Matching with Trade-offs:

Revealed Preferences over Competing Characteristics

Alfred Galichon[1] Bernard Salanié [2]

First version dated December 6, 2008. The present version is of March 22, 2010[3].

[1]Economics Department, École polytechnique; e-mail: alfred.galichon@polytechnique.edu

[2]Department of Economics, Columbia University; e-mail: bsalanie@columbia.edu.

[3]The authors are grateful to Guillaume Carlier, Pierre-André Chiappori, Piet Gauthier, Jerry Hausman, Jim Heckman, Kevin Lang, Guy Laroque, Christian Léonard, Sonia Oreffice, Rob Shimer as well as seminar participants at Crest, Ecole Polytechnique, séminaire Roy, University of Chicago, Harvard-MIT, the Toulouse School of Economics, the University of New South Wales, the University of Melbourne, Australian National University and the University of Alicante for useful comments and discussions. Much of this paper was written while Salanié was visiting the Toulouse School of Economics; he is grateful to the Georges Meyer endowment for its support. Galichon gratefully acknowledges support from Chaire EDF-Calyon "Finance and Développement Durable," and Chaire Axa "Assurance et Risques Majeurs" and FiME, Laboratoire de Finance des Marchés de l'Energie (www.fime-lab.org).

Abstract

We investigate in this paper the theory and econometrics of optimal matchings with competing criteria. The surplus from a marriage match, for instance, may depend both on the incomes and on the educations of the partners, as well as on characteristics that the analyst does not observe. The social optimum must therefore trade off matching on incomes and matching on educations. Given a flexible specification of the surplus function, we characterize under mild assumptions the properties of the set of feasible matchings and of the socially optimal matching. Then we show how data on the covariation of the types of the partners in observed matches can be used to estimate the parameters that define social preferences over matches. We provide both nonparametric and parametric procedures that are very easy to use in applications.

Keywords: matching, marriage, assignment.

JEL codes: C78, D61, C13.

Introduction

Starting with Becker (1973), most of the economic theory of one-to-one matching has focused on the case when the surplus created by a match is a function of just two numbers: the one-dimensional types of the two partners. As is well-known, if the types of the partners are one-dimensional and are complementary in producing surplus then the socially optimal matches exhibit positive assortative matching. Moreover, the resulting configuration is stable, it is in the core of the corresponding matching game, and it can be implemented by the celebrated Gale and Shapley (1962) deferred acceptance algorithm.

While this result is both simple and powerful, its implications are also quite unrealistic. If we focus on marriage and type is education for instance, then positive assortative matching has the most educated woman marrying the most educated man, then the second most educated woman marrying marrying the second most educated man, and so on. In practice the most educated woman would weigh several criteria in deciding upon a match; even in the frictionless world studied by theory, the social surplus her match creates may be higher if she marries a man with less education but, say, a similar income. Income and education are only imperfectly correlated; and the correlation patterns differ for men and women. Then the optimal match must trade off assortative matching along these two dimensions. This point is quite general: with multiple types, the stark predictions of the one-dimensional case break down.

Analysts of matching have long felt the need to accommodate the imperfect assortative matching observed in the data, of course. One possibility is to introduce search frictions, as in Shimer and Smith (2000); but the resulting model is hard to handle, and it still implies assortative matching, under stronger conditions. In our view, a simpler solution consists in allowing the joint surplus of a match to incorporate heterogeneity that is unobserved by the analyst. As explained below, this was pioneered by Choo and Siow (2006) and extended by Chiappori, Salanié, Tillman, and Weiss (2008); a different variant, used by Chiappori, Oreffice, and Quintana-Domeque (2009), assumes that the observed heterogeneity has a one-dimensional index structure. Our contribution here is twofold. First, we exhibit a very

simple nonparametric estimator of the surplus function that is at the heart of matching models with transferable utility; second, we explore the properties of optimal or equilibrium matchings[1] and we use our results to describe an empirical strategy and to obtain parametric estimators.

For simplicity, we use the language of the economic theory of marriage in our illustrations; yet nothing we do actually depends on it. The methods proposed in this paper apply just as well to any one-to-one matching problem—or bipartite matchings. In fact, we can even extend them to problems in which the sets of partners are determined endogenously—as with same-sex unions. This is investigated in Section 8, where we consider possible extensions of our setting.

We do require, however, that utility be transferable across partners. Our primitive function is indeed the surplus created by a match, defined as

$$\tilde{\Phi}(\tilde{x}, \tilde{y})$$

where $\tilde{x}$ is the full type of a man and $\tilde{y}$ the full type of the woman who is his partner in this match. A full type has components that are observed by the econometrician, and others that are only observed by the participants on the market. We denote x the observable type of a man, and y that of a woman.

As is well-known, this model is too general to be empirically testable: even without unobserved heterogeneity (when x coincides with $\tilde{x}$ and y coincides with $\tilde{y}$), any observed assignment can be rationalized by a well-chosen surplus function. This is a consequence of a more general theorem by Blair (1984). Echenique (2008) shows that on the other hand, some collections of matchings are not rationalizable: if the analyst can observe identical populations on several assignments, then these assignments must be consistent with each other in a sense that his paper makes precise. But we are unlikely to have such data at hand in general.

[1] A word on terminology: like most of the literature, we call a "match" the pairing of two partners, and a "matching" the list of all realized matches.

Relatedly, analysts sometimes observe several subpopulations which are matching independently and yet have the same surplus function. Fox (2009) shows that under a "rank-order condition" on the unobserved heterogeneity, it is then possible to identify several important features of the surplus function, and in particular how important complementarity is on various dimensions.

While analyzing complementarity is also one of our goals here, many of the applications we have in mind do not fit Fox's assumption that there be enough variation across subpopulations with identical surplus. Marriage markets, for instance, seem to be either so disconnected that their surplus functions are unlikely to be similar, or too connected to make it possible to ignore matching across markets. In this paper, we only assume that we have data on one instance of a matching problem, such as the marriage market in the US in the 1980s, or the market for CEOs. Our data will consist of the values of the observable types of both partners in each realized match, and of the types of unmatched individuals.

Since the optimal/equilibrium matching is determined on the basis of both the observable and the (to us) unobservable types, we will need to impose assumptions that allow us to integrate over the distribution of the unobservable types in a manageable way. Our aim is to start from the observable matching (the distribution of matches across observable types) and to recover as much information as we can on the surplus function. That is, the econometrician observes a distribution $\pi(x, y)$ over the observable types of both partners in observed matches; and he seeks to recover $\tilde{\Phi}$.

To do this, we have to impose assumptions. First, we restrict our analysis to the case when observable types take a finite number of values. Then we impose a separability assumption that excludes interactions between the unobservable types of the partners in the production of joint surplus:

$$\tilde{\Phi}(\tilde{x}, \tilde{y}) = \Phi(x, y) + \chi(\tilde{x}, y) + \xi(\tilde{y}, x),$$

where the χ's and ξ's have conditional mean zero. One interpretation is that the surplus created by a potential match is an unknown function of the types of the partners only, plus preference shocks that are observed by all participants but not by the analyst—in the

nature of unobserved heterogeneity.

This separability assumption was used by Choo and Siow (2006) and then generalized by Chiappori, Salanié, Tillman, and Weiss (2008), who showed that the matching equilibrium then boils down to a series of parallel discrete choice models. While this is an important step on the way to a solution, the resulting model is still too rich to be taken to the data. We need to restrict the distribution of unobserved heterogeneity further. To do this, we adopt Choo and Siow (2006)'s assumption that the terms χ and ξ above are type-I extreme values—giving the model the structure of a multinomial logit.

If the analyst is lucky enough to have very rich data, unobserved heterogeneity is almost irrelevant and the observable matching π maximizes the observable surplus function. A bit more formally, let P and Q denote the marginal distributions over observed types of men and women respectively. Then if there is no unobserved heterogeneity, the observed matching π must maximize

$$E_\pi \Phi(X, Y)$$

over the set $\mathcal{M}(P, Q)$ of all joint distributions π that have marginals P and Q.

On the other hand, if data is so poor that unobserved heterogeneity dominates, then the analyst should observe something that, to him, looks like completely random matching: π should be the product $P \otimes Q$ of the marginal distributions. As is well known, independence maximizes relative entropy: $\pi = P \otimes Q$ minimizes the *mutual information*

$$I(\pi) = E_\pi \log \frac{\pi(X, Y)}{P(X)Q(Y)}$$

over $\mathcal{M}(P, Q)$.

We show that under our assumptions, for any intermediate amount of unobserved heterogeneity, the observable matching π maximizes a straightforward linear combination of the observable surplus and of the mutual information above:

$$E_\pi \Phi(X, Y) - \sigma E_\pi \log \frac{\pi(X, Y)}{P(X)Q(Y)},$$

where σ measures the size of unobserved heterogeneity.

Apart from its simplicity, this objective function has two very nice properties when $\sigma > 0$: it is globally strictly concave, and it has an infinite derivative in zero. The former implies that the optimal matching is unique, and the latter that all observable matches have positive probability: $0 < \pi(x, y) < 1$ for all (x, y). Finally, for any given Φ and σ (and (P, Q)) the optimal matching π is much easier to compute than in the homogeneous case: we show that the well-known Iterative Projection Fitting Procedure is easily adapted to the structure of this problem. Since IPFP is a very fast, very stable and very simple algorithm (known to some economists as RAS), we consider this to be another attractive property of our method.

Under our assumptions, we prove that the surplus function over observable types Φ is nonparametrically partially identified from the data; and that the features that are identified hold considerable economic interest. In fact, the log-likelihood function of the observed matches is one of the surplus functions that can rationalize the data:

$$\Phi \equiv \log \pi;$$

and we can use it to test for complementarities between any two observable dimensions of the types of the partners, such as the education of the wife and the income of the husband. We can also identify the relative strengths of such complementarities across different dimensions.

While this is a remarkably simple and useful result, even discrete types may take a large number of values, making nonparametric estimation impractical—all the more so that π is a joint distribution of types. Parametric analysis will often be necessary in practice. Our IPFP algorithm makes maximum likelihood estimation quite simple even for nonlinear models; but models in which the observable surplus function is linear in the parameters turn out to have quite interesting properties. Consider, for instance, approximating the observable surplus function with a linear expansion over some known basis functions (ϕ^k), with unknown *assorting weights* Λ:

$$\Phi(x, y) = \sum_k \Lambda_k \phi^k(x, y).$$

If the true model in fact belongs to this class, then all relevant information can be expressed

in terms of the mutual information I of the joint distribution of types and of the average values of these basis functions ϕ^k across couples, which we call *covariations*; more formally, the numbers

$$\hat{I} = E_{\hat{\pi}} \log \frac{\hat{\pi}(X,Y)}{\hat{p}(X)\hat{q}(Y)}$$

and

$$\hat{C}^k = E_{\hat{\pi}} \phi^k(X,Y)$$

are sufficient statistics for estimation of Λ and specification testing. This is a very significant reduction in complexity, from the joint distribution $\hat{\pi}(x,y)$ to just these $(K+1)$ numbers.

We first show that if the true model was generated by the assumed basis functions, then in the mutual information $\hat{I}$ should be minimal given the vector of covariations $\hat{C}$, in a sense that we made precise. This gives us a specification test. If the test does not reject the null hypothesis, then there exists a vector of assorting weights Λ for which the optimal matching generates exactly these covariations; and if the true model has positive heterogeneity ($\sigma > 0$), this vector is unique up to multiplication by a positive constant. Moreover, we can test that a correctly specified model is homogeneous ($\sigma = 0$.)

These results lead us to propose a moment matching estimator of the assorting weights that has very desirable properties if the model is correctly specified: it is consistent, asymptotically normal, and asymptotically efficient, and it is also very easy to compute as if maximizes a globally strictly convex function. If the model may have been misspecified, then we can still use this estimator to compute the implied joint distribution of types, and compare it to the nonparametric estimator of π; this gives an additional test for misspecification. Moreover, we can use standard techniques to select among potential sets of basis functions.

This paper thus proves both a negative and a positive result. The negative part is that even if we assume separable heterogeneity with a multinomial logit structure, the model still cannot be rejected since we can always rationalize it by the parameters $\Phi = \log \pi$ and $\sigma = 1$ for instance. The positive part is that given any theory about the way the observable types enter the surplus function (as embodied in a set of basis functions (ϕ^k)), we exhibit well-

behaved estimators of the unknown parameters Λ; we can test whether heterogeneity $\sigma > 0$ is needed to rationalize the data; and we can test whether the basis functions adequately describe matching patterns.

