# Estimating Separable Matching Models*

Alfred Galichon[†]   Bernard Salanié[‡]

March 24, 2022

**Abstract**

In this paper we propose two simple methods to estimate models of matching with transferable and separable utility introduced in Galichon and Salanié (2022). The first method is a minimum distance estimator that relies on the generalized entropy of matching. The second relies on a reformulation of the more special but popular Choo and Siow (2006) model; it uses generalized linear models (GLMs) with two-way fixed effects.

Keywords: matching, marriage, assignment, estimations comparison.

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. They used their model to estimate the effect of the 1973 liberalization of abortion in the US on marriage outcomes. In doing so, they used a nonparametric estimator of the matching patterns. Their specification is a natural extension of the multinomial logit model, and it has become quite popular.

*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.

The Choo and Siow specification rests on three main assumptions that will be defined later in the paper: separability; large market; and standard type I extreme value random utility. In Galichon and Salanié (2022), we showed that the third, distributional assumption is not necessary: for any (separable) distribution of the errors, the joint surplus is nonparametrically identified. The nonparametric estimator of Choo and Siow was feasible in their case as they only conditioned on the ages of the partners in a couple. It 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 all of these cases, the analyst must resort to parametric models. This note shows 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 (again, to be formally defined later in the paper).

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.

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 the 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.

We conclude with a brief discussion of the pros and cons of these two methods. Both are coded in a Python package called `cupid_matching` that is available on the standard repositories[1].

[1]See http://bsalanie.github.io for more information, and https://share.streamlit.io/bsalanie/cupid_matching_st/main/cupid_streamlit.py for an interactive Streamlit app that demonstrates solving and estimating a Choo and Siow (2006) model.

# 1 The Model

This paper applies to a bipartite matching market with perfectly transferable utility. For simplicity, we refer to potential partners as “men” and “women”. We use the same notation as in Galichon and Salanié (2022). We assume that the analyst can only observe which of a finite set of *types* each individual belongs to. Men and women of a given type differ along some dimensions that they all observe, while the analyst does not. Each man $i \in \mathcal{I}$ belongs to one group of (observable) type $x_i \in \mathcal{X}$; and, similarly, each woman $j \in \mathcal{J}$ belongs to one (observable) type $y_j \in \mathcal{Y}$. We will say that “man $i$ is of type $x$” and “woman $j$ is of type $y$.” We denote $\mu_{ij}$ the indicator function for a matching between man $i$ and woman $j$, which is equal to 1 if $i$ and $j$ are matched and to 0 otherwise. Similarly, $\mu_{i0}$ and $\mu_{0j}$ are the indicator of $i$ or $j$ to remain unmatched, respectively. Without loss of generality, we assimilate $\mathcal{X}$ to $\{1,\ldots,X\}$ and $\mathcal{Y}$ to $\{1,\ldots,Y\}$. As $\mathcal{X}$ and $\mathcal{Y}$ will later serve as choice sets of partners types for women and men, respectively, and as the marital options needs to include remaining unmatched, we shall add the option to remain unmatched 0 to these sets and 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.

We denote $n_x$ the mass of men of type $x \in \mathcal{X}$, and $m_y$ the mass of women of type $y \in \mathcal{Y}$. We denote $\boldsymbol{q} = (\boldsymbol{n}, \boldsymbol{m})$ the vector that collects the margins $\boldsymbol{n}$ and $\boldsymbol{m}$ of the problem

In addition to the margins $\boldsymbol{q}$, the analyst observes matchings at the type level. We denote $\mu_{xy}$ the mass of the couples where the man belongs to type $x$, and where the woman belongs to type $y$, which is formally defined as $\mu_{xy} = \sum_{i\in\mathcal{I},j\in\mathcal{J}} \mu_{ij} 1\{x_i = x\}1\{y_j = y\}$. We also denote $\mu_{x0} = \sum_{i\in\mathcal{I}} \mu_{i0} 1\{x_i = x\}$, and $\mu_{0y} = \sum_{j\in\mathcal{J}} \mu_{0j} 1\{y_j = y\}$ the mass of single individuals who are respectively men of type $x$ and women of type $y$. We will be interested in the limiting market with a large number of men in any type $x$, and of women in any type $y$. Since the problem is homogeneous, we shall normalize the total mass $N$ of households to one; that is, we rescale $\boldsymbol{\mu}$ and $\boldsymbol{q}$ by a multiplicative factor $N$ such that $\sum_{x,y} \mu_{xy} + \sum_x \mu_{x0} + \sum_y \mu_{0y} = 1$. Again, we use the boldface notation $\boldsymbol{\mu}$ to denote the vector of matching numbers. We denote $\mathcal{A} = \mathcal{X} \times \mathcal{Y} \cup \mathcal{X} \times \{0\} \cup \{0\} \times \mathcal{Y}$ the set of possible marital arrangements (matched household of type $xy$, or single households of type $x0$ or $0y$), so that $\boldsymbol{\mu}$ is a vector of $\mathbb{R}^{\mathcal{A}}$.

