend on the nest: $\rho^x_n \equiv \rho_n$ and $\delta^y_{n'} \equiv \delta_{n'}$. Equation (2.8) can be rewritten as follows, for $y \in n$ and $x \in n'$:

$$\mu_{xy}^{\rho_n+\delta_{n'}} = \exp(\Phi_{xy})\mu_{x0}\mu_{0y}\mu_{xn}^{\rho_n-1}\mu_{n'y}^{\delta_{n'}-1}. \tag{A.1}$$

Since $\mu_{xn} = \sum_{y\in n}\mu_{xy}$, we get

$$\mu_{xn} = \mu_{x0}^{1/(\rho_n+\delta_{n'})}\mu_{xn}^{(\rho_n-1)/(\rho_n+\delta_{n'})}\sum_{y\in n}\exp\left(\Phi_{xy}/(\rho_n+\delta_{n'})\right)\mu_{0y}^{1/(\rho_n+\delta_{n'})}\mu_{n'y}^{(\delta_{n'}-1)/(\rho_n+\delta_{n'})},$$

and, denoting $K_{xy} = \exp\left(\Phi_{xy}/(\rho_n+\delta_{n'})\right)$:

$$\mu_{xn}^{(\delta_{n'}+1)/(\rho_n+\delta_{n'})} = \mu_{x0}^{1/(\rho_n+\delta_{n'})}\sum_{y\in n}K_{xy}\mu_{0y}^{1/(\rho_n+\delta_{n'})}\mu_{n'y}^{(\delta_{n'}-1)/(\rho_n+\delta_{n'})}. \tag{A.2}$$

Substituting in the adding up constraint $\mu_{x0} + \sum_{y=1}^{Y}\mu_{xy} = n_x$ gives

$$\begin{aligned} n_x &= \mu_{x0} + \sum_{n\in\mathcal{N}}\mu_{xn} \\ &= \mu_{x0} + \sum_{n\in\mathcal{N}}\mu_{x0}^{1/(\delta_{n'}+1)}\left(\sum_{y\in n}K_{xy}\mu_{0y}^{1/(\rho_n+\delta_{n'})}\mu_{n'y}^{(\delta_{n'}-1)/(\rho_n+\delta_{n'})}\right)^{(\rho_n+\delta_{n'})/(\delta_{n'}+1)}. \end{aligned} \tag{A.3}$$

For given values of $(\mu_{0y}, \mu_{n'y})$ for all y, (A.3) defines μ_{x0} uniquely[14] . Once μ_{x0} is known, we can plug it in (A.2) to obtain the values of μ_{xn} for all n. We do this for all values of x.

Then we can apply similar equations to the y side:

$$\mu_{n'y}^{(\rho_n+1)/(\rho_n+\delta_{n'})} = \mu_{0y}^{1/(\rho_n+\delta_{n'})}\sum_{x\in n'}K_{xy}\mu_{x0}^{1/(\rho_n+\delta_{n'})}\mu_{xn}^{(\rho_n-1)/(\rho_n+\delta_{n'})}$$

$$m_y = \mu_{0y} + \sum_{n'\in\mathcal{N}'}\mu_{0y}^{1/(\rho_n+1)}\left(\sum_{x\in n'}K_{xy}\mu_{x0}^{1/(\rho_n+\delta_{n'})}\mu_{xn}^{(\rho_n-1)/(\rho_n+\delta_{n'})}\right)^{(\rho_n+\delta_{n'})/(\rho_n+1)}$$

to solve for μ_{0y} and $\mu_{n'y}$ given the values of (μ_{x0}, μ_{xn}) for all x. We iterate until convergence and we use (A.1) to compute the matching patterns μ_{xy}.

B Bias Correction for the Minimum Distance Estimator

As explained in Section 6, the range of variation of cell counts can be very large in applications. This dispersion may be an issue with the minimum distance estimator, as the derivative of the generalized

[14] Since $\delta_{n'} \geq 0$, the right-hand side is an increasing function of μ_{x0} whose values go from zero to infinity.

entropy often is highly nonlinear for small cells. To illustrate this, remember that in the Choo and Siow model, this derivative equals

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

If types x and y have a low propensity to marry each other, the observed values of both fractions will be very small. Since the logarithm function is very concave at zero, the logarithm $\hat{r}$ of the ratio $\hat{\mu}_{xy}/\hat{\mu}_{x0}$, for instance, is likely to underestimate the logarithm r of the population ratio μ_{xy}/μ_{x0}.