Our methods can also be used heuristically, to explore ways to understand what goes on in matching markets—and how they change across time and space. Standard statistical techniques could for instance be put to work to find the basis functions that explain the largest share of the variation in the data. Such a methodological stance is reminiscent of revealed preferences in consumer theory; in fact the analogy is very sharp, as the underlying theoretical structure is the same. The theoretical work done by Hatfield and Milgrom (2005) and Chiappori, McCann, and Nesheim (2008) also suggest exploring analogies with auctions and hedonic models, respectively.

Our depiction of matching markets of course abstracts from many features of real-world markets. We focus on static, frictionless markets, as in much of the literature on marriage markets. Our data merely consists in the knowledge of "who is married to whom" at a given date. Most models of matching on job markets, for instance, have adopted a much more dynamic perspective, in which job flows in fact provide a lot of information on the underlying parameters. This is hard to do on many matching markets (it would require, for instance, a good theory of divorce) and we leave this for further work. Another recent trend in the economic literature on matching has been to focus on matching technology, such as platforms; on the marriage market, online dating sites are an example (see e.g. Hitsch, Hortacsu, and Ariely (2010) and Lee (2009).) By construction, features of the matching technology such as frictions are encompassed in our matching surplus function, which thus reflects the *social distance* between the observable characteristics of two individuals, reflecting technological and social accessibility. In a recent paper, Echenique, Lee, and Shum (2009) take the dual approach to both transferrable utility and non-transferrable utility models of matching, combining homogeneity in preferences and flexible matching frictions. Our view is that at this level of generality, it is hard to choose between the two approaches: any particular application will have to face the standard issue of finding instruments to

identify frictions and preferences separately (when they do not merge, as may be the case with intercaste marriage for instance.)

Section 1 sets up the matching model we study in the paper, along with our assumptions on the specification of the observable surplus and the process that drives unobserved heterogeneity. In section 2 we build on these assumptions to derive our main analytical results, and we prove (partial) nonparametric identification in section 3. Section 4 discusses possible empirical strategies; and it shows that the presence of heterogeneity in fact makes the computation of the optimal matching much easier. Under the assumption that the surplus function Φ is unknown up to a linear parametrization, we give our results a geometric interpretation in section 5. Section 6 introduces our tests and estimators and derives their asymptotic properties; and section 7 illustrates our methods on a subsample of 2008 US Census data. We conclude by sketching extensions of our methods.

Since much of what we do uses convexity, we recall some definitions and basic results in Appendix A. All proofs are collected in Appendix B. Finally, we should note that there are close parallels between the analysis we develop in the present paper and familiar notions in thermodynamics and statistical physics. E.g the social utility of a matching evokes (minus) the internal energy of a physical system, and the standard error of unobservable heterogeneity parallels its physical temperature. Since the analogy may prove to be as useful to others as it was to us, we elaborate on it in Appendix C.

1 The Assignment Problem

Throughout the paper, we assume that two subpopulations M and W of equal size must be matched, and that utility transfers between partners are unconstrained. Each man (as we will call the members of M) must be matched with one and only one member of W (we will call them women.) Thus we do not model the determination of the unmatched population (the singles) in this paper; we take it as data. We elaborate on this point in our concluding

remarks. Note also that we assumed bipartite matching: the two subpopulations which define admissible partners are exogenously given. This assumption can also be relaxed; see Section 8.

Throughout the paper, we illustrate results on the education/income example sketched in the Introduction, which we denote (ER).

1.1 Population characteristics

Each man m has characteristics $\tilde{x}_m$, of which a subset x_m is observed by the econometrician. We call $\tilde{x}_m$ the full type and x_m the observable type. Similarly, each woman w similarly has a full type $\tilde{y}_w$ and an observable type y_w.

We denote $\tilde{P}$ (resp. $\tilde{Q}$) the distribution of *full* types $\tilde{x}$ (resp. $\tilde{y}$) in the subpopulation M (resp. W), and P (resp. Q) the distribution of *observable* types x (resp. y.) In observed datasets we will have a finite number N of men and women, so that P and Q are the empirical distributions over the types observed in the sample, $\{x_1, ..., x_N\}$ and $\{y_1, ..., y_N\}$ respectively.

Take the education/income example: there a first dimension of observable types is education $E \in \{D, G\}$ (dropout or graduate), and a second dimension is income class R, which takes values 1 to n_R . P describes both the number of graduates among men and the distributions of income among graduate men and among dropout men. In addition, full types may incorporate physical characteristics, religion, and so on.

1.2 Matching

The intuitive definition of a matching is the specification of "who marries whom": given a man of index $m \in \{1, ..., N\}$, it is simply the index of the woman he marries, $w = \sigma(m) \in \{1, ..., N\}$. Imposing that each man be married to one and only one woman at a given time translates into the requirement that σ be a permutation of $\{1, ..., N\}$. This definition is too restrictive in so far as we would like to allow for some randomization. This could arise

because a given type is indifferent between several partner types; or because the analyst only observes a subset of relevant characteristics, and the unobserved heterogeneity induces apparent randomness.

A *feasible matching* (or *assignment*) is therefore defined in all generality as a joint distribution $\tilde{\Pi}$ over types of partners $\tilde{X}$ and $\tilde{Y}$, such that the marginal distribution of $\tilde{X}$ is $\tilde{P}$ and the marginal distribution of $\tilde{Y}$ is $\tilde{Q}$. We denote $\mathcal{M}\left(\tilde{P}, \tilde{Q}\right)$ the set of such joint distributions.

1.3 Surplus of a match

The basic assumption of the model is that matching man m of full type $\tilde{x}_m$ and woman w of full type $\tilde{y}_w$ generates a joint surplus $\tilde{\Phi}(\tilde{x}_m, \tilde{y}_w)$, where $\tilde{\Phi}$ is a deterministic function. Along with most of the matching literature, we assume that

Assumption (O): Observability. Each agent observes the full types $\tilde{x}$ and $\tilde{y}$ of all men and all women, but the econometrician only observes their components x and y.

Assumption (O) rules out asymmetric information between participants in the market, as the economics of matching with incomplete information is a subject of its own. On the other hand, we do not really need to assume full information: $\tilde{\Phi}$ could for instance be reinterpreted as the expectation of a random variable conditional on $\tilde{x}, \tilde{y}$, as long as all participants evaluate it in the same way.

Matching markets are all about complementarities in the generation of surplus. If we are to identify such complementarities between observable types, we have to exclude complementarities between the unobserved components of full types. Following the insight of Choo and Siow (2006), formalized by Chiappori, Salanié, Tillman, and Weiss (2008), we therefore assume:

Assumption (S): Separability. Let $\tilde{x}$ and $\tilde{x}'$ have the same observable type: $x = x'$. Similarly, let $\tilde{y}$ and $\tilde{y}'$ be such that $y = y'$. Then

$$\tilde{\Phi}(\tilde{x}, \tilde{y}) + \tilde{\Phi}(\tilde{x}', \tilde{y}') = \tilde{\Phi}(\tilde{x}, \tilde{y}') + \tilde{\Phi}(\tilde{x}', \tilde{y}).$$

Assumption (S) requires that conditional on observable types, the surplus exhibit no complementarity across unobservable types. It is easy to see that imposing assumption (S) is equivalent to requiring that the idiosyncratic surplus from a match must be additively separable, in the following sense:

$$\tilde{\Phi}(\tilde{x}, \tilde{y}) = \Phi(x, y) + \chi(\tilde{x}, y) + \xi(\tilde{y}, x)$$

with $E_{\tilde{P}}(\chi\left(\tilde{X}, y\right)|X = x) \equiv 0$ and $E_{\tilde{Q}}(\xi\left(\tilde{Y}, x\right)|Y = y) \equiv 0$ for every (x, y).

Given Assumption (S), we call $\Phi(x, y)$ the *observable surplus*. Note that the model is invariant if one rescales the three terms on the right-hand side by the same positive constant. Later on we will normalize these three components.

As proved in Chiappori, Salanié, Tillman, and Weiss (2008), assumption (S) implies that at the optimum (or equilibrium), a given individual (say, a man $\tilde{x}$) has a preference $\xi(\tilde{x}, y)$ for a particular class of observable characteristics (say y), but he is indifferent between all partners which have the same y but a different $\tilde{y}$. More precisely, the optimal matching is characterized by two functions of observable characteristics $U(x, y)$ and $V(x, y)$ that sum up to $\Phi(x, y)$ such that if a man $\tilde{x}$ is matched with a woman of characteristics $\tilde{y}$, he will get utility

$$U(x, y) + \chi(\tilde{x}, y)$$

while his match gets utility

$$V(x, y) + \xi(\tilde{y}, x).$$

Chiappori, Salanié, Tillman, and Weiss (2008) showed that given assumption (S), the matching problem boils down to a set of single-agent choice problems for each type of man and of woman: for instance, man $\tilde{x}$ is matched in equilibrium to a woman $\tilde{y}$ whose observable

type y maximizes

$$U(x, y) + \chi(\tilde{x}, y)$$

over all values in the support of Q.

One justification for assumption (S) would be that in a hypothetical match between man $\tilde{x}$ and woman $\tilde{y}$ that results in a transfer of t from the man to the woman, the man gets utility

$$U_0(x, y) + \chi_0 - t$$

and the woman gets

$$V_0(x, y) + \xi_0 + t$$

with the restrictions that

$$\chi_0 + \xi_0 = \chi(\tilde{x}, y) + \xi(\tilde{y}, x)$$

and $U_0 + V_0 = \Phi$. Note that because transfers are endogenous at the optimum, U and V may be quite different from U_0 and V_0.

It may be useful to resort to an analogy with the specification of demand systems, as used for instance in empirical industrial organization. The utility a consumer with observed type x and full type $\tilde{x}$ gets from consuming a product with observed characteristics y and full characteristics $\tilde{y}$ can be decomposed into a sum of four terms:

$$U_0(x, y) + \chi(\tilde{x}, y) + \xi(\tilde{y}, x) + \zeta(\tilde{x}, \tilde{y}).$$

The first one describes the average taste for observed product characteristics among consumers of a given observed type; the second one allows for unobserved variation in taste for observed characteristics; the third one allows for unobserved product effects; and the fourth one is the idiosyncratic term. Assumption (S) rules out this last term. Its strength depends on the quality of the data. There is clear evidence, for instance, that American couples produce more surplus when the partners have similar religious background. If religious affiliation is not in our dataset, then assumption (S) will not hold. Physical characteristics are a more subtle case. If the dataset contains no information on them, then assumption

(S) does not rule out a preference for good looks, nor variation in such preferences; but it does rule out a correlation between the preference for good looks and one's own good looks.

While assumption (S) is already quite powerful, it still allows for very complex patterns: the covariance matrix of the $\chi(\tilde{x}, y)$ for a given man $\tilde{x}$ is an unwieldy object—not to mention other distributional characteristics. To go further, we need to add more restrictions on the specification of the components of the idiosyncratic surplus $\chi(\tilde{x}, y)$ and $\xi(\tilde{y}, x)$.

1.4 Specifying the idiosyncratic surplus

Following Choo and Siow (2006) and Chiappori, Salanié, Tillman, and Weiss (2008), we introduce the following assumption[2]:

Assumption GUI: Gumbel Unobserved Interactions

1. The distributions of observed types P and Q are discrete, with probability mass functions $p(x)$ and $q(y)$

2. There are an infinite number of individuals with a given observable type in the population

3. Fix the observable characteristics x of a man, and let $\left(y^1, ..., y^{n_Q}\right)$ be the possible values of the observable characteristics of women. Then the preference shocks $\chi\left(\tilde{x}, y^1\right), ..., \chi\left(\tilde{x}, y^{n_Q}\right)$ are distributed as n_Q independent and centered Gumbel (type-I extreme value) random variables with scale factor σ_1;

   similarly,

4. Fix the observable characteristics y of a woman, and let $\left(x^1, ..., x^{n_P}\right)$ be the possible values of the observable characteristics of men. Then the preference shocks $\xi\left(\tilde{y}, x^1\right), ..., \xi\left(\tilde{y}, x^{n_P}\right)$ are distributed as n_P independent and centered Gumbel random variables with scale factor σ_2.

[2] We define the scale factor to be 1 for the standard type-I extreme value distribution, which has variance $\pi^2/6$; thus e.g. χ has variance $\sigma_1^2\pi^2/6$.