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

We model the *joint surplus* $\tilde{\Phi}_{ij}$, which is the sum of the cardinal utilities that both a man $i$ and a woman $j$ jointly obtain by being matched together, and we assume a *separable* matching surplus:

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

*(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* $\tilde{\Phi}_{0j} = \eta_{0j}$*,*

*where, conditional on* $x_i = x$*, the random vector* $\boldsymbol{\varepsilon}_i = (\varepsilon_{iy})_{y\in\mathcal{Y}_0}$ *has probability distribution* $\mathbb{P}_x$*, and, conditional on* $y_j = y$*, the random vector* $\boldsymbol{\eta}_j = (\eta_{xj})_{x\in\mathcal{X}_0}$ *has probability distribution* $\mathbb{Q}_y$*. The distributions* $\mathbb{P}_x$ *and* $\mathbb{Q}_y$ *have full support and a density with respect to the Lebesgue measure. The variables*

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

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

Separability allows for a restricted form of "matching on unobservables"; it rules out interaction terms on characteristics that are unobserved on both sides of the market, e.g. some unobserved preference of man $i$ for some unobserved characteristics of woman $j$.

Chiappori, Salanié, and Weiss (2017) and Galichon and Salanié (2022) showed that under separability, at any stable matching $\boldsymbol{\mu}$ there exist two matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ such that for all $(x, y)$, $U_{xy}+V_{xy} = \Phi_{xy}$, and $U_{x0} = V_{0y} = 0$, and such that man $i$ of type $x$ is assigned option $y = 0, 1, \dots, Y$ which maximizes $U_{xy}+\varepsilon_{iy}$ (where option 0 means remaining unmatched, and option $y \neq 0$ means being matched with a woman of type $y$); similarly woman $j$ of type $y$ is assigned option $x = 0, 1, \dots, X$ which maximizes $V_{xy} + \eta_{xj}$.

## 1.1 Generalized Entropy

Consider the classic "Emax" function $G_x$ defined as follows. In this paragraph we let $\boldsymbol{U} = (U_1, \dots, U_Y)$ be a $Y$-dimensional vector. Then we define

$$G_x(\boldsymbol{U}) = E_{\mathbb{P}_x} \max\left(\max_{y\in\mathcal{Y}}(U_y + \varepsilon_{iy}), \varepsilon_{i0}\right).$$

As a maximum of linear functions, $G_x$ is a convex function. We denote $\partial G_x(\boldsymbol{U})$ its subgradient; because of the assumptions made on $\mathbb{P}_x$, it is a singleton almost everywhere.

Now take the Legendre-Fenchel transform of $G_x$: for any $(\nu_1, \ldots, \nu_Y)$ such that $\sum_{y \in \mathcal{Y}} \nu_y \leq 1$, we define

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

It is another convex function; and since $G_x$ is convex, $G_x$ is the Legendre-Fenchel transform of $G_x^*$. As a consequence,

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

This convex duality is at the core of the identification and inference results in Galichon and Salanié (2022).

Defining $H_y$ and $H_y^*$ in the same way, we get the *generalized entropy*: for any feasible matching $\boldsymbol{\mu}$,

$$\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{q}) = -\sum_{x \in \mathcal{X}} n_x G_x^* \left( \frac{\boldsymbol{\mu}_{x\cdot}}{n_x} \right) - \sum_{y \in \mathcal{Y}} m_y H_y^* \left( \frac{\boldsymbol{\mu}_{\cdot y}}{m_y} \right). \tag{1.2}$$

The function $\mathcal{E}$ only depends on the matching 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 $\boldsymbol{\eta}$.

## 1.2 The Data

We assume that the analyst observes a random sample of size $N$ from a large population of households. By simple counting (possibly using sampling weights), she obtains estimators of the matching patterns $\hat{\mu}_{xy}, \hat{\mu}_{x0}$, and $\hat{\mu}_{0y}$, as well as the margins:

$$\hat{n}_x = \hat{\mu}_{x0} + \sum_{y \in \mathcal{Y}} \hat{\mu}_{xy}$$

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

and a consistent estimator $\Sigma_{\hat{\boldsymbol{\mu}}}$ of their asymptotic variance-covariance matrix, given by

$$\Sigma_{\hat{\boldsymbol{\mu}}} = \operatorname{diag}(\hat{\boldsymbol{\mu}}) - \hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^\top.$$

## 2 Minimum-distance Estimation

Recall that we have assumed that each $\mathbb{P}_x$ (resp. each $\mathbb{Q}_y$) has full support on $\mathbb{R}^{Y+1}$ (resp. $\mathbb{R}^{X+1}$). Then all $\mu_{xy}, \mu_{x0}, \mu_{0y}$ must be positive; as a consequence, the $G_x, H_y, G_x^*, H_y^*$ functions are continuously differentiable everywhere, as is the generalized entropy function $\mathcal{E}$.