A simple solution to this problem is to use second-order Taylor expansions. Asymptotically $\hat{\mu}_{xy}$ is distributed as a normal with mean μ_{xy} and variance $\mu_{xy}(1 - \mu_{xy}/N)$. For any function f with three bounded derivatives around μ_{xy},

$$Ef\left(\hat{\mu}_{xy}\right) \simeq f(\mu_{xy}) + \frac{1}{2} f''(\mu_{xy}) V\hat{\mu}_{xy} = f(\mu_{xy}) + \frac{1}{2} f''(\mu_{xy})\, \mu_{xy} \left(1 - \frac{\mu_{xy}}{N}\right).$$

Thus we may hope to get a better estimate of $f(\boldsymbol{\mu})$ with

$$f\left(\hat{\mu}_{xy}\right) - \frac{1}{2} f''(\hat{\mu}_{xy})\, \mu_{xy} \left(1 - \frac{\hat{\mu}_{xy}}{N}\right).$$

This correction could be applied to the derivative f of the generalized entropy.

In the Choo and Siow model and its variants, this won't work: if $f(\boldsymbol{\mu}) = \log \mu_{xy}$ for instance, the correction is infinite when $\hat{\mu}_{xy} = 0$. However, it is easy to see that $Ef(\hat{\mu}_{xy} + d_{xy})$ is also almost unbiased for $f(\mu_{xy})$ if we take

$$d_{xy} = -\frac{1}{2} \frac{f''(\hat{\mu}_{xy})}{f'(\hat{\mu}_{xy})}\, \hat{\mu}_{xy} \left(1 - \frac{\hat{\mu}_{xy}}{n_x}\right).$$

If f is the logarithm function, this gives the finite correction

$$d_{xy} = \frac{1}{2} \left(1 - \frac{\hat{\mu}_{xy}}{N}\right),$$

so that we only need to replace $\log \hat{\mu}_{xy}$ with

$$\log \left(\hat{\mu}_{xy} + \frac{1}{2} \left(1 - \frac{\hat{\mu}_{xy}}{N}\right)\right).$$

This also solves the "zero cell" problem as it is well-defined at $\hat{\mu}_{xy} = 0$. It is easy to check that this is equivalent to substituting the proportion $\hat{\mu}_{xy}/N$ with

$$\frac{\hat{\mu}_{xy} + 1/2}{N + 1/2},$$

a formula that is often used in discrete choice problems.

C Proofs

C.1 Proof of Proposition 3

Consider men of type x. Their Emax function under identical-scale unobserved heterogeneity is

$$\begin{aligned} G_x(\boldsymbol{U}_{x\cdot};\boldsymbol{\alpha}) &= E_{\mathbb{P}^0_x} \max_{y\in\mathcal{Y}_0} \left(U_{xy} + \boldsymbol{a}_{xy}\cdot\boldsymbol{\alpha} + \boldsymbol{b}_x\cdot\boldsymbol{\alpha}\,\varepsilon^0_{iy}\right) \\ &= (\boldsymbol{b}_x\cdot\boldsymbol{\alpha}) \times G^0_x\left((\boldsymbol{U}_{x\cdot} + \boldsymbol{a}_{x\cdot}\cdot\boldsymbol{\alpha})/(\boldsymbol{b}_x\cdot\boldsymbol{\alpha})\right) \end{aligned}$$

where we denote G^0_x the Emax function for the distribution $\varepsilon^0|x$, which is independent of $\boldsymbol{\alpha}$.