(GUI) underlies the standard multinomial logit model of discrete choice. We use it for the Independence of Irrelevant Alternatives property: without it, the odds ratio of the probability that a man with observable type x ends up in a match with a woman of observable type y rather than with z would also depend on the types of other women, and the model would become unmanageable.

This assumption has well-known limitations. The first one is that it does not extend directly to continuous choice. We are currently exploring alternative specifications that would allow us to deal with continuous characteristics. It also restricts both heteroskedasticity and correlation patterns. We discuss extensions in section 8.

Finally, part 2 of assumption (GUI) is made strictly for notational simplicity: it allows us to replace averages with expectations. If there are, say a finite number m of members of each observed type, then our main results in the next two sections only hold asymptotically in m. When we describe our estimators in section 6, we of course take into account the fact that we only have a finite sample.

Under assumptions (O), (S), and (GUI), the model is fully parametrized; its parameters can be collected in a vector

$$\theta = (\Phi, \sigma_1, \sigma_2),$$

where Φ is the observable surplus function and σ_1 (resp. σ_2) is the scale factor of the unobservable characteristics of the men (resp. of women). Without loss of generality, all components of θ can be multiplied by any positive number; hence we shall need to impose some normalization on θ. We return to this later.

As we will see, the *total heterogeneity* $(\sigma_1 + \sigma_2)$ plays a key role in our results; thus we introduce a specific notation for it:

$$\sigma = \sigma_1 + \sigma_2.$$

2 Solving for the Optimal Matching

In this section we assume (O), (S), and (GUI), and we consider the problem of optimal matching:

$$\mathcal{W}(\theta) = \sup_{\tilde{\Pi} \in \mathcal{M}(\tilde{P},\tilde{Q})} E_{\tilde{\Pi}} \tilde{\Phi}\left(\tilde{X}, \tilde{Y}\right). \tag{2.1}$$

As the tilde signs in the formula suggest, none of the relevant quantities is observed; our main aim in this section is to prove that the formula can be rewritten entirely in terms of observable quantities, making inference possible. We examine first the dual, and then the primal version of the problem.

2.1 The Dual

Let us provide some intuition before we state a formal theorem. Under (O), (S) and (GUI), standard formulæ for the multinomial logit model give the expected utility of a man of observable type x at the optimal matching:

$$E\left[\max_y \left(U(x,y) + \chi(\tilde{X}, y)\right) | X = x\right] = \sigma_1 \log \sum_y \exp\left(U(x,y)/\sigma_1\right).$$

Therefore the expected social surplus from the optimal matching is simply[3] (adding the equivalent formula for women of observable type y):

$$\sigma_1 E_P \log \sum_y \exp(U(X,y)/\sigma_1) + \sigma_2 E_Q \log \sum_x \exp(V(x,Y)/\sigma_2).$$

Now recall that $U(x,y)$ is the mean utility of a man with observable type x who ends up being matched to a woman with observable type y at the optimum. As in the general development of the theory of matching, U is the value of the multiplier of the population constraints; and as such, it (along with V) is the unknown function in the dual program

[3]Since this formula may not be entirely transparent, we develop one term below:

$$E_P \log \sum_y \exp(U(X,y)/\sigma_1) = \sum_x p(x) \log \sum_y \exp\left(U(x,y)/\sigma_1\right).$$

in which the expression for the social surplus above is minimized over all U, V such that $U + V \geq \Phi$. We now state this as a theorem (proved in Appendix B):

Theorem 1 (Social welfare: dual version) *Assume (O), (S), and (GUI). Then*

$$\mathcal{W}(\theta) = \inf_{(U,V)\in A} \left(\sigma_1 E_P \log \sum_y \exp(U(X,y)/\sigma_1) + \sigma_2 E_Q \log \sum_x \exp(V(x,Y)/\sigma_2) \right) \quad (2.2)$$

where the constraint set A is defined by the inequalities

$$\forall x, y,\ U(x,y) + V(x,y) \geq \Phi(x,y).$$

At an optimal matching, men with observable type x will be found in matches with women with observable types y such that $U(x,y) + V(x,y) = \Phi(x,y)$. The expected utility of men with observable type x matched with women of observable type y is $U(x,y)$.

2.2 The Primal

Theorem 1 also has a primal version, of course; and it is in fact our most useful result, as it will lead directly to an empirical strategy. While deriving the theorem takes a bit more work (again, see Appendix B), the intuition is simple. First, if there were no unobserved heterogeneity (with σ close to zero) the optimal matching would coincide with the optimal observable matching Π, which solves

$$\mathcal{W}(\theta) = \sup_{\Pi\in\mathcal{M}(P,Q)} E_\Pi \Phi(X,Y).$$

Going to the polar opposite, in the limit when σ goes to infinity only unobserved heterogeneity would count; and since it is just noise, the optimal matching would simply assign partners randomly, yielding the product measure $P \otimes Q$.

As it turns out, when σ takes any intermediate value the optimal matching maximizes a weighted sum of these two extreme cases:

Theorem 2 (Social welfare: primal version) *Under the assumptions of Theorem 1*

$$\mathcal{W}(\theta) = \sup_{\Pi \in \mathcal{M}(P,Q)} \left(\sum_{x,y} \pi(x,y) \Phi(x,y) - \sigma I(\Pi) \right) + \sigma_1 S(Q) + \sigma_2 S(P), \tag{2.3}$$

where $S(P)$ *and* $S(Q)$ *are the* entropies *of* P *and* Q *given by*

$$S(P) = -\sum_x p(x) \log p(x); \quad \textit{and } S(Q) = -\sum_y p(y) \log p(y);$$

and $I(\Pi)$ *is the* mutual information *of the joint distribution* Π *, given by*

$$I(\Pi) = \sum_{x,y} \pi(x,y) \log \frac{\pi(x,y)}{p(x)q(y)}.$$

The mutual information $I(\Pi)$ is just the Kullback-Leibler divergence of Π from the independent product $P \otimes Q$ to Π. It is easy to see that I is a strictly convex function of Π. Moreover,

$$I(\Pi) = S(P) + S(Q) - S(\Pi);$$

and since Π has marginals P and Q,

$$0 \leq S(\Pi) \leq S(P) + S(Q),$$

so that we also have

$$0 \leq I(\Pi) \leq S(P) + S(Q).$$

The left hand-side is an equality at any pure matching (when all π's are 0 or 1), and the right hand-side inequality becomes an equality when where $\Pi = P \otimes Q$.

Mutual information is a measure of the covariation of types x and y. Now $P \otimes Q$ is the independent product of P and Q, which corresponds to a completely random matching $\Pi = P \otimes Q$. Thus a large positive $I(\Pi)$ indicates that the matching Π induces strong correlation across types; $I(\Pi) = S(P) + S(Q)$ if and only if $\Pi = P \otimes Q$. If σ is very large then the Theorem suggests that $I(\Pi)$ should be minimized, which can only occur for the independent matching $\Pi = P \otimes Q$; whereas if σ is negligible then Π should be chosen so as to maximize the expected *observable* surplus $E_\Pi \Phi(X,Y)$. This corroborates the intuition given earlier.

The optimal matchings coincide with the solutions to this maximization problem. Since we only observe the realized Π over observable variables, Theorem 2 defines the empirical content of the model: a combination of the parameters $\theta = (\Phi, \sigma_1, \sigma_2)$ is identified if and only if the solution Π depends non-trivially on it.

We already knew that θ can be rescaled by any positive constant without altering the solution. We can now go one step further: while all components of θ figure in this theorem, σ_1 and σ_2 only enter through their sum σ (the terms $\sigma_1 S(Q)$ and $\sigma_2 S(P)$ do not depend on Π and therefore do not help for identification.) Thus and as announced, σ_1 and σ_2 are not separately identified: only the total heterogeneity σ is.

Accordingly, we redefine the parameter vector θ as

$$\theta = (\Phi, \sigma)\,.$$

Taking the limit when $\sigma \longrightarrow 0$ in Theorems 1 and 2 and denoting $\mathcal{W}_0(\Phi) \equiv \mathcal{W}(\Phi, 0)$, we obtain as a corollary the classical duality of optimal matching:

Corollary 1 (Homogeneous social welfare) *Assume (O); then*

a) The value of the social optimum when $\theta = (\Phi, 0)$ is given both by

$$\mathcal{W}_0(\Phi) = \max_{\Pi \in \mathcal{M}(P,Q)} \sum_{x,y} \pi(x,y)\Phi(x,y)\,, \tag{2.4}$$

and by

$$\mathcal{W}_0(\Phi) = \inf_{(u,v)\in A^0} \left(\sum_x p(x)u(x) + \sum_y q(y)v(y) \right) \tag{2.5}$$

where the constraint set A^0 is given by

$$\forall x, y, \;\; u(x) + v(y) \geq \Phi(x,y)\,;$$

A matching $(X, Y) \sim \Pi$ is optimal for Φ if and only if the equality

$$u(X) + v(Y) = \Phi(X, Y)$$

holds Π-almost surely, where u and v solve the optimization problem (2.5).

Since all men with observable characteristics x have the same tastes in the homogeneous limit, they all obtain the same utility at the optimum. The utility $U(x, y)$ becomes a function of x only, which we denoted $u(x)$ above; and this is just the Lagrange multiplier on the population constraint

$$\sum_y \pi(x, y) = p(x)$$

which is implicit in the notation $\Pi \in \mathcal{M}(P, Q)$.

3 Nonparametric Identification

The results in the previous sections give a very useful description of the optimal matchings, and they show that σ_1 and σ_2 cannot be identified separately. On the other hand, we have not provided a proof of identification of the remaining parameters yet. We now set out to do so.

First note that if $\sigma > 0$ and since mutual information I is strictly convex, the objective function in Theorem 2 is strictly concave. Thus the optimal observable matching maximizes a strictly concave function over a compact convex set, and it must be unique[4].

Now remember that given assumptions (O) and (S), there exist two functions $U(x, y) + V(x, y) = \Phi(x, y)$ such that the optimal matching obtains when man $\tilde{x}$ maximizes $U(x, y) + \chi(\tilde{x}, y)$ over y and woman $\tilde{x}$ maximizes $V(x, y) + \xi(\tilde{y}, x)$ over x. Now if π is the observable component of an optimal matching, and given assumption (GUI),

$$U(x, y) = \sigma_1 \log \pi(x, y) + u(x),$$

where

$$u(x) = \sigma_1 \log \sum_y \exp\left(\frac{U(x, y)}{\sigma_1}\right).$$

In the literature on discrete choice, u is called the *inclusive value*: here $u(x)$ is the expected utility of a man of observed type x on the marriage market. Similarly,

$$V(x, y) = \sigma_2 \log \pi(x, y) + v(y).$$

[4]Related results are given in Decker, Stephens, and McCann (2009).

Now U and V depend on θ and are not easy to characterize as we will see; but we know that they must sum up to Φ, so that

$$\Phi(x,y) = \sigma \log \pi(x,y) + u(x) + v(y).$$

In this formula u and v still depend on θ in a complex way; but they only appear in terms that depend only on characteristics of one partner. This implies that the surplus function Φ is identified up to an additive function of the form $a(x) + b(y)$.

To state this more formally, define the *cross-difference operator* as

$$\Delta_2 F(x,y;x',y') = \big(F(x',y') - F(x',y)\big) - \big(F(x,y') - F(x,y)\big),$$

for any function F of (x,y). Then:

Theorem 3 (Cross-differences are identified up to scale) *Assume (O), (S), and (GUI). Take any $\theta = (\Phi, \sigma_1, \sigma_2)$ with $\sigma = \sigma_1 + \sigma_2 > 0$. Then*

1. *There exists a unique observable matching π which maximizes the social welfare (2.3).*

2. *There exists a unique 4-tuple (π, u, v, c) that solves the following system:*

$$\begin{cases} \pi(x,y) \equiv p(x)\, q(y) \exp\left(\frac{\Phi(x,y) - u(x) - v(y) - c}{\sigma}\right), \\ \pi \in \mathcal{M}(P,Q) \\ E_P u(X) = E_Q v(Y) = 0. \end{cases} \tag{3.1}$$

   $u(x)$ and v are both finite-valued functions, and the constant c coincides with the value of the social welfare $c = \mathcal{W}(\theta)$.