Galichon and Salanié (2022) showed that at the stable matching $\boldsymbol{\mu}$, the joint surplus matrix $\boldsymbol{\Phi}$ can be obtained by the following simple formula:

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

These are the first-order conditions of the maximization of the total joint surplus

$$\mathcal{W} = \max_{\boldsymbol{\mu}} \left( \sum_{x,y} \mu_{xy} \Phi_{xy} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{q}) \right).$$

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 write the generalized entropy function $\mathcal{E}^{\boldsymbol{\alpha}}$ and the parameterized surplus vector $\boldsymbol{\Phi}^{\boldsymbol{\beta}}$. Then one can use (2.1) as the basis for a minimum distance estimator[2]. That is, we write a mixed hypothesis as

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

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

We choose $\hat{\boldsymbol{\lambda}}$ to minimize $\|\boldsymbol{D}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})\|^2_{\boldsymbol{S}}$ for some positive definite $(X \times Y, X \times Y)$ matrix $\boldsymbol{S}$. By the general theory of minimum distance estimators, we know that this yields a consistent estimator of $\boldsymbol{\lambda}$ if the model is well specified, and that if we choose $\boldsymbol{S} = \hat{\boldsymbol{\Omega}}^{-1}$ where $\hat{\boldsymbol{\Omega}}$ consistently estimates $V\boldsymbol{D}^{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ (and can be obtained by the delta method), the minimum distance estimator will reach its efficiency bound. Further, if the model is well specified and the choice of $\boldsymbol{S}$ is the efficient one, the minimized value of the squared norm follows a $\chi^2$ of degree $X \times Y - d_{\boldsymbol{\alpha}} - d_{\boldsymbol{\beta}}$. Note that this optimization problem is *not* a convex optimization problem in general.

### 2.1 The Linear Case

Minimum-distance estimation is a particularly appealing strategy if 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:

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

[2]Note that in general, one should choose $d_{\boldsymbol{\alpha}} + d_{\boldsymbol{\beta}} \leq X \times Y$ to ensure identification.

and

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

for some vectors of basis functions $\boldsymbol{e}(\boldsymbol{\mu}, \boldsymbol{q})$ and $\boldsymbol{\phi}$. Then

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

is linear in the parameters $\boldsymbol{\lambda}$. (Recall that, for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$, the vector $e_{xy}(\boldsymbol{\mu}, \boldsymbol{q})$ is of size $d_{\boldsymbol{\alpha}}$, and $\boldsymbol{\phi}_{xy}$ is of size $d_{\boldsymbol{\beta}}$.)

These two conditions call for several remarks. Condition (2.3) is a natural choice for a flexible specification. 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. As we will see, it holds in several other leading examples. Note also that the parameter-free part $\boldsymbol{e}^0$ is necessary in order to normalize the scale of the error terms, which is otherwise not identified in this discrete-choice model.

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 $\boldsymbol{e}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$ 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

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

Since $\hat{\boldsymbol{e}}^0$ and $\hat{\boldsymbol{F}}$ are functions of $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{q}})$, the variance $\hat{\boldsymbol{\Omega}}(\boldsymbol{\lambda})$ of $\hat{\boldsymbol{D}}^{\boldsymbol{\lambda}}$ can be computed from $\hat{\boldsymbol{V}}$ using the delta method. Again, taking $\boldsymbol{S}$ to be the inverse of $\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\lambda}})$ is the efficient choice. This procedure is summarized in Box 1.

**Box 1: min-distance estimation, linear case**

1. Choose any positive definite matrix $\boldsymbol{S}$ and solve (2.4) for a consistent estimator $\boldsymbol{\lambda}^*$
2. Use the delta method to estimate the variance $\boldsymbol{\Omega}^*$ of $\hat{\boldsymbol{D}}^{\boldsymbol{\lambda}}$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$; let $\boldsymbol{S}^* = (\boldsymbol{\Omega}^*)^{-1}$
3. Take $\boldsymbol{S} = \boldsymbol{S}^*$ and solve (2.4) again for $\hat{\boldsymbol{\lambda}}$
4. The variance-covariance matrix of $\hat{\boldsymbol{\lambda}}$ is consistently estimated by
$$\left(\hat{\boldsymbol{F}}^\top \boldsymbol{S}^* \hat{\boldsymbol{F}}\right)^{-1}$$
5. Under the null of correct specification, the statistic
$$\hat{\boldsymbol{T}} = \left(\hat{\boldsymbol{D}}^{\hat{\boldsymbol{\lambda}}}\right)^\top \boldsymbol{S}^* \hat{\boldsymbol{D}}^{\hat{\boldsymbol{\lambda}}}$$
converges to a $\chi^2(X \times Y - d_{\boldsymbol{\alpha}} - d_{\boldsymbol{\beta}})$ distribution.

If the distributions $(\mathbb{P}_x)$ and $(\mathbb{Q}_y)$ are parameter-free, the matrix $\boldsymbol{\Omega}^*$ does not depend on $\boldsymbol{\lambda}$ any more, and $\hat{\boldsymbol{F}}$ is simply the matrix $\boldsymbol{\phi}$. The estimators of $\boldsymbol{\lambda} = \boldsymbol{\beta}$ can be obtained following the procedure described in Box 2.