Now take the Legendre-Fenchel transform:

$$\begin{aligned} G^*_x(\boldsymbol{\mu}_{\cdot|x};\boldsymbol{\alpha}) &= \max_{\boldsymbol{U}_{x\cdot}} \left(\boldsymbol{\mu}_{\cdot|x}\boldsymbol{U}_{x\cdot} - G_x(\boldsymbol{U}_{x\cdot};\boldsymbol{\alpha})\right) \\ &= (\boldsymbol{b}_x\cdot\boldsymbol{\alpha}) \times \max_{\hat{\boldsymbol{U}}_{x\cdot}} \left(\boldsymbol{\mu}_{\cdot|x}\hat{\boldsymbol{U}}_{x\cdot} - G^0_x(\hat{\boldsymbol{U}}_{x\cdot})\right) - \sum_{y\in\mathcal{Y}_0} \mu_{y|x}\boldsymbol{a}_{xy}\boldsymbol{\alpha} \end{aligned}$$

where we used the change of variables

$$\hat{U}_{xy} = \frac{U_{xy} + \boldsymbol{a}_{xy}\cdot\boldsymbol{\alpha}}{\boldsymbol{b}_x\cdot\boldsymbol{\alpha}}.$$

The maximum in the second line is simply $(G^0)^*_x(\boldsymbol{\mu}_{\cdot|x}$, and is independent of $\boldsymbol{\alpha}$. Therefore

$$\frac{\partial G^*_x}{\partial \boldsymbol{\mu}_{\cdot|x}}(\boldsymbol{\mu}_{\cdot|x};\boldsymbol{\alpha}) = (\boldsymbol{b}_x\cdot\boldsymbol{\alpha}) \times \frac{\partial (G^0)^*_x}{\partial \boldsymbol{\mu}_{\cdot|x}}(\boldsymbol{\mu}_{\cdot|x}) - \boldsymbol{a}_{x\cdot}\boldsymbol{\alpha}$$

which is clearly linear in $\boldsymbol{\alpha}$. By (1.6), the derivatives of the generalized entropy are a linear combination of the derivatives of G^*_x and of H^*_y. Therefore they also are linear in $\boldsymbol{\alpha}$.

C.2 Proof of Theorem 4

Recall that

$$N = \sum_{x,y} \mu^{\boldsymbol{\beta}}_{xy} + \sum_x \mu^{\boldsymbol{\beta}}_{x0} + \sum_y \mu^{\boldsymbol{\beta}}_{0y}$$

is the total mass of households in the sample. For the Choo and Siow (2006) specification we know that at the stable matching $(\boldsymbol{\mu}, \boldsymbol{u}, \boldsymbol{v})$ for a joint surplus $\boldsymbol{\Phi}$ the following obtain:

$$\begin{aligned} \mu_{x0} &= \hat{n}_x \exp(-u_x) \\ \mu_{0y} &= \hat{m}_y \exp(-v_y) \\ \mu_{xy} &= \sqrt{\hat{n}_x \hat{m}_y}\exp((\Phi_{xy} - u_x - v_y)/2). \end{aligned} \tag{C.1}$$

Now consider the maximization of the following expression:

$$L(\boldsymbol{u},\boldsymbol{v},\boldsymbol{\beta}) = \sum_{x,y}\hat{\mu}_{xy}\boldsymbol{\phi}_{xy}\boldsymbol{\beta} - 2\sum_{x,y}\sqrt{\hat{n}_x\hat{m}_y}\exp((\boldsymbol{\phi}_{xy}\boldsymbol{\beta}-u_x-v_y)/2) \\ -\sum_x \hat{n}_x\exp(-u_x) - \sum_y \hat{m}_y\exp(-v_y) - \sum_x \hat{n}_x u_x - \sum_y \hat{m}_y v_y \tag{C.2}$$

over $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{\beta}$. The first order conditions with respect to u_x and v_y give

$$\sum_y \sqrt{\hat{n}_x\hat{m}_y}\exp((\boldsymbol{\phi}_{xy}\boldsymbol{\beta}-u_x-v_y)/2) + \hat{n}_x\exp(-u_x) = \hat{n}_x$$
$$\sum_x \sqrt{\hat{n}_x\hat{m}_y}\exp((\boldsymbol{\phi}_{xy}\boldsymbol{\beta}-u_x-v_y)/2) + \hat{m}_y\exp(-v_y) = \hat{m}_y.$$