3. *The probability π defined in 2. coincides with the optimal observable matching defined in 1.*

4. *As a consequence, $\Delta_2 \Phi \equiv \sigma \Delta_2 \log \pi$; and this is a necessary and sufficient condition for θ to rationalize π.*

5. At the optimal matching, each possible match has positive probability:

$$0 < \pi(x, y) < 1 \ \textit{ for all } x, y \textit{ such that } p(x)q(y) > 0.$$

Given Theorem 3, the complementarity of various components of the observable types (x, y) of the partners can be tested directly on $\log \pi$, since $\Delta_2 \log \pi$ and $\Delta_2 \Phi$ have the same sign. Moreover, the relative strengths of complementarities along several dimensions (say education and income on example (ER)) at a point (x, y) can be estimated by evaluating $\Delta_2 \log \pi$ for values of (x', y') that differ from (x, y) along these dimensions.

These results are reminiscent of those in Fox (2009), although we obtained them under quite a different set of assumptions: we do not use variation across subpopulations, neither does his rank-order condition apply to our model. Note also that when specialized to one-dimensional types, our result yields that of Siow (2009), who tests complementarity of the surplus function by examining log-supermodularity of the match distribution.

Theorem 3 immediately gives us an estimator of the observable joint surplus function Φ:

$$\hat{\Phi}(x, y) = \log \hat{\pi}(x, y),$$

with $\hat{\pi}$ an estimator of π. This estimator corresponds to an assumed $\sigma = 1$; the last part of the Theorem shows that multiplying it by any positive factor (σ) and adding any pair of functions of x and of y would also yield a perfectly valid estimator. The positive scale factor σ is obviously irrelevant; the indeterminacy up to additive functions of x and y may seem more surprising. These additive functions represent the expected utilities of men of observed type x and women of type y on the marriage market, which could only be identified by relating the proportion of individuals to remain single to their types; since we are focusing on the population of matched individuals, we cannot identify them.

While $u(x)$ and $v(y)$ can be chosen arbitrarily, $U(x, y)$ and $V(x, y)$ are partially identified. To be more precise, recall the equations

$$U(x, y) = \sigma_1 \log \pi(x, y) + u(x) \quad \text{and} \quad V(x, y) = \sigma_2 \log \pi(x, y) + v(y);$$

the extent to which U and V are identified is directly implied by the fact that π is identified but neither σ_1, nor σ_2, nor u or v are. Therefore only ratios of the form

$$\frac{U(x,y_1)-U(x,y_2)}{U(x,y_3)-U(x,y_4)}$$

are point identified, with the obvious analog statement for V. Note that this only makes sense if $\sigma_1 > 0$ and $\pi(x,y_3) \neq \pi(x,y_4)$; while the latter is directly testable, the former is not.

These results have another surprising consequence: if Φ, U, V rationalize the data, then for any $\mu = (\mu_1, \mu_2) \gg 0$, the linear combination

$$\Phi_\mu(x,y) \equiv \mu_1 U(x,y) + \mu_2 V(x,y)$$

and the functions $U_\mu \equiv \mu_1 U$, $V_\mu \equiv \mu_2 V$ also rationalize the data. This is a by-product of assumptions (S) and (GUI).

4 Empirical and computational strategies

Theorem 3 and its corollary immediately suggest a very simple nonparametric approach. In this discrete case, a nonparametric estimator $\hat{\pi}_N(x,y)$ is readily obtained, by counting the proportion of matches between a man of characteristics x and a woman of characteristics y. We can pick arbitrary functions $a(x)$ and $b(y)$ and a number $\sigma > 0$ and define

$$\hat{\Phi}_N(x,y) = \sigma \log \hat{\pi}_N(x,y) + a(x) + b(y),$$

without any reference to basis functions—imposing $\sigma = 1$ on the way.

A nonparametric approach will often be unsuitable for applied purposes, when the aim is to test for stylized facts about the matching patterns. We could, however, take this nonparametric estimator as the basis for a parametric approach.

4.1 Parametric approach

Suppose for instance that the researcher specifies a parametric family of observable surplus functions ($\Phi(x, y; \beta)$). Then he may choose $(\hat{\beta}, \hat{\sigma})$ to minimize a distance

$$\|\hat{\pi}_N - \pi_{\beta,\sigma}\|,$$

with $\pi_{\beta,\sigma}$ the optimal matching given by Theorem 2. Since the problem is invariant to rescaling, some normalization (e.g. imposing the value of σ) will be required, unless it is already imposed by the parametrization.

An alternative approach would start from a parametric specification of the surplus function as above and choose (β, σ) to maximize the likelihood function

$$\sum_{i=1}^{N} \log \pi_{\beta.\sigma}(x_i, y_i)$$

where each observation i is a couple. This amounts, of course, to maximizing

$$\sum_{x,y} \hat{\pi}_N(x, y) \log \pi_{\beta,\sigma}(x, y),$$

where the sum now runs over all (x, y) cells.

One problem with these two-step approaches is that they require solving for the optimal matching for potentially large populations, and a large number of parameter vectors during optimization. This may seem to be a forbidding task: there exist well-known algorithms to find an optimal matching, and they are reasonably fast; but with large populations the required computer resources may still be large. Fortunately, it turns out that introducing (our type of) heterogeneity actually makes computing optimal matchings much simpler.

4.2 Computation

Choose a parameter vector $\theta = (\Phi, \sigma)$ and return to the characterization of optimal matchings in equation 2.3. Dividing by σ and taking the logarithm, optimal matchings can also be obtained by solving the following minimization program:

$$\min_{\Pi \in \mathcal{M}(P,Q)} \sum_{x,y} \pi(x, y) \log \frac{\pi(x, y)}{p(x)q(y) \exp(\Phi(x, y)/\sigma)}.$$

Now define a set of probabilities r by

$$r(x,y) = \frac{p(x)q(y)\exp(\Phi(x,y)/\sigma)}{\sum_{x,y} p(x)q(y)\exp(\Phi(x,y)/\sigma)};$$

and note that given any choice of parameters θ and known marginals (p,q), the probability r itself is known.

Determining the optimal matchings therefore boils down to finding the joint probabilities π with known marginals p and q which minimize the Kullback-Leibler distance to r:

$$\sum_{x,y} \pi(x,y)\log\frac{\pi(x,y)}{r(x,y)}. \tag{4.1}$$

Equivalently, we are looking for the Kullback-Leibler projection of r on $\mathcal{M}(P,Q)$.

This is a well-known problem in various fields, and algorithms to solve it have been around for a long time. National accountants, for instance, use RAS algorithms to fill cells of a two-dimensional table whose margins are known; here the choice of r reflects prior notions of the correlations of the two dimensions of the table. These RAS algorithms belong to a family called Iterative Projection Fitting Procedures (IPFP). They are very fast, and are guaranteed to converge under weak conditions. We only describe the application of IPFP to our model here; we direct the reader to Rüschendorf (1995) for more information.

The intuition of equation 4.1 is quite clear: the random matching, which is optimal when σ is very large, has $\pi(x,y) = p(x)q(y)$. For smaller σ' s the probability of a match between x and y must increase with the surplus it creates, $\Phi(x,y)$; and given our assumption (GUI) on the distribution of unobserved heterogeneity, it should not come as a surprise that the corresponding factor is multiplicative and exponential.

To describe the algorithm, we split π into[5]

$$\pi(x,y) = r(x,y)\exp(-(u(x)+v(y))/\sigma).$$

The functions u and v of course will only be determined up to a common constant. The algorithm iterates over values (u^k, v^k). We start from $u^0 \equiv -\sigma\log p$ and $v^0 \equiv 0$. Then at

[5]It can be shown that at the optimum $\pi(x,y) = 0$ where $r(x,y) = 0$.

step $(k+1)$ we compute

$$\exp(-v^{k+1}(y)/\sigma) = \frac{q(y)}{\sum_x r(x,y)\exp(-u^k(x)/\sigma)}$$

and

$$\exp(-u^{k+1}(x)/\sigma) = \frac{p(x)}{\sum_y r(x,y)\exp(-v^{k+1}(y)/\sigma)}.$$

Two remarks are in order here: first, we could just as well start from $u^0 \equiv 0$ and $v^0 = -\sigma \log q$ and modify the iteration formulæ accordingly. Second and just as in other Gauss-Seidel algorithms, it is important to update one component based on the other updated component: the right-hand sides have u^k and v^{k+1}.

If (u, v) is a fixed point of the algorithm, then

$$\frac{\pi(x,y)}{p(x)q(y)} = \exp\left(\frac{\Phi(x,y) - u(x) - v(y)}{\sigma}\right).$$

Comparing this formula to Theorem 3 shows that $u(x)$ and $v(y)$ have a simple interpretation: they represent (up to a common additive constant) the expected utilities of a man of observable characteristics x and of a woman of observable characteristics y.

Thus the IPFP algorithm gives us not only the optimal matching, but also these expected utilities. Of course, their values depend on the normalization of $\theta = (\Phi, \sigma)$, both because of the scale factor σ and because Φ could be translated by a sum of a function of x and a function of y without changing the optimal matching.

The formulæ above can be simplified further. Given data on N couples, the marginal p assigns $1/N$ probability to each of $(x_1, \ldots, x_N)$, and similarly for women. Define a matrix Ψ by $\Psi_{ij} = \exp(\Phi(x_i, y_j)/\sigma)$, and vectors $a_i^k = \exp(-u^k(x_i)/\sigma)$, $b_j^k = \exp(-v^k(y_j)/\sigma)$. Then we end up with the shockingly simple and inexpensive formulæ for the IPFP algorithm:

$$b^{k+1} = \frac{N}{\Psi' a^k} \quad \text{and} \quad a^{k+1} = \frac{N}{\Psi b^{k+1}}.$$

Once the iterations have converged to some (a, b), the optimal matching is simply

$$\pi_{ij} = \frac{1}{N^2}\Psi_{ij} a_i b_j.$$

As will be shown in our application in section 7, the IPFP algorithm is remarkably fast. Yet it cannot substitute for the limitations of the data: even with large datasets, nonparametric estimation must face the fact that there are many possible (x, y) cells. Take the education-income example, and assume that we distinguish three levels of education and five income classes. Then x and y can each take 15 different values, and there are $15^2 = 225$ cells. For some of these cells, the estimator $\hat{\pi}_N$ will be zero; but more importantly, it will likely be rather imprecise in general, and so will any parametric estimator obtained by the minimum distance method described above.

Among all parametric specifications of the observable surplus function Φ, linear models are the most natural. As we will see in the next section, they also yield both illuminating insights into the properties of the optimal matchings and an alternative, very appealing estimation method.

5 The Semilinear Case

In this section, we assume that the analyst has chosen K *basis assorting functions*

$$\phi^1(x, y), ..., \phi^K(x, y);$$

and that she specifies the observable surplus function $\Phi_\Lambda(x, y)$ as a linear combination of these basis assorting functions, with unknown *assorting weights* $\Lambda \in \mathbb{R}^K$:

Model (SLOI): Semilinear Observable Interactions. The analyst specifies the observable surplus function as

$$\Phi_\Lambda(x, y) = \sum_{k=1}^{K} \Lambda_k \phi^k(x, y) \tag{5.1}$$

where the sign of each Λ_k is unrestricted, and not all are zero.■

Note that in the discrete case which we focus on in this paper, this specification is only restrictive if K is small enough. Indeed, choosing $K = n_P \times n_Q$ and the family of basis

functions

$$\phi^{ij}(x,y) = 1_{\{x=x_i, y=y_j\}}$$

for $i = 1, \ldots, n_P$ and $j = 1, \ldots, n_Q$ generates all possible surplus functions.

In most applications, the analyst will want to choose a value of K that is much smaller than $n_P \times n_Q$, and so (SLOI) does restrict the specification. To return to the education/income example (ER): we could for instance assume that a match between man m and woman w creates a surplus that depends on the similarity of the partners in both education and income dimensions. The corresponding specification would be (with education levels $E = (D, G)$ coded as $(0, 1)$):

$$\Phi(x_m, y_w) = \sum_{e_m=0,1; e_w=0,1} \Lambda_{e_m,e_w} \mathbb{1}(E_m = e_m, E_w = e_w) + \sum_{i=1,\ldots,n_r; j=1,\ldots,n_r} \Lambda_{ij} \mathbb{1}(R_m = i, R_w = j).$$

This specification only has $(n_r^2 + 4)$ parameters, while an unrestricted specification would have $4n_r^2$. Such an unrestricted specification would for instance allow the effect of matching partners in income class 3 to depend on both of their education levels.