**Box 2: min-distance estimator, linear case with parameter-free heterogeneity**

1. Evaluate $\boldsymbol{\Omega}^* = V\hat{\boldsymbol{e}}^0$ and $\boldsymbol{S}^* = (\boldsymbol{\Omega}^*)^{-1}$
2. Solve the linear system $\left(\boldsymbol{\phi}^\top \boldsymbol{S}^* \boldsymbol{\phi}\right) \boldsymbol{\beta} = -\boldsymbol{\phi}^\top \boldsymbol{S}^* \hat{\boldsymbol{e}}^0$
3. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is consistently estimated by
$$\left(\boldsymbol{\phi}^\top \boldsymbol{S}^* \boldsymbol{\phi}\right)^{-1}$$
4. Under the null of correct specification, the statistic
$$\hat{\boldsymbol{T}} = \left(\boldsymbol{\phi}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}^0\right)^\top \boldsymbol{S}^* \left(\boldsymbol{\phi}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}^0\right)$$
converges to a $\chi^2(X \times Y - d_{\boldsymbol{\beta}})$ distribution.

Note that since $\boldsymbol{\phi}$ is non-random, the variance of $\hat{\boldsymbol{D}}^{\boldsymbol{\lambda}}$ is the variance of the derivative of the generalized entropy. Step 2 therefore requires evaluating the second derivatives of the generalized entropy $\mathcal{E}$: by the delta method,

$$V\hat{\boldsymbol{D}}^{\boldsymbol{\lambda}} = \begin{pmatrix} \mathcal{E}_{\boldsymbol{\mu\mu}}^{\top} & \mathcal{E}_{\boldsymbol{\mu q}}^{\top} \end{pmatrix} V \begin{pmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\boldsymbol{q}} \end{pmatrix} \begin{pmatrix} \mathcal{E}_{\boldsymbol{\mu\mu}} \\ \mathcal{E}_{\boldsymbol{\mu q}} \end{pmatrix}.$$

It is easy to see from the definition in (1.2) that the first derivative of $\mathcal{E}$ with respect to $\mu_{xy}$ 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 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 $\boldsymbol{\alpha}$. In our third example, the calculation requires finding the fixed point of a contraction, in a way that is familiar from empirical industrial organization.

### 2.2.1 The Heteroskedastic Logit Model

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 $\sigma_x$ and $\tau_y$ respectively. Then $\boldsymbol{\alpha} = (\boldsymbol{\sigma}, \boldsymbol{\tau})$ and the derivatives of the generalized entropy function 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 $\mu_{0y} = m_y - \sum_{x \in \mathcal{X}} \mu_{xy}$. The second derivatives of the generalized entropy take a very simple form:

$$\frac{\partial^2 \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy} \partial \mu_{zt}}(\boldsymbol{\mu}, \boldsymbol{q}) = -\frac{\sigma_x}{\mu_{x0}} \mathbb{1}(z = x) - \frac{\tau_y}{\mu_{0y}} \mathbb{1}(t = y) - \frac{\sigma_x + \tau_y}{\mu_{xy}} \mathbb{1}(z = x, t = y) \tag{2.5}$$

and

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

Scale normalization is done by fixing the value of one of the parameters in $\boldsymbol{\alpha}$. The Choo and Siow homoskedastic model obtains when all $\sigma_x$ and $\tau_y$ equal one; a gender-heteroskedastic model would have all $\sigma_x$ equal to one and all $\tau_y$ equal to an unknown $\tau$. Chiappori, Salanié, and Weiss (2017) applied a minimum distance estimator to the homoskedastic and heteroskedastic logit models.

### 2.2.2 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 $1-(\rho_n^x)^2$ (with $\rho_0^x = 1$ for the nest made of alternative 0). Similarly, for women of type $y$, alternative 0 is in a nest by itself with parameter $\delta_0^y = 1$ 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

$$\begin{aligned}\frac{\partial \mathcal{E}^{\boldsymbol{\alpha}}}{\partial \mu_{xy}}(\boldsymbol{\mu}, \boldsymbol{q}) = &-\rho_n^x \log \frac{\mu_{xy}}{\mu_{x0}} - (1-\rho_n^x)\log\frac{\mu_{xn}}{\mu_{x0}} \\ &-\delta^y_{n'} \log \frac{\mu_{xy}}{\mu_{0y}} - (1-\delta^y_{n'})\log\frac{\mu_{n'y}}{\mu_{0y}},\end{aligned} \tag{2.7}$$

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 $\rho_{xn}$ only depends on $n$).

### 2.2.3 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 (Berry, Levinsohn, and Pakes, 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.

Let individual preferences $\boldsymbol{e}$ of men of type $x$ have distribution $\mathbb{P}^e_x$. We will seek to estimate the parameters $\boldsymbol{\beta}$ of the joint surplus, the scale factor $s$, and the parameters of the distributions $\mathbb{P}^e_x$. We collect $s$ and the parameters of $\mathbb{P}^e_x$ in a vector $\boldsymbol{\alpha}$.