Substituting (C.1), these are simply the margin equations

$$\sum_y \mu_{xy} + \mu_{x0} = n_x \tag{C.3}$$
$$\sum_x \mu_{xy} + \mu_{0y} = m_y. \tag{C.4}$$

The first order conditions with respect to β_k become

$$\sum_{xy}\hat{\mu}_{xy}\phi^k_{xy} = \sum_{xy}\mu_{xy}\phi^k_{xy}.$$

Therefore maximizing (C.2) gives the moment matching estimator $\hat{\boldsymbol{\beta}}$ and the associated stable matching $(\boldsymbol{u}^{\boldsymbol{\beta}},\boldsymbol{v}^{\boldsymbol{\beta}})$.

Now remember that given observations $(\hat{N}_a,\boldsymbol{Z}_a)_{a\in\mathcal{A}}$ weighted by a vector $\boldsymbol{w}$, the log-likelihood function of a Poisson count model with parameter $\exp(\boldsymbol{Z}_a'\boldsymbol{\gamma})$ is

$$l\left(\hat{\boldsymbol{\mu}},\boldsymbol{\gamma};\boldsymbol{w}\right) = \sum_{a\in\mathcal{A}} w_a\left(\hat{\mu}_a\boldsymbol{Z}_a'\boldsymbol{\gamma} - \exp\left(\boldsymbol{Z}_a'\boldsymbol{\gamma}\right) - \log(\hat{\mu}_a!)\right). \tag{C.5}$$

. Define $\boldsymbol{\gamma} = (\boldsymbol{\beta}',\boldsymbol{a}',\boldsymbol{b}')$ with

$$\boldsymbol{a} = \boldsymbol{u} - \log\hat{\boldsymbol{n}}, \quad \boldsymbol{b} = \boldsymbol{v} - \log\hat{\boldsymbol{m}}.$$

Then with $\boldsymbol{Z}$ and $\boldsymbol{w}$ defined in Theorem 4, and denoting $\boldsymbol{Z}_a$ the row a of the matrix $\boldsymbol{Z}$, we have

$$(\boldsymbol{Z}\boldsymbol{\gamma})_{xy} = (\boldsymbol{\phi}_{xy}\boldsymbol{\beta} - u_x + \log\hat{n}_x - v_y + \log\hat{m}_y)/2$$
$$(\boldsymbol{Z}\boldsymbol{\gamma})_{x} = -u_x + \log\hat{n}_x$$
$$(\boldsymbol{Z}\boldsymbol{\gamma})_{y} = -v_y + \log\hat{m}_y.$$

This proves that up to constant terms, the objective functions l and L are identical.

C.3 Proof of Theorem 5

The variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ follows directly from the fact that it maximizes (C.5), and hence is an M-estimator, see chapter 5 of van der Vaart (1998). The maximization of (C.5) gives first-order conditions

$$\sum_{a\in\mathcal{A}} w_a \exp(\boldsymbol{Z}_a\hat{\boldsymbol{\gamma}})\boldsymbol{Z}_a = \sum_{a\in\mathcal{A}} w_a\hat{\mu}_a\boldsymbol{Z}_a,$$

so that, applying the delta method, we get at first order

$$\left(\sum_{a\in\mathcal{A}} w_a \exp(\boldsymbol{Z}_a\hat{\boldsymbol{\gamma}})\boldsymbol{Z}_a'\boldsymbol{Z}_a\right)(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}) = \sum_{a\in\mathcal{A}} w_a\boldsymbol{Z}_a(\hat{\mu}_a-\mu_a).$$

so we obtain a consistent estimator of the variance of $\hat{\boldsymbol{\gamma}}$ as

$$\hat{V}\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{A}}^{-1}\,\hat{\boldsymbol{B}}\,\hat{\boldsymbol{A}}^{-1}$$

where

$$\hat{\boldsymbol{A}} = \sum_{a\in\mathcal{A}} w_a \exp(\boldsymbol{Z}_a\hat{\boldsymbol{\gamma}})\boldsymbol{Z}_a'\boldsymbol{Z}_a$$

and

$$\hat{\boldsymbol{B}} = \sum_{a,a'\in\mathcal{A}} w_a w_{a'}\mathrm{cov}(\hat{\mu}_a,\hat{\mu}_{a'})\boldsymbol{Z}_a'\boldsymbol{Z}_{a'}.$$