An even more restrictive, "diagonal" specification would be

$$\Phi(x_m, y_w) = \sum_{e=0,1} \Lambda_e^E \mathbb{1}(E_m = E_w = e) \; + \sum_{i=1,\ldots,n_r} \Lambda_i^R \mathbb{1}(R_m = R_w = i).$$

In this last form, it is clear that the relative importance of the Λ's reflects the relative importance of the criteria. Thus Λ_i^R measures the preference for matching partners who are both in income class i, while Λ_0^E measures the preference for matching dropouts. The relative values of these numbers indicate how social preferences value complementarity of incomes of partners more, relative to complementarity in educations.

In model (SLOI), the set of parameters becomes $\theta = (\Lambda, \sigma)$. If the model is correctly specified, then Theorem 3 gives us an estimator of the assorting weights Λ and the total heterogeneity σ[6]. In fact, the cross-difference operator is linear and so under (SLOI),

$$\Delta_2 \log \pi = \frac{\Delta_2 \Phi}{\sigma} = \sum_{k=1}^{K} \frac{\Lambda_k}{\sigma} \Delta_2 \phi^k;$$

[6]Recall that σ_1 and σ_2 are not separately identified.

if the cross-differences of the ϕ^k are linearly independent, as they are in both education-income examples above, then observing π gives us the Λ's (along with overidentifying restrictions.) This is a very weak requirement; having linearly dependent basis functions would indeed be a modeling mistake. However, this estimator is likely to be very imprecise, for the reasons discussed in section 4.

The semilinear structure underlying model (SLOI) makes it very easy to analyze optimal matchings, as all inference can be based on only $(K+1)$ numbers. These numbers are sufficient statistics for testing the model; and if we cannot reject that it is well-specified (so that the true data-generating process satisfies (SLOI) for the set of basis functions chosen by the analyst), then K of these numbers form a sufficient statistic for estimating $\theta = (\Lambda, \sigma)$. We now set out to substantiate these claims.

5.1 The Covariogram

Consider a hypothetical observed matching Π; under this matching, the basis functions have expected values

$$C^k(\Pi) = \sum_{x,y} \pi(x,y)\phi^k(x,y)\,.$$

We call each C^k a *covariation*. Take the diagonal (ER) example; then for the matching considered,

- $C^{e_m,e_w}(\Pi)$ is the proportion of matches in which the man has education e_m and the woman has education e_w;
- $C^{ij}(\Pi)$ is the proportion of matches in which the man is in income class i and the woman is in income class j.

Random matching, as represented by $\Pi_\infty = P \otimes Q$, plays a special role in our analysis, as it obtains in the limit when heterogeneity becomes very large. We denote the corresponding covariations C^k_∞. At the polar opposite is the matching Π_0 which obtains in the homogenous

limit $\sigma = 0$; we denote the implied covariations $C_0^k(\Lambda)$. Note that C_∞ does not depend on Λ, but C_0 does.

5.2 Feasible summaries

Define the function

$$\mathcal{Z}(C, I, \Lambda, \sigma) = \Lambda \cdot C - \sigma I.$$

We know from Theorem 2 that if model (SLOI) is correctly specified, then for some $\theta = (\Lambda, \sigma)$ the observable optimal matching maximizes $\mathcal{Z}(C(\Pi), I(\Pi), \Lambda, \sigma)$ over $\Pi \in \mathcal{M}(P, Q)$. As in previous sections, we denote $\mathcal{W}(\Lambda, \sigma)$ the value of this program.

Thus in model (SLOI) the vector $(C(\Pi), -I(\Pi))$ summarizes all the relevant information about a matching Π. We call each such vector a *matching summary*; matching summary vectors belong to $\mathbb{R}^K \times \mathbb{R}^-$.

Given an observed matching with couples $(x_i, y_i)_{i=1}^N$, it is of course very easy to estimate the associated summary:

$$\hat{C}_N^k = \sum_{i=1}^N \phi^k(x_i, y_i) \quad \text{and} \quad \hat{I}_N = \sum_{i=1}^N \log \frac{\hat{\pi}_N(x_i, y_i)}{\hat{p}_N(x_i)\hat{q}_N(y_i)},$$

given estimators $\hat{\pi}_N, \hat{p}_N$ and $\hat{q}_N$ of the joint and marginal distributions of types.

Given population distributions P and Q, we define the *set of feasible summaries* $\mathcal{F}$ as the set of summary vectors $(C, -I)$ that are generated by some feasible matching $\Pi \in \mathcal{M}(P, Q)$, that is

$$\mathcal{F} = \left\{ (C, -I) \in \mathbb{R}^K \times [-S(P) - S(Q), 0] : \exists \Pi \in \mathcal{M}(P, Q),\ C^k = C^k(\Pi),\ I = I(\Pi) \right\}$$

The projection of $\mathcal{F}$ on its first K dimensions is of particular interest to us. We accordingly define the *covariogram* $\mathcal{F}_c$ as the set of covariations C that are implied by some feasible matching; that is,

$$\mathcal{F}_c = \left\{ C : \exists \Pi \in \mathcal{M}(P, Q),\ C^k = C^k(\Pi) \right\}.$$

Covariograms provide us with a nice graphical representation of the properties of a matching. Figure 1 illustrates their relevant properties, and the reader should refer to it as we go along. To fit it within two dimensions, we assume that there are only two basis functions; e.g. in the (ER) example we could have

$$\Phi(E_m, E_w, R_m, R_w) = \Lambda_1 \mathbb{1}(E_m = E_w) + \Lambda_2 \mathbb{1}(R_m = R_w),$$

so that Λ_1 (resp. Λ_2) measures the preference for assortative matching on educations (resp. income classes.)

Proposition 1 *The sets $\mathcal{F}$ and $\mathcal{F}_c$ are nonempty closed convex sets, and their support functions are $(\Lambda, \sigma) \to \mathcal{W}(\Lambda, \sigma)$ and $\Lambda \to \mathcal{W}(\Lambda, 0)$, respectively.*

As will soon become clear, the boundaries of the convex sets $\mathcal{F}$ and $\mathcal{F}_c$ have special significance in our analysis. For now, let us simply note that the boundary of $\mathcal{F}_c$ exhibits kinks when these distributions of characteristics are discrete—which is always the case in our setting. The reason for these kinks is that in the discrete case, the optimal matching for homogenous types is generically stable under a small perturbation of the assorting weights Λ; starting from almost every Λ's, a small change in Λ leaves covariations unchanged. Any such value of Λ generates a covariation vector on a vertex of the polytope. On the other hand, there exist a finite number of values of Λ where the optimal matching problem has multiple solutions, with corresponding multiple covariations; each such value of Λ generates a facet of the polytope. This is shown on Figure 1 with all $\lambda = \Lambda_2/\Lambda_1$ in an interval $[\lambda_i, \lambda_{i+1}]$ generating the same covariations in the homogeneous case. (We will describe the other objects on Figure 1 as we go along.)

The heterogeneous case $\sigma > 0$ is much better-behaved. We already know from Theorem 3 that the optimal matching is unique and cannot be pure; we will see in Proposition 5 that it is also a smooth function of the parameters, so that the kinks disappear as soon as there is any positive amount of heterogeneity.

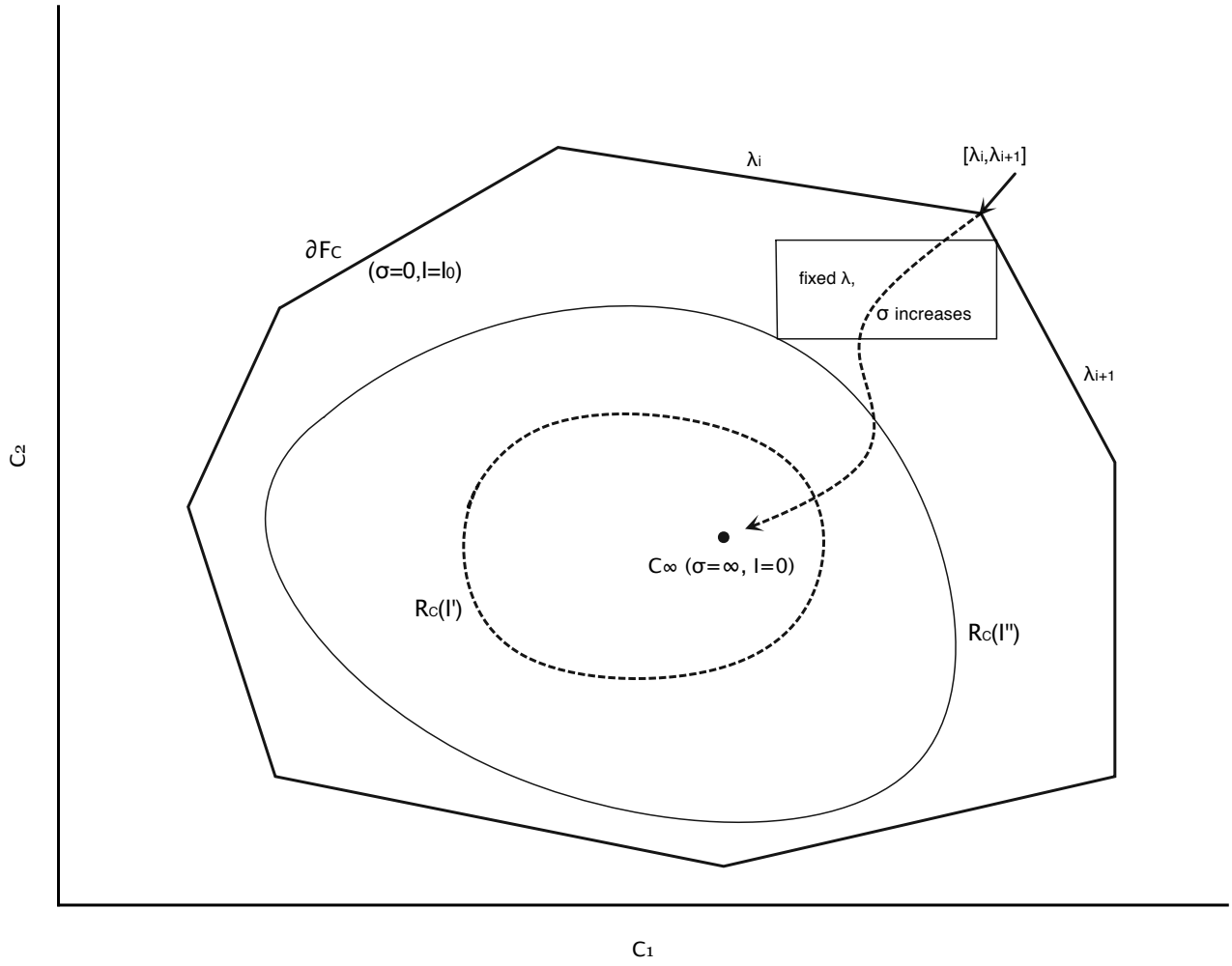

Figure 1: The covariogram and related objects

5.3 Rationalizable Summaries

For any $\theta = (\Lambda, \sigma)$ with $\sigma > 0$, denote $\Pi(\theta)$ the corresponding optimal matching under (SLOI): $\Pi(\theta)$ maximizes $\mathcal{Z}(C(\Pi), I(\Pi), \theta)$ over $\Pi \in \mathcal{M}(P, Q)$. And define

$$\bar{C}(\theta) = C(\Pi(\theta)), \quad -\bar{I}(\theta) = -I(\Pi(\theta))$$

the corresponding summary. These definitions are valid since by Theorem 3, the optimal matching is unique for $\sigma > 0$. When $\sigma = 0$, for a finite number of values of Λ the optimal matching will not be unique, and so these functions are correspondences. To simplify notation, we treat them as functions except when it matters.

We now define the *set of rationalizable summaries* $\mathcal{R}$ as the set of $(K+1)$-uples $(C, -I)$ such that $C = \bar{C}(\theta)$ and $I = \bar{I}(\theta)$ for some parameter values $\theta = (\Lambda, \sigma)$. Equivalently, since $\Pi(\theta)$ maximizes $\mathcal{Z}(C(\Pi), I(\Pi), \theta)$ and achieves the value $\mathcal{W}(\theta)$:

$$\mathcal{R} = \left\{(C, -I) : \exists\, (\Lambda, \sigma) \in \mathbb{R}^K \times \mathbb{R}^+,\ \Lambda \cdot C - \sigma I = \mathcal{W}(\Lambda, \sigma)\right\}.$$

Obviously, rationalizable summaries are feasible: $\mathcal{R} \subset \mathcal{F}$. In fact, since $\mathcal{W}$ is the support function of $\mathcal{F}$,

Proposition 2 *$\mathcal{R}$ is the frontier of $\mathcal{F}$.*

Proposition 2 points towards a specification test: given a choice of basis functions (ϕ^k) and an observed matching $(\hat{\pi}_N)$,

1. construct the set of feasible summaries $\hat{\mathcal{F}}_N$ (which depends on $\hat{p}_N$ and $\hat{q}_N$)
2. compute the observed summary $(\hat{C}_N, \hat{I}_N)$
3. and use its distance to the frontier of $\hat{\mathcal{F}}_N$ to compute a test statistic.