To compute the derivative of the generalized entropy function, we recall from Galichon and Salanié (2022) that

$$G_x^*(\boldsymbol{\nu};\boldsymbol{\alpha}) = -\min_{U_0=0,\boldsymbol{U}\in\mathbb{R}^Y}\left[\int s\log\sum_{y=0,1,\ldots,Y}\exp\left(\frac{U_y+(\boldsymbol{Ze})_y}{s}\right)d\mathbb{P}_x^e(\boldsymbol{e}) - \sum_{y\in\mathcal{Y}}\nu_y U_y\right].$$

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

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

This is exactly isomorphic to the inversion problem in Berry, Levinsohn, and Pakes (1995), with the unknown $\boldsymbol{U}$ standing for the product effects and $\boldsymbol{\nu}$ playing the role of the product market shares. 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 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:

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

The limit case $s=0$ yields the pure characteristics model of Berry and Pakes (2007). Then the system to be solved for row $x$ of $\boldsymbol{U}$ is

$$\nu_y = \mathbb{P}_x^e\left(y\in\arg\max_{t=0,1,\ldots,Y}(U_t+\boldsymbol{Z}_t\boldsymbol{e})\right) \;\; \forall y=1,\ldots,Y.$$

If each $\boldsymbol{Z}_t$ is a scalar, the inequalities boil down to

$$e^-(y;\boldsymbol{U},\boldsymbol{Z}) \equiv \max_{t|Z_t<Z_y}\frac{U_t-U_y}{Z_y-Z_t} \leq e \leq \min_{t|Z_t>Z_y}\frac{U_y-U_t}{Z_t-Z_y} \equiv e^+(y;\boldsymbol{U},\boldsymbol{Z}),$$

and the system of equations to be solved for $\boldsymbol{U}$ is

$$\nu_y = \mathbb{P}_x^e(e^+(y;\boldsymbol{U},\boldsymbol{Z})) - \mathbb{P}_x^e(e^-(y;\boldsymbol{U},\boldsymbol{Z})) \;\; \forall y=1,\ldots,Y.$$

## 3 Moment-based Estimation by Poisson Regression

Now take the generalized entropy function $\mathcal{E}$ as known/assumed; and assume that the joint surplus vector $\boldsymbol{\Phi}\in\mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ is semilinear: $\boldsymbol{\Phi}^{\boldsymbol{\beta}} = \boldsymbol{\phi\beta}$, where $\boldsymbol{\beta}$ is a vector of dimension $d_{\boldsymbol{\beta}}$ and $\phi$ is a $|\mathcal{X}||\mathcal{Y}|\times d_{\boldsymbol{\beta}}$ matrix. 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:

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

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 $\mathcal{W}(\boldsymbol{\beta}, \boldsymbol{q}) = \max_{\boldsymbol{\mu}} \left( \boldsymbol{\mu} \boldsymbol{\Phi}^{\boldsymbol{\beta}} + \mathcal{E}(\boldsymbol{\mu}, \boldsymbol{q}) \right)$ is the value of the total joint surplus. With a semilinear specification for $\boldsymbol{\Phi}^{\boldsymbol{\beta}}$, both of these 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. Define a vector $w \in \mathbb{R}^{\mathcal{A}}$ by $w_{xy} = 2$ if $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and $w_{xy} = 1$ if $x = 0$ or if $y = 0$, so that $w_{xy}$ is the size of household $xy$, namely 2 if matched, 1 if single. The following theorem summarizes our results.

**Theorem 1** (Estimating the logit model with a Poisson regression)**.** *In the Choo and Siow model, the moment-matching estimator $\hat{\beta}$ is the solution to a Poisson regression of $\left(\hat{\mu}_{xy}\right)_{xy \in \mathcal{A}}$ on $\left(\Phi^{\beta}_{xy}/w_{xy}\right)_{xy \in \mathcal{A}}$, with 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, $\beta$ 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 1 is given in Appendix B. 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.

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.

**Box 3: 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 $k = 1, \ldots, K$, represent the vector$(\boldsymbol{\phi}^k_{xy})_{xy \in \mathcal{A}}$ in the same order. Then represent the $|\mathcal{A}| \times K$ matrix $\phi$ from these $K$ column vectors of size $|\mathcal{A}|$.

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

$$\boldsymbol{w} = (\boldsymbol{2}^\top_{X \times Y}, \boldsymbol{1}^\top_X, \boldsymbol{1}^\top_Y)^\top.$$

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

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

5. Run a Poisson regression of $\hat{\mu}$ on $\boldsymbol{Z}$ with weights $\boldsymbol{w}$. Do not add 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 $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{a}}^\top, \hat{\boldsymbol{b}}^\top)^\top \in \mathbb{R}^{K+X+Y}$. 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 2.** *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* $\boldsymbol{W} = diag(w)$*, we have*

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

*and*

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

# 4 Monte Carlo Simulation

We coded these two estimation methods in a `Python` package called `cupid_matching` that is available from the standard repositories[3]. To test the quality of the estimators, we generated data both from a Choo and Siow model and from a semilinear nested logit model. We use both the Poisson estimator and the minimum-distance estimator on the former model, and only the minimum-distance estimator of course on the latter.