D Tripartite Matching

Suppose that each match at most one partner in each of $p=3$ sub-populations. We denote $i_1\in\mathcal{X}_1, i_2\in\mathcal{X}_2$, and $i_3\in\mathcal{X}_3$ the types, with corresponding observed types x_1,x_2,x_3. There may now be seven types of matches:

- complete matches, with a joint surplus

$$\tilde{\Phi}_{i_1i_2i_3} = \Phi_{x_1x_2x_3} + \varepsilon^1_{x_2x_3} + \varepsilon^2_{x_1x_3} + \varepsilon^3_{x_1x_2}$$

- three types of incomplete matches, for instance

$$\tilde{\Phi}_{i_10i_3} = \Phi_{x_10x_3} + \varepsilon^1_{0x_3} + \varepsilon^3_{x_10}$$

- and three types of singles, with for instance $\tilde{\Phi}_{i_100} = \varepsilon^1_{00}$.

The margin conditions have for instance

$$\sum_{x_2x_3} \mu_{x_1x_2x_3} + \sum_{x_2} \mu_{x_1x_20} + \sum_{x_3} \mu_{x_10x_3} + \mu_{x_100} = n^1_{x_1}.$$

Of course, some of these types of matches may be ruled out by the context.

If the distribution of $\boldsymbol{\varepsilon}^1$ is $\mathbb{P}^1_{x_1}$, we define

$$G^1_{x_1}(U^1_{x_1\cdot\cdot}) = E_{\mathbb{P}^1_{x_1}} \max\Bigg(\max_{x_2,x_3}(U^1_{x_1x_2x_3} + \varepsilon^1_{x_2x_3}), \\ \max_{x_2}(U_{x_1x_20} + \varepsilon^1_{x_20}), \max_{x_3}(U_{x_10x_3} + \varepsilon^1_{0x_3}), \varepsilon^1_{00}\Bigg)$$

and

$$(G^1)^*(\nu_{\cdot\cdot}) = \max_{\boldsymbol{U}} \left(\sum_{x_2x_3} \nu_{x_2x_3} U_{x_2x_3} + \sum_{x_2} \nu_{x_20} U_{x_20} + \sum_{x_3} \nu_{0x_3} U_{0x_3} - G^1(U) \right)$$

The generalized entropy is

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{n}) = -\sum_{x_1} n^1_{x_1} (G^1)^*(\mu_{\cdot\cdot|x_1}) - \sum_{x_2} n^2_{x_2} (G^2)^*(\mu_{\cdot\cdot|x_2}) - \sum_{x_3} n^3_{x_3} (G^3)^*(\mu_{\cdot\cdot|x_3})$$

and the joint surplus is identified by

$$\begin{aligned}
\Phi_{x_1x_2x_3} &= -\frac{\partial \mathcal{E}}{\partial \mu_{x_1x_2x_3}} \\
&= -\frac{\partial (G^1)^*}{\partial \mu_{x_2x_3|x_1}} - \frac{\partial (G^2)^*}{\partial \mu_{x_1x_3|x_2}} - \frac{\partial (G^3)^*}{\partial \mu_{x_1x_2|x_3}} \\
\Phi_{x_1x_20} &= -\frac{\partial \mathcal{E}}{\partial \mu_{x_1x_20}} \\
&= -\frac{\partial (G^1)^*}{\partial \mu_{x_20|x_1}} - \frac{\partial (G^2)^*}{\partial \mu_{x_10|x_2}} \\
\Phi_{x_100} &= 0
\end{aligned}$$

and the obvious permutations.

Applying the minimum-distance estimator to these equalities is straightforward. For the Poisson estimator, we need to change a few things as the set of possible match configurations is larger, and each match may have 1, 2, or 3 members.