While this is correct in principle, it is a fairly cumbersome way to proceed. We now turn to more practical approaches to inference.

5.4 Mutual information level sets

Whether model (SLOI) is well-specified or not, the frontier of $\mathcal{F}$ can itself be decomposed into sections on which the mutual information I is constant. To see this, consider a covariation vector C in the covariogram $\mathcal{F}_c$. By definition, there exists a feasible matching Π such that $C = C(\Pi)$. Since the set of such matchings is defined by linear equalities and mutual information $I(.)$ is a strictly convex function, we can define

$$I_r(C) := \min\{I(\Pi) : \Pi \in \mathcal{M}(P,Q) \text{ and } C = C(\Pi)\};$$

we call $I_r(C)$ the *rationalizing mutual information* of C.

By construction, the point $(C, -I_r(C))$ must be on the frontier of $\mathcal{F}$. Since $\mathcal{W}$ is the support function of $\mathcal{F}$, it follows that there exists a $\theta = (\Lambda, \sigma)$ such that

$$\mathcal{W}(\Lambda, \sigma) = \Lambda \cdot C - \sigma I_r(C);$$

and for this value of the parameter vector,

$$C = \bar{C}(\theta) \text{ and } I_r(C) = \bar{I}(\theta).$$

If $\sigma > 0$, then taking the Legendre-Fenchel transform and using the homogeneity of $\mathcal{W}$, the rationalizing mutual information is also

$$I_r(C) = \sup_{\lambda}\{\lambda \cdot C - \mathcal{W}(\lambda, 1)\}; \tag{5.2}$$

therefore I_r is a strictly convex and C^1 function.

Conversely, for any mutual information $0 \leq I \leq S(P) + S(Q)$ we define the *set of rationalizable covariations* by

$$\mathcal{R}_c(I) = I_r^{-1}(I).$$

Thus each set $\mathcal{R}_c(I)$ is a level set of the rationalizable mutual information function I_r.

Note the two limiting cases: when mutual information I is zero (corresponding to random matching), then $\mathcal{R}_c(0) = \{C_\infty\}$, where $C_\infty^k = E_{p\otimes q}\phi^k(X,Y)$. When $I = S(P)+S(Q)$, $\mathcal{R}_c(S(P)+S(Q))$ consists of the extreme points of the covariogram $\mathcal{F}_c$.

The following result sums up our results so far for the semilinear model:

Proposition 3 *Under (O), (S) and (GUI),*

a) The social welfare function $\mathcal{W}$ is positive homogeneous of degree one in $\theta = (\Lambda, \sigma)$. It is convex on $\mathbb{R}^K \times [0, +\infty)$ and it is strictly convex on its interior.

b) If $\sigma > 0$, the derivatives of $\mathcal{W}$ at $\theta = (\Lambda, \sigma)$ are $(\bar{C}(\theta), -\bar{I}(\theta))$.

c) The function $I_r(C)$ is C^1 on the interior of $\mathcal{F}_c$. Let $\theta = (\Lambda, \sigma)$ be such that $C = \bar{C}(\theta)$ and $I_r(C) = \bar{I}(\theta)$; then

$$\frac{\partial I_r}{\partial C^k} = \frac{\Lambda_k}{\sigma}.$$

d) For the homogeneous model with $\sigma = 0$, the function $\mathcal{W}_0 \equiv \mathcal{W}(., 0)$ has a subdifferential in Λ given by the set of K-uples

$$\partial \mathcal{W}_0 = \{C(\Pi)\}$$

generated by the matchings in $\Pi(\Lambda, 0)$. The boundary of $\mathcal{F}_c$ is constituted by the covariation vectors in $\bar{C}(\Lambda, 0)$.

5.5 Inference in the Semilinear Model

Our most important result on the semilinear model is that any feasible summary $(C, -I)$ can be rationalized by model (SLOI), provided only that $I = I_r(C)$; and that if C is in the interior of $\mathcal{F}_c$ then the corresponding assorting weights are unique, up to a scale factor.

We already proved that if $I = I_r(C)$, the summary $(C, -I)$ is rationalizable. Take a summary $(C, -I)$ such that $I \neq I_r(C)$. If it is a feasible summary, then by construction $I \geq I_r(C)$. Assume that the inequality is strict, and θ rationalizes the summary $(C, -I)$; then given $\sigma > 0$,

$$\mathcal{Z}(C, -I_r(C); \theta) > \mathcal{Z}(C, -I; \theta),$$

which contradicts the optimality of $\Pi(\theta)$ for θ. Thus $I = I_r(C)$ is a necessary and sufficient condition for $(C, -I)$ to be a rationalizable summary.

The following result, which sums up the relationships between the sets we introduced:

Proposition 4 *Under (O), (S), and (GUI),*

a) The sets $\mathcal{R}_c(I)$ are the frontiers of nested closed and convex sets that expand from $\{C_\infty\}$ to $\mathcal{F}_c$ as mutual information I increases from 0 to $S(P)+S(Q)$.

b) Any point C in the interior of $\mathcal{F}_c$ belongs to exactly one level set $\mathcal{R}_c(I)$, associated to the mutual information level $I = I_r(C)$.

c) For any such C, define

$$\lambda(C) = \frac{\partial I_r}{\partial C}(C);$$

then on the tangent space to $\mathcal{R}_c(I)$,

$$\frac{dC^i}{dC^j} = -\frac{\lambda_j(C)}{\lambda_i(C)} \quad \forall i,j = 1,\dots,K. \tag{5.3}$$

d) For any C in the interior of $\mathcal{F}_c$, the equations

$$C = \bar{C}(\theta) \quad \textit{and } I = \bar{I}(\theta)$$

have a solution if and only if $I = I_r(C)$; and then the set of solutions is the half-line

$$\theta = \sigma \times (\lambda(C), 1) \quad \textit{for } \sigma > 0.$$

e) If C is on the boundary of $\mathcal{F}_c$, then let $\lambda(C)$ be the normal cone (the set of vectors that are normal to $\partial\mathcal{F}_c$ in C); the set of solutions to the inclusion equation

$$C \in \bar{C}(\theta)$$

is the set of $\theta = (\lambda, 0)$ such that $\lambda \in \lambda(C)$.

Proposition 4 is illustrated on figure 1. Note that when we fix Λ and increase σ from 0 to $+\infty$, the summary vector $(C, -I)$ for the optimal matching moves continuously from $(C_0(\Lambda), -I_0(\Lambda))$ to $(C_\infty, 0)$; thus part a) tells us that increasing σ for given Λ moves us from a point on the boundary of $\mathcal{F}_C$ to C_∞. This is represented on Figure 1 by the dashed trajectory.

The interpretation of part c) is simplest when the matrix Λ is diagonal. With several dimensions for types, the optimal matching must sacrifice some covariation in one dimension to the benefit of some covariation in another. The implied sacrifice ratio, quite naturally, is exactly the ratio of the assorting weights along these dimensions. Take for instance the homogeneous case with only two characteristics, and set $\Lambda_{11} = 1$ and $\Lambda_{22} = \varepsilon$. Then the function $\varepsilon \to C^{11}(1, \varepsilon)$ is decreasing, and the function $\varepsilon \to C^{22}(1, \varepsilon)$ is increasing. Therefore, when one puts more weight on the second dimension, the covariation of the characteristics in the second dimension increases, while the covariation on the first dimension decreases. Quite intuitively, in the limit where all the weights are put on one dimension, the classical Beckerian theory of positive assortative matching obtains.

Proposition 4 has direct implications for identification. Neglecting sampling variation, let us observe π, p, q and therefore I. Then, given a model (SLOI), compute the observed covariations C and the function I_r. The above suggests an empirical strategy (assuming $\sigma > 0$):

- If $I \neq I_r(C)$, we reject model (SLOI);
- If $I = I_r(C)$, then we identify the parameters of (SLOI) as

$$\Lambda = \frac{\partial I_r}{\partial C}(C)$$

  up to a scale factor.

6 Parametric Inference

We now turn to the problem of parametric inference. Our data will consist of matched characteristics of N pairs $\{(x_1, y_1), ..., (x_N, y_N)\}$, and our null hypothesis is that they were generated by an optimal matching consistent with assumptions (O), (S), (GUI). Given a proposed specification for model (SLOI) with basis functions ϕ^k, and our estimates of the marginal distributions of types $\hat{P}_N$ and $\hat{Q}_N$, we would like to test for correct specification and to infer the values of Λ and σ which come closest to rationalizing the observed matching.

Our empirical strategy is based on the knowledge of the matching summaries $\left(\hat{C}, \hat{I}\right)$, which are the sufficient statistics for our model, or of just the covariations $\hat{C}$ when $\left(\hat{C}, \hat{I}\right)$ lies on the efficient frontier, that is $\hat{I} = I_r\left(\hat{C}\right)$. In either cases, positive homogeneity imposes the need for a normalization of the estimator $\left(\hat{\Lambda}, \hat{\sigma}\right)$. We assume throughout that $\sigma > 0$, which can be tested that checking whether $\hat{C}$ is on the frontier of $\mathcal{F}_c$.

The normalization rule

We choose to normalize (Λ, σ) by

$$\textbf{Normalization convention}\text{: } \sigma \bar{I}(\Lambda, \sigma) = 1, \tag{6.1}$$

although any other choice would do just as well. This one has the advantage that it relates heterogeneity σ and mutual information I in a natural way.

By construction, $I_r\left(\hat{C}\right) = \bar{I}(\lambda, 1) = \hat{I}$ if λ rationalizes the data. If the null of correct specification $\hat{I} = I_r(\hat{C})$ is not rejected, then Proposition 4 implies that

$$\hat{\lambda} = \frac{\partial I_r}{\partial C}\left(\hat{C}\right),$$

and given our normalization rule,

$$\hat{\Lambda} = \frac{\partial \log I_r}{\partial C}\left(\hat{C}\right). \tag{6.2}$$

Our general approach will be to identify the parameter value $\left(\hat{\lambda}, 1\right)$, and then rescale

$$\hat{\Lambda} = \frac{\hat{\lambda}}{\hat{I}}, \quad \hat{\sigma} = \frac{1}{\hat{I}}.$$

6.1 The Efficiency Bound

We start by computing the Fisher information bound of model (SLOI); then we will introduce a very simple estimator that achieves it.

Given a function h of two random variables (X, Y) which have joint cdf F, we write the two-way ANOVA decomposition

$$h(X, Y) = E_F h(X, Y) + a(X; h, F) + b(Y; h, F) + \varepsilon(X, Y; h, F),$$

with

$$a(X;h,F) = E_F\left(h(X,Y)|X\right) - E_F h(X,Y) \text{ and } b(Y;h,F) = E_F\left(h(X,Y)|Y\right) - E_F h(X,Y).$$

By construction, the *ANOVA residue* $\varepsilon(X,Y;h,F)$ has zero conditional means:

$$E_F\left(\varepsilon\left(X,Y;h,F\right)|X\right) = 0 \text{ and } E_F\left(\varepsilon\left(X,Y;h,F\right)|Y\right) = 0.$$

The following proposition will be our fundamental tool for inference. It states that the ANOVA residue of $\phi^k(X,Y)$ under the optimal matching $\Pi(\theta)$ is proportional to the score function $\frac{\partial \log \pi_\theta}{\partial \Lambda_k}$, where $\pi_\theta(x,y)$ denotes the proportion of matches between observed types x and y for the optimal matching in θ. We denote $\mathcal{D}$ the image of the interior of the covariogram $\mathcal{F}_c$ by $\partial I_r / \partial C$.