In both cases, we take $X = Y = 20$ and we use $K = 8$ basis functions: $1, x, y, x^2, xy, y^2, \mathbb{1}(x \geq y), \max(x-y, 0)$. The true data-generating process has

$$\Phi_{xy} = 1 - \frac{(x-y)^2}{100} + 0.5\mathbb{1}(x \geq y),$$

so that the true $\boldsymbol{\beta}$ is $(1.0, 0.0, 0.0, -0.01, 0.02, -0.01, 0.5, 0.0)$. This could be interpreted as the joint surplus from marriage as a function of the ages of the husband $x$ and of the wife $y$. It is highest when the partners have the same age; if they don't, it is larger when the husband is the older partner. We use equal numbers of men and women; and we choose vectors $\boldsymbol{n} = \boldsymbol{m}$ whose elements form a decreasing geometric sequence with rate 0.8 (there are fewer individuals available for marriages at higher ages).

## 4.1 Semilinear Logit

The semilinear logit model is entirely described above. We use the IPFP algorithm described in Section 4.2 of Galichon and Salanié (2022) to solve for the stable matching patterns $\boldsymbol{\mu}$ for the

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

margins $\boldsymbol{n}$ and $\boldsymbol{m}$. To generate a sample, we draw randomly $N = 10,000$ households from the multinomial probability distribution generated by $\boldsymbol{\mu}$. We generated $S = 1,000$ such samples. We used minimum distance estimation and Poisson GLM on each sample. While the minimum distance estimator only uses a linear regression, the Poisson GLM method uses numerical optimization under the hood. In our simulations using the `sklearn` Python package, the algorithm went astray on 50 of our 1,000 samples, mostly because of overflow errors. We discarded these samples from our analysis.

As Figure 1 shows, on the remaining 950 samples the two estimators perform about equally well. Both estimators exploit the same $X \times Y$ moment conditions

$$\boldsymbol{\phi} \cdot \boldsymbol{\beta} + \frac{\partial \mathcal{E}}{\partial \boldsymbol{\mu}} = \mathbf{0},$$

and both minimize a quadratic form of these conditions. The difference is in the weighting matrix. We saw in Section 2.1 that the minimum-distance estimator uses the variance of the derivative of the entropy at the observed matching. On the other hand, the Poisson estimator uses the Hessian of the entropy at the current parameter values. While the two estimators are quite close in our simulations, one can imagine situations in which the divergence would be larger.

### 4.2 Semilinear Nested Logit

For both men and women, we defined three nests that consist of $\{0\}$, $\{1, \ldots, 10\}$, and $\{11, \ldots, 20\}$. We take the true nest parameters to be all equal to 0.5 (that is, $\rho_n^x = \delta_{n'}^y = 0.5$ for all $n, n' : x \in n$ and $y \in n'$).

To generate samples from the nested logit model, we proceed as with the logit model. The only difference is that setting up the system to be solved for equilibrium requires a bit more work. We describe our IPFP algorithm in Appendix A.

The minimum distance estimator converges fast on all samples. However, we found that a sample size of 10,000 households was much too small to get reliable estimates of the parameters. Figure 2 gives the distribution of the estimates of the four nest parameters (first two rows) and the eight coefficients of the bases for larger sample sizes: respectively $N = 100,000$ and $N = 1,000,000$. There is a clear downwards bias on the nest parameters $\rho$ and $\delta$ when $N = 100,000$, to the point that some estimates are negative. Some of the coefficients of the bases are also badly estimated. With $N = 1,000,000$, the minimum distance estimator performs much better.