The function L of Appendix C becomes

$$\begin{aligned}
L(\boldsymbol{u}^1, \boldsymbol{u}^2, \boldsymbol{u}^3, \boldsymbol{\beta}) = & \sum_{x_1 x_2 x_3} \hat{\mu}_{x_1 x_2 x_3} \boldsymbol{\phi}_{x_1 x_2 x_3} \boldsymbol{\beta} \\
& + \sum_{x_1 x_2} \hat{\mu}_{x_1 x_2 0} \boldsymbol{\phi}_{x_1 x_2 0} \boldsymbol{\beta} \sum_{x_1 x_3} \hat{\mu}_{x_1 0 x_3} \boldsymbol{\phi}_{x_1 0 x_3} \boldsymbol{\beta} \sum_{x_2 x_3} \hat{\mu}_{0 x_2 x_3} \boldsymbol{\phi}_{0 x_2 x_3} \boldsymbol{\beta} \\
& - 3 \sum_{x_1 x_2 x_3} \left(\hat{n}^1_{x_1} \hat{n}^2_{x_2} \hat{n}^3_{x_3} \right)^{1/3} \exp\left((\boldsymbol{\phi}_{x_1 x_2 x_3} \boldsymbol{\beta} - u^1_{x_1} - u^2_{x_2} - u^3_{x_3})/3 \right) \\
& - 2 \sum_{x_1 x_2} \sqrt{\hat{n}^1_{x_1} \hat{n}^2_{x_2}} \exp\left((\boldsymbol{\phi}_{x_1 x_2 0} \boldsymbol{\beta} - u^1_{x_1} - u^2_{x_2})/2 \right) - 2 \sum_{x_1 x_3} \sqrt{\hat{n}^1_{x_1} \hat{n}^3_{x_3}} \exp\left((\boldsymbol{\phi}_{x_1 0 x_3} \boldsymbol{\beta} - u^1_{x_1} - u^3_{x_3})/2 \right) \\
& - 2 \sum_{x_2 x_3} \sqrt{\hat{n}^2_{x_2} \hat{n}^3_{x_3}} \exp\left((\boldsymbol{\phi}_{0 x_2 x_3} \boldsymbol{\beta} - u^2_{x_2} - u^3_{x_3})/2 \right) - \sum_{k=1,2,3} \sum_{x_k} \hat{n}^k_{x_k} \left(\exp(-u^k_{x_k}) + u^k_{x_k} \right).
\end{aligned}$$

To identify it to the function $l(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma}; \boldsymbol{w})$, we define $\boldsymbol{\gamma} = (\boldsymbol{\beta}', \boldsymbol{a^{1}}', \boldsymbol{a^{2}}', \boldsymbol{a^{3}}')'$; the set of observations a is

$$\mathcal{A} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \cup \mathcal{X}_1 \times \mathcal{X}_2 \times \{0\} \cup \mathcal{X}_1 \times \{0\} \times \mathcal{X}_3 \cup \{0\} \times \mathcal{X}_2 \times \mathcal{X}_3 \cup \mathcal{X}_1 \times \{0\} \times \{0\} \cup \{0\} \times \mathcal{X}_2 \times \{0\} \cup \{0\} \times \{0\} \times \mathcal{X}_3;$$

and w_a is the number of partners in match a (3, 2, or 1). We denote by $\boldsymbol{\phi}^{123}$ the matrix whose k-th column contains all values of $\phi^k_{x_1 x_2 x_3}$, stacked in a vector; $\boldsymbol{\phi}^{120}$ the matrix whose k-th column contains all values of $\phi^k_{x_1 x_2 0}$, stacked in a vector; etc. Then the matrix $\boldsymbol{Z}$ has $|\mathcal{A}|$ rows and $(K + X_1 + X_2 + X_3)$ columns and