Proposition 5 (Score function) *Fix $\sigma > 0$. Under (O), (S), and (GUI) the likelihood function of model (SLOI) is infinitely differentiable with respect to Λ on $\sigma\mathcal{D}$, and its score function is given by*

$$\frac{\partial \log \pi_\theta}{\partial \Lambda_k}(x,y) = \frac{\varepsilon(x,y;\phi^k,\pi_\theta)}{\sigma}.$$

As a result, we get a very simple relationship between the Fisher information matrix, the ANOVA residues, and the Hessian of the social welfare function at fixed σ.

Proposition 6 (Fisher information matrix) *Under (O), (S), and (GUI), the Fisher information matrix of model (SLOI)*

$$\mathcal{I}^{kl}(\theta) := E\left(\frac{\partial \log \pi}{\partial \Lambda_k}(X,Y)\frac{\partial \log \pi}{\partial \Lambda_l}(X,Y)\right)$$

is proportional to the ANOVA residues on $\mathcal{D}$:

$$\mathcal{I}^{kl}(\theta) := \frac{E\left(\varepsilon(X,Y;\phi^k,\pi_\theta)\varepsilon(X,Y;\phi^l,\pi_\theta)\right)}{\sigma^2}. \tag{6.3}$$

Moreover, it is also proportional to the Hessian of $\mathcal{W}$ on $\mathcal{D}$:

$$\frac{\partial^2 \mathcal{W}}{\partial \Lambda_k \partial \Lambda_l}(\Lambda,\sigma) = \sigma \mathcal{I}^{kl}(\theta).$$

6.2 The Asymptotic of Covariations

Denote

$$\hat{\pi}_N(x,y) = \frac{1}{N}\sum_{n=1}^{N} 1\{x_n = x, y_n = y\}$$

the sample estimator of $\pi(x,y)$. Standard asymptotic theory ensures the distributional convergence

$$N^{1/2}\hat{\pi}_N(x,y) \Longrightarrow G(x,y)$$

where G is a Gaussian process such that $\sum_{x,y} G(x,y) = 0$, $var(G(x,y)) = \pi(x,y)(1-\pi(x,y))$, and for $(x,y) \neq (x',y')$

$$cov\left(G(x,y), G\left(x',y'\right)\right) = -\pi(x,y)\pi\left(x',y'\right).$$

The basis of our investigation will be the empirical moments of ϕ^k,

$$\hat{C}_N^k = \frac{1}{N}\sum_{n=1}^{N}\phi^k(x_n,y_n) = \sum_{x,y}\phi^k(x,y)\hat{\pi}_N(x,y).$$

Let C^k denote the expectation of $\phi^k(X,Y)$ under the joint limit distribution π of (X,Y). Then

$$\sqrt{N}\left(\hat{C}_N^k - C^k\right) \Longrightarrow \xi^k \tag{6.4}$$

where $\xi^k = \sum_{x,y}\phi^k(x,y)G(x,y)$. In particular,

$$cov\left(\sqrt{N}\left(\hat{C}_N^k - C^k\right), \sqrt{N}\left(\hat{C}_N^l - C^l\right)\right) = cov_\pi\left(\phi^k(X,Y), \phi^l(X,Y)\right)$$

for all $1 \leq k,l \leq K$.

6.3 Testing

Our testing strategy will be based on the comparison between

- $\hat{I}_N = \sum_{x,y}\hat{\pi}_N(x,y)\log\frac{\hat{\pi}_N(x,y)}{\hat{p}_N(x)\hat{q}_N(y)}$, which is the mutual information measured in the data, and

- $\hat{I}_r\left(\hat{C}_N\right)$ which is the rationalizing mutual information computed in the estimated model.

Assume that the model is well-specified and λ is the true value of the parameters. Then since λ is the derivative of I_r in C,

$$I_r\left(\hat{C}_N\right) = I_r\left(C\right) + \lambda.\left(\hat{C}_N - C\right) + o\left(N^{-1/2}\right)$$

thus

$$I_r\left(\hat{C}_N\right) = I_r\left(C\right) + N^{-1/2}\sum_{x,y}\sum_{k}\lambda_k\phi^k\left(x,y\right)G\left(x,y\right) + o\left(N^{-1/2}\right)$$

while

$$\hat{I}_N = I + N^{-1/2}\sum_{x,y}\log\frac{\pi\left(x,y\right)}{p\left(x\right)q\left(y\right)}G\left(x,y\right) + o\left(N^{-1/2}\right)$$

Our test is thus based on the fact that, under the null of correct specification,

$$N^{1/2}\left(I_r\left(\hat{C}_N\right) - \hat{I}_N\right) \Longrightarrow \sum_{x,y}\left(\sum_{k}\lambda_k\phi^k\left(x,y\right) - \log\frac{\pi\left(x,y\right)}{p\left(x\right)q\left(y\right)}\right)G\left(x,y\right)$$

thus

$$N^{1/2}\left(I_r\left(\hat{C}_N\right) - \hat{I}_N\right) \sim N\left(0, var_\pi\left(\sum_{k}\lambda_k\phi^k\left(X,Y\right) - \log\frac{\pi\left(X,Y\right)}{p\left(X\right)q\left(Y\right)}\right)\right).$$

6.4 The Moment Matching Estimator

We shall call $\mathcal{W}_N\left(\theta\right)$ the value of the social surplus at parameter θ obtained with the empirical distributions of observable types $\hat{p}_N$ and $\hat{q}_N$.

While we could use maximum-likelihood to estimate λ, the results in section 5 suggest a much simpler method based solely on the observed covariations $\hat{C}$. Since λ is the derivative of I_r and I_r is the Legendre-Fenchel transform of $\mathcal{W}$, we take $\hat{\lambda}$ to minimize

$$\lambda\cdot\hat{C} - \mathcal{W}_N\left(\lambda,1\right) \tag{6.5}$$

over $\mathbb{R}^K$. This objective function is strictly convex, so that its minimizer is unique; and as shown in section 4, $\mathcal{W}_N$ can be computed very efficiently with our IPFP algorithm.

Letting $\hat{I}$ be the value of expression (6.5) at the optimal value $\hat{\lambda}$, we obtain the **Moment Matching (MM) estimator, denoted** $\hat{\Lambda}^{MM}$ and $\hat{\sigma}^{MM}$, by setting

$$\hat{\Lambda}^{MM} = \frac{\hat{\lambda}}{\hat{I}}, \quad \hat{\sigma}^{MM} = \frac{1}{\hat{I}}.$$

If the data was actually generated by model (SLOI) with parameters $\left(\hat{\Lambda}^{MM}, \hat{\sigma}^{MM}\right)$, the empirical covariations $\hat{C}_N$ would coincide with the optimal correlations $\bar{C}\left(\hat{\Lambda}^{MM}, \hat{\sigma}^{MM}\right)$. By construction, the Moment Matching estimator assigns the assorting weights values such that the predicted covariations coincide with the observed covariations. It is consistent and asymptotically Gaussian, and moreover:

Theorem 4 *Under (O), (S) and (GUI), if the model is correctly specified*

$$\sqrt{N}\left(\hat{\lambda}_N - \lambda\right) \Longrightarrow \mathcal{I}^{-1}\xi$$

where ξ is the Brownian bridge characterized in (6.4) and the matrix $\mathcal{I}^{kl}$ is the Fisher information matrix in (6.3). Therefore the MM estimator is asymptotically efficient.

7 An Illustration on US Census Data

To explore the fruitfulness of our proposed approach, we extracted data on married couples in the US from the American Community Survey (ACS) 2008 survey of the Census Bureau.The ACS collects data from a 1/100 representative sample of the US population every year. We downloaded the "small" sample from the Minnesota Population Center (Ruggles, Sobek, Alexander, Fitch, Goeken, Hall, King, and C.Ronnander (2008).) This contains about 75,000 individual records. To avoid having to deal with complex marital histories, we focus on couples which are the first marriage for both partners. At this stage of our data selection, we had 10,466 couples. To reduce heterogeneity, we chose to focus on "young" couples, where both partners are aged 20 to 35 and are out of school. This selection leaves us with 1,353 couples.

We chose to focus on the trade-off between matching on race and matching on education levels. We recoded the education variable in the CPS so that it takes five values: high school dropout (HSD), high school graduate (HSG), some college (SC; which includes two-year programs) college graduate (COLL), and postgraduate (POST). Similarly, we defined four values for race: white non hispanic, hispanic, black, and "others'.

Tables 1 and 2 describes the marginal distributions of characteristics $\hat{p}_N$ and $\hat{q}_N$ in the data. The numbers in these tables are all significantly larger than 0 at the 5% level, except for the two that are surrounded by parentheses. More than one third of hispanic men are high school dropouts, as against fewer than 7% for the three other race categories; at the other end of the educational ladder, close to one third of "other" men have a postgraduate degree (10% in the population of men.) The proportions are similar for women, although the discrepancies are less pronounced. We should also note here that our sample is not representative of the whole US population: by construction, it is younger and it consists of men and women in a first marriage.

Education: Race:	HS Dropout	HS Graduate	Some Coll.	Coll. Graduate	Postgraduate	Total
White non hispanic	0.041	0.177	0.216	0.231	0.075	0.740
Hispanic	0.050	0.046	0.028	0.016	(0.002)	0.142
Black	0.004	0.016	0.022	0.010	0.007	0.058
Others	0.004	0.009	0.009	0.019	0.019	0.060
Total	0.099	0.248	0.275	0.276	0.103	1.000

Table 1: Marginal characteristics $\hat{p}_N$ of husbands

The joint distribution of characteristics $\hat{\pi}_N$ is quite unwieldy: even considering just race and education, there are already $20^2 = 400$ cells to consider for possible matches. In fact, no fewer than 259 of these cells contain no match at all; and only 16 have more than 5, which makes nonparametric estimates very uninformative.

Education: Race:	HS Dropout	HS Graduate	Some Coll.	Coll. Graduate	Postgraduate	Total
White non hispanic	0.024	0.145	0.239	0.245	0.090	0.744
Hispanic	0.036	0.041	0.038	0.024	0.006	0.146
Black	0.003	0.011	0.021	0.011	0.006	0.052
Others	(0.002)	0.007	0.010	0.024	0.016	0.059
Total	0.066	0.205	0.307	0.305	0.118	1.000

Table 2: Marginal characteristics $\hat{q}_N$ of wives

On the other hand, looking at endogamy on each of our two dimensions separately is both easy and instructive. There are several measures of endogamy indices in the literature; we could for instance define endogamy indices by the diagonal of the matrix

$$\hat{E}_N(x,y) = \frac{\hat{\pi}_N(x,y)}{\hat{p}_N(x)\hat{q}_N(y)}.$$

This matrix is a natural choice for us since it figures prominently in our theoretical analysis, via the definition of mutual information

$$\hat{I}_N = \sum_{x,y} \hat{\pi}_N(x,y) \log \hat{E}_N(x,y).$$

Applying these two definitions to race gives us table 3, in which the number in each cell is $\hat{E}_N$. Random matching would make all $\hat{E}_N$ equal one in this table; in fact, they are all significantly different from one (the $0^\star$ reflects the absence of any couple with a Hispanic husband and a Black wife in our sample.)

The numbers on the diagonal (in bold) stand out. The $\hat{E}_N$ numbers on the diagonal are strikingly high, especially for Blacks and Others; but since these categories are less numerous than whites and hispanics, they end up contributing less to mutual information.

Table 4 gives endogamy indices on education. Only one of the $\hat{E}_N$ terms is non-significantly different from one, and there is again a $0^\star$. Here also the diagonal strongly dominates, but the effect is less striking.

Race of the wife: / Race of the husband:	White non hispanic	Hispanic	Black	Others
White non hispanic	**1.27**	0.21	0.17	0.24
Hispanic	0.21	**5.76**	$0^\star$	0.09
Black	0.24	0.35	**14.68**	0.21
Others	0.23	0.17	0.24	**13.36**

Table 3: Endogamy on races

The mutual information based on race is 0.484, while that on education is 0.264. Based on this measure, endogamy on race trumps endogamy on education.

Wife: / Husband:	HSD	HSG	SC	COLL	POST
HSD	**5.89**	1.75	0.63	0.15	0.13
HSG	(1.04)	**2.04**	1.18	0.44	0.13
SC	0.49	0.81	**1.57**	0.86	0.50
COLL	0.08	0.31	0.65	**1.77**	1.62
POST	$0^\star$	0.11	0.33	1.49	**3.61**

Table 4: Endogamy on educations

We could apply Theorem 3 and the Δ_2 operator to recover nonparametric estimates of complementarities; but given the paucity of data in most cells, we go directly to the semilinear specification. In order to be able to plot the covariogram in two dimensions, we only use the two most natural basis functions:

1. similarity on races: $\phi^1(R_m, E_m, R_w, E_w) = \mathbb{1}(R_m = R_w)$;
2. similarity on educations: $\phi^2(R_m, E_m, R_w, E_w) = \mathbb{1}(E_m = E_w)$.