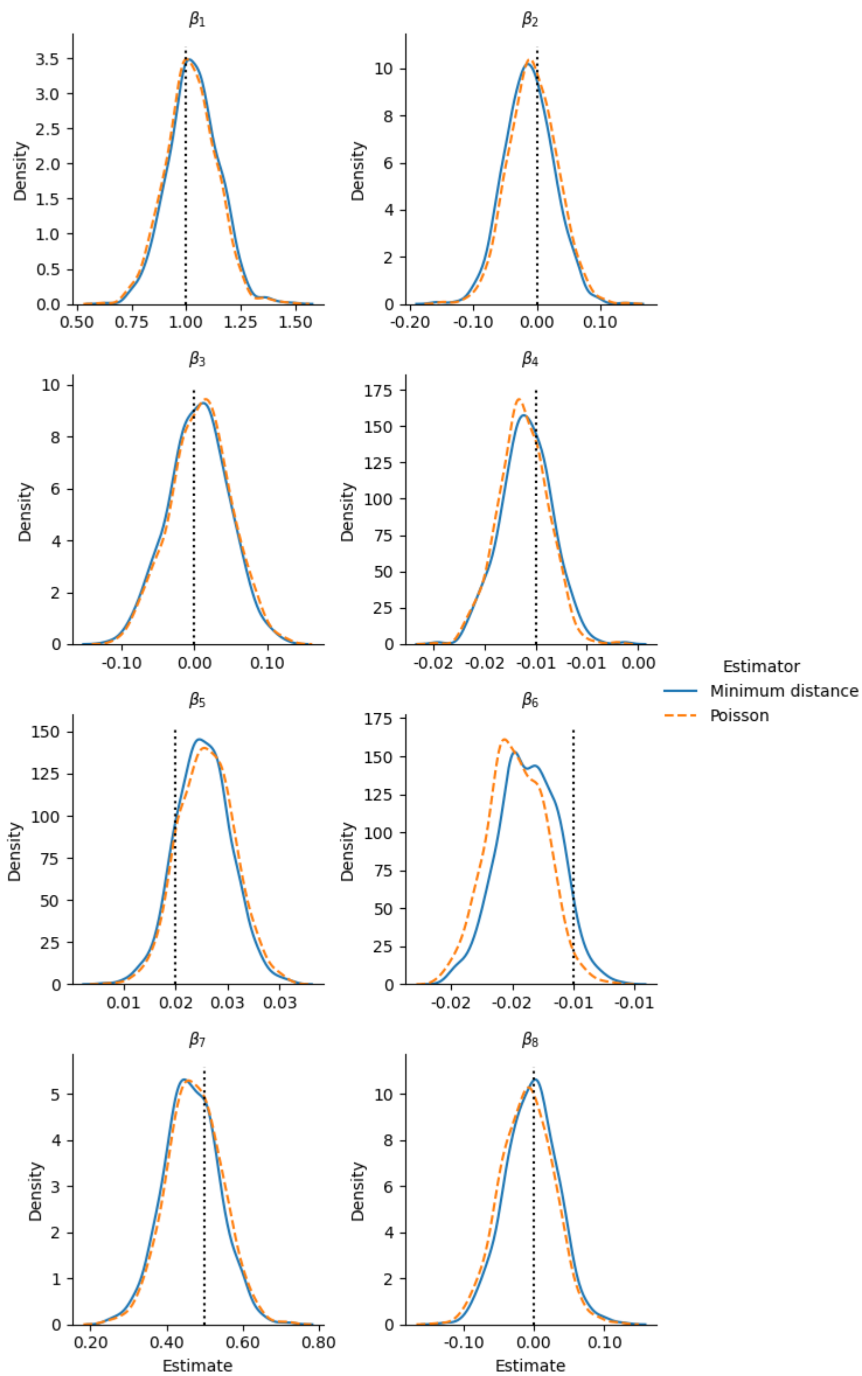


Figure 1: Estimating the Choo and Siow Model

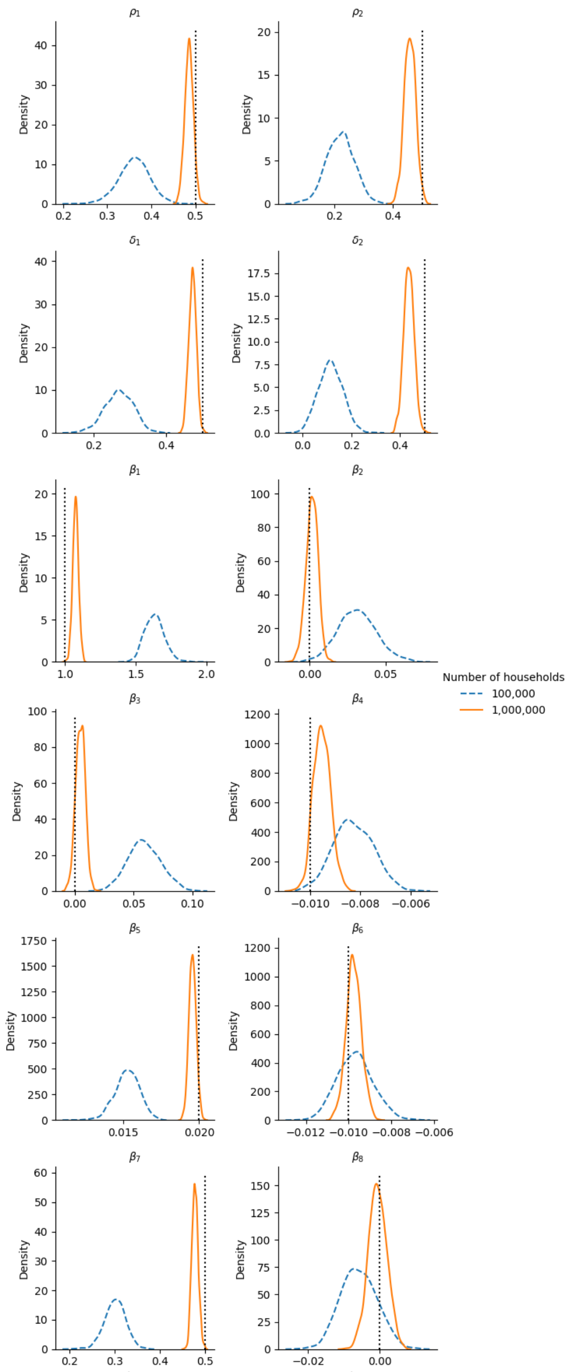


Figure 2: Estimating the Nested Logit Model

## Concluding remarks

Each of the two methods we presented here 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, the data often contains zero cells—some $\hat{\mu}_{xy}$ may be zero. Then the corresponding equation in (2.1) is only an inequality and it must be dropped from the system of estimating equations. An alternative is to add a small positive number $\delta$ to each $\hat{\mu}_{xy}$, to increase the margins $\hat{n}_x$ and $\hat{m}_y$ accordingly, and to estimate on this adjusted data.

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 more robust to zero cells: nothing in Section 3 relied on taking derivatives with respect to $\boldsymbol{\mu}$ at the observed matching patterns.

Our simulations suggest that it takes large sample sizes to get reliable estimates of distributional parameters (our $\boldsymbol{\alpha}$). In labor markets or in marriage markets, large samples are readily available. When they are not (as with matching between firms), it may be better to stick to the Choo and Siow (2006) specification. Fortunately, the simulations reported in Chiappori, Nguyen, and Salanié (2019) are encouraging as to its robustness.

## References

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

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

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

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

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

Galichon, A., and B. Salanié (2022): “Cupid’s Invisible Hand: Social Surplus and Identification in Matching Models,” *Review of Economic Studies*, forthcoming.

Santos Silva, J. M. C., and S. Tenreyro (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: $\rho_n^x \equiv \rho_n$ and $\delta_{n'}^y \equiv \delta_{n'}$. Equation (2.7) 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 $\mu_{x0}$ uniquely[4] . Once $\mu_{x0}$ is known, we can plug it in (A.2) to obtain the values of $\mu_{xn}$ 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 $\mu_{0y}$ and $\mu_{n'y}$ given the values of $(\mu_{x0}, \mu_{xn})$ for all $x$. We iterate until convergence and we use (A.1) to compute the matching patterns $\mu_{xy}$.

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

## B Proofs

### B.1 Proof of theorem 1

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 have at the stable matching $(\boldsymbol{\mu}, \boldsymbol{u}, \boldsymbol{v})$ for a joint surplus $\boldsymbol{\Phi}^{\boldsymbol{\beta}}$:

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

Consider the maximization of the following expression:

$$\begin{aligned} &\sum_{x,y} \hat{\mu}_{xy} \boldsymbol{\phi}_{xy} \boldsymbol{\beta} - 2 \sum_{x,y} \sqrt{\hat{n}_x \hat{m}_y} \exp((\boldsymbol{\phi}_x \boldsymbol{\beta} - u_x - v_y)/2) \\ &- \sum_{x} \hat{n}_x \exp(-u_x) - \sum_{y} \hat{m}_y \exp(-v_y) - \sum_{x} \hat{n}_x u_x - \sum_{y} \hat{m}_y v_y \end{aligned}$$

over $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{\beta}$. We see that the first order conditions yield that $\boldsymbol{\mu}$ defined in B.1 satisfies the margin equations

$$\sum_{y} \mu_{xy} + \mu_{x0} = n_x \tag{B.2}$$

$$\sum_{x} \mu_{xy} + \mu_{0y} = m_y \tag{B.3}$$

for the first order conditions with respect to $u_x$ and $v_y$, and

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

for the first order conditions with respect to $\beta_k$.

Now remember that the log-likelihood function of a Poisson count model with parameter $\exp(\boldsymbol{Z}_a^\top \boldsymbol{\gamma})$ is

$$l\left(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma}; \boldsymbol{w}\right) = \sum_{a \in \mathcal{A}} w_a \left( \hat{\mu}_a \boldsymbol{Z}_a^\top \boldsymbol{\gamma} - \exp\left(\boldsymbol{Z}^t op_a \boldsymbol{\gamma}\right) - \log(\hat{\mu}_a!) \right). \tag{B.4}$$

if the observations $(\hat{N}_a, \boldsymbol{Z}_a)_{a \in \mathcal{A}}$ are weighted by a vector $\boldsymbol{w}$. Define $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{a}^\top, \boldsymbol{b}^\top)$ 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 (1), we have[5]

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

and up to constant terms, $l$ and $L$ are identical.

## B.2 Proof of theorem 2

The variance-covariance matrix of $\hat{\boldsymbol{\gamma}}$ follows directly from the fact that it maximizes (B.4), and hence is an M-estimator, see chapter 5 of van der Vaart (1998). The maximization of (B.4) gives first-order conditions

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

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

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

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

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

and

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

[5]Note that $\boldsymbol{Z}_i$ should be interpreted here as row $i$ of the matrix $\boldsymbol{Z}$.