$$\boldsymbol{Z} = \begin{pmatrix}
\boldsymbol{\phi^{123}}/3 & -\frac{1}{3} \boldsymbol{I}_{X_1} \otimes \mathbf{1}_{(X_2 X_3, 1)} & -\frac{1}{3} \mathbf{1}_{(X_1,1)} \otimes \boldsymbol{I}_{X_2} \otimes \mathbf{1}_{(X_3,1)} & -\frac{1}{3} \mathbf{1}_{(X_1 X_2, 1)} \otimes \boldsymbol{I}_{X_3} \\
\boldsymbol{\phi^{120}}/2 & -\frac{1}{2} \boldsymbol{I}_{X_1} \otimes \mathbf{1}_{(X_2,1)} & -\frac{1}{2} \mathbf{1}_{(X_1,1)} \otimes \boldsymbol{I}_{X_2} & \mathbf{0}_{(X_1 X_2, X_3)} \\
\boldsymbol{\phi^{103}}/2 & -\frac{1}{2} \boldsymbol{I}_{X_1} \otimes \mathbf{1}_{(X_3,1)} & \mathbf{0}_{(X_1 X_3, X_2)} & -\frac{1}{2} \mathbf{1}_{(X_1,1)} \otimes \boldsymbol{I}_{X_3} \\
\boldsymbol{\phi^{023}}/2 & \mathbf{0}_{(X_2 X_3, X_1)} & -\frac{1}{2} \boldsymbol{I}_{X_2} \otimes \mathbf{1}_{(X_3,1)} & -\frac{1}{2} \mathbf{1}_{(X_2,1)} \otimes \boldsymbol{I}_{X_3} \\
\mathbf{0}_{(X_1, K)} & -\boldsymbol{I}_{X_1} & \mathbf{0}_{(X_1, X_2)} & \mathbf{0}_{(X_1, X_3)} \\
\mathbf{0}_{(X_2, K)} & \mathbf{0}_{(X_2, X_1)} & -\boldsymbol{I}_{X_2} & \mathbf{0}_{(X_2, X_3)} \\
\mathbf{0}_{(X_3, K)} & \mathbf{0}_{(X_3, X_1)} & \mathbf{0}_{(X_3, X_2)} & -\boldsymbol{I}_{X_3}
\end{pmatrix}.$$

The Poisson-GLM estimates of $\boldsymbol{\gamma}$ give estimates of $\boldsymbol{\beta}$ and of the average expected utilities $u^k_{x_k} = a^k_{x_k} + \log \hat{n}^k_{x_k}$.

E The Poisson-GLM Estimator for the unipartite Model

For the linear Choo and Siow unipartite model, we have

$$\Phi_{xy} = \log \frac{\mu^2_{xy}}{\mu_{x0} \mu_{y0}} \quad \text{and} \quad u_x = -\log(\mu_{x0}/n_x).$$

To define the Poisson estimator we start by replacing $\hat{\boldsymbol{\mu}}$ with $\tilde{\boldsymbol{\mu}}$, defined by

$$\tilde{\mu}_{xy} = \begin{cases} \hat{\mu}_{xy}/2 & \text{if } x \neq y \\ \hat{\mu}_{xx} & \text{if } x = y. \end{cases}$$

The function L becomes

$$L(\boldsymbol{u}, \boldsymbol{\beta}) = \sum_{x,y} \tilde{\mu}_{xy} \boldsymbol{\phi}_{xy} \boldsymbol{\beta} - 2 \sum_{x,y} \sqrt{\hat{n}_x \hat{n}_y} \exp((\boldsymbol{\phi}_{xy}\boldsymbol{\beta} - u_x - u_y)/2) - \sum_x \hat{n}_x \exp(-u_x) - \sum_x \hat{n}_x u_x.$$

The corresponding Poisson model has only one-way fixed effects, with $w_{xy} = 2$, $w_x = 1$, and

$$\begin{aligned} (\boldsymbol{Z}\boldsymbol{\gamma})_{xy} &= (\boldsymbol{\phi}_{xy}\boldsymbol{\beta} - u_x + \log \hat{n}_x - u_y + \log \hat{n}_y)/2 \\ (\boldsymbol{Z}\boldsymbol{\gamma})_x &= -u_x + \log \hat{n}_x. \end{aligned}$$

This is achieved by defining $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{a})$ with $a_x = u_x - \log \hat{n}_x$ and the following matrix $\boldsymbol{Z}$, with $(X^2 + X)$ rows and $(K + X)$ columns:

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{\phi}/2 & -\frac{1}{2}(\mathbf{1}_X \otimes \boldsymbol{I}_X + \boldsymbol{I}_X \otimes \mathbf{1}_X) \\ \mathbf{0}_{(X,K)} & -\boldsymbol{I}_X \end{pmatrix}.$$