The mean observed values of these two functions are simply the proportion of matches in

which both partners have the same race (resp. education). They form the vector of observed covariations

$$\hat{C}_N = (0.911, 0.466).$$

We first traced the frontier of the covariogram in (C_1, C_2) space, using the Munkres algorithm to generate the optimal matching for the surplus function

$$\lambda_1 \phi^1(R_m, E_m, R_w, E_w) + \lambda_2 \phi^2(R_m, E_m, R_w, E_w)$$

and drawing $1,000$ values of (λ_1, λ_2) randomly from the uniform distribution on the circle $\|\lambda\| = 1$, for $\sigma = 0$ since we are drawing the frontier.

While the implementation of the Munkres algorithm we used is quite efficient, it still takes 8 seconds on average to generate each optimal matching. The results are striking: somewhat to our surprise, the frontier of the covariogram is almost a perfect square. More detailed investigation shows that the four corners correspond to the four combination of the signs of λ_1 and λ_2: their values hardly seem to matter. The reason is that the conditional distributions of education within each race do not differ much for men and women, and that of course the marginal distribution of races are very close. Take $\lambda_1 > 0$ for instance, so that the optimal matching would maximize matching on race if λ_2 were zero. If λ_2 is not in fact zero, then it is possible to maintain the maximal matching on race while shuffling partners so that matching on educations is also maximized (if $\lambda_2 > 0$) or minimized (if $\lambda_2 < 0$.) In that sense, there is in fact little trade off between matching on race and matching on education in the homogeneous model with $\sigma = 0$.

This is not the case any more when $\sigma > 0$. We illustrate it on Figure 2, where we plotted the homogeneous frontier, the observed covariations $\hat{C}_N$ and the mutual information level set $I = I_r(\hat{C})$ that goes through it. This curve corresponds to a mutual information of 0.588, which is much smaller than $\hat{I}_N = 0.800$, the value of the nonparametric mutual information on race and education. The difference between these two values, multiplied by $N = 1,353$, generates a misspecification test, which clearly rejects the semilinear specification at any reasonable level. This is not surprising, as we are trying to explain 400 numbers with

just two parameters. The values of the respective mutual informations are summarized in Table 5.

To estimate $\hat{\lambda}_N$, we used our Moment Matching estimator, based on the IPFP algorithm. The objective function is strictly concave, and its gradient is extremely simple, so that it can be computed analytically to speed up the process. In fact this is hardly needed, as the estimator obtains in less than half a second; computing the optimal matching for any value of λ takes much less than a tenth of a second, which is about 1,000 times faster than in the homogeneous model.

The resulting estimators (normalizing $\sigma = 1$) are

$$\hat{\lambda}_N = (2.88, 1.03).$$

As expected, this is also the slope of the normal vector to the mutual information level set that goes through the observed covariations $\hat{C}_N$, as represented on Figure 2. The figure also shows the covariations implied by random matching, $C_\infty = (0.577, 0.238)$, which correspond to the mutual information level set $I = 0$, and an intermediate level set.

Specification	Value
Nonparametric, race	0.484
Nonparametric, education	0.264
Nonparametric, race and education	0.800
Semilinear, race and education	0.588

Table 5: Mutual informations

8 Possible Extensions and Concluding Remarks

Our theory so far relies on several strong assumptions.

The multinomial logit structure Our results rely heavily on assumption (GUI); yet it is well-known in applied econometrics that the multinomial logit model implies strong

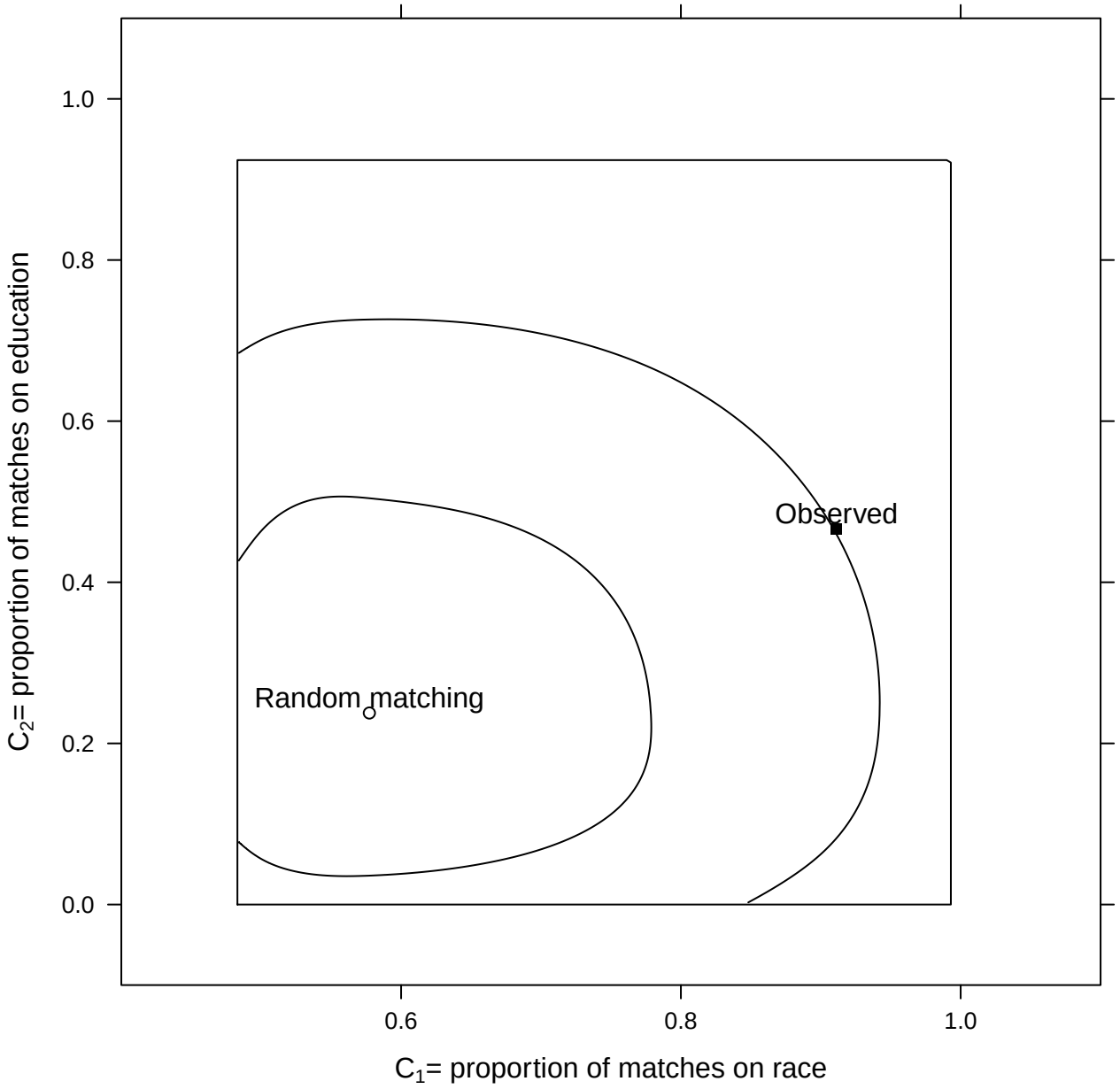

Figure 2: The Covariogram for Race and Education

restrictions. Note that the assumption can be tested along the lines of the procedures suggested by Hausman and McFadden (1984): a Hausman test based on the difference between our estimator (or the implied matching or covariations) and an estimator that only uses a restricted set of (x, y) types would have power against alternatives that violate IIA.

While we do not explore this here, it is also possible to obtain estimates of at least part of the surplus function Φ within a nested logit setting. More generally, it is easy to extend the results of the paper to the heteroskedastic case (in the form of functions $\sigma_1(x)$ and $\sigma_2(y)$). The expression in Theorem 1 turns into

$$\mathcal{W}(\theta) = \inf_{(U,V)\in A} \left(E_P \sigma_1(X) \log \sum_y \exp(U(X,y)/\sigma_1(X)) + E_Q \sigma_2(Y) \log \sum_x \exp(V(x,Y)/\sigma_2(Y)) \right), \tag{8.1}$$

while the expression in Theorem 2 becomes

$$\mathcal{W}(\theta) = \sup_{\Pi\in\mathcal{M}(P,Q)} \left(\sum_{x,y} \pi(x,y) \left\{ \Phi(x,y) - (\sigma_1(x) + \sigma_2(y)) \log \pi(x,y) \right\} \right). \tag{8.2}$$

Thus if we knew the functions σ_1 and σ_2, a nonparametric estimator of Φ in the heteroskedastic case would be

$$\hat{\Phi}_N(x,y) = (\sigma_1(x) + \sigma_2(y)) \log \hat{\pi}_N(x,y) + a(x) + b(y).$$

Typically we do not know σ_1 and σ_2, but we are willing to restrict their dependence on x and y; in the nested logit case for instance, σ_1 (res. σ_2) would only depend on the nest of x (resp. y), and it is enough to identify the variation of Φ within a nest.

Continuous distributions. While we have assumed discrete characteristics, we expect the main thrust of our arguments to carry over to the case where the distributions of the characteristics are continuous. We are working on such an extension; this will require adapting the (GUI) assumption to one that is better-suited to continuous choice.

Single households. So far we have not allowed for unmatched individuals. In an optimal matching, some men and/or women may remain single, as of course some must if

there are more individuals on one side of the market. The choice of the socially optimal matching can be broken down into the choice of the set of individuals who participate in matches and the choice of actual matches between the selected men and women. Our theory applies without any change to the second subproblem; that is, all of our results extend to M and W as selected in the first subproblem.

From the point of view of statistical inference, we may lose some efficiency in doing so; we note here that when the unobserved heterogeneity in preferences over partners is separable from the utility of marriage itself, our method does not incur any efficiency loss.

Non-bipartite matching. Bipartite matching refers to the fact that each individual is exogenously assigned in one category—in our terminology, husband or wife. Our analysis in fact is very easy to extend so as to incorporate same-sex unions, and thus to rationalize endogamy in the gender dimension.

To do so, we just need to add one (observed) characteristic, in the form of gender. If for instance gender becomes the first dimension of the characteristics vector, then the observed surplus has an assorting weight $\Lambda_{11} < 0$ that reflects the more typical preference for the opposite sex; while heterogenous preferences χ and η will automatically take into account the dispersion of individual preference for same-sex unions.

References

Becker, G. (1973): "A theory of marriage, part I," *Journal of Political Economy*, 81, 813–846.

Blair, C. (1984): "Every finite distributive lattice is a set of stable matchings," *Journal of Combinatorial Theory, Series A*, 37, 353—356.

Chiappori, P., S. Oreffice, and C. Quintana-Domeque (2009): "Fatter Attraction: Anthropometric and Socioeconomic Characteristics in the Marriage Market," DP 4594, IZA.

Chiappori, P.-A., R. McCann, and L. Nesheim (2008): "Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness," *Economic Theory*.

Chiappori, P.-A., B. Salanié, A. Tillman, and Y. Weiss (2008): "Assortative Matching on the Marriage Market: A Structural Investigation," mimeo Columbia University.

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

Decker, C., B. Stephens, and R. McCann (2009): "When do systematic gains uniquely determine the number of marriages between different types in the Choo-Siow matching model? Sufficient conditions for a unique equilibrium," mimeo University of Toronto.

Dudley, R. M. (2002): *Real Analysis and Probability*. Cambridge University Press.

Echenique, F. (2008): "What matchings can be stable? The testable implications of matching theory," *Mathematics of Operations Research*, 33, 757–768.

Echenique, F., K. Lee, and M. Shum (2009): "Aggregate Matchings," mimeo.

Fox, J. (2009): "Identification in Matching Games," Discussion paper, NBER.

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

Hatfield, J., and P. Milgrom (2005): "Matching with Contracts," *American Economic Review*, 95, 913–955.

Hausman, J., and D. McFadden (1984): "Specification Tests for the Multinomial Logit Model," *Econometrica*, 52, 1219–1240.

Hiriart-Urrut, J.-B., and C. Lemaréchal (2001): *Fundamental of Convex Analysis*. Springer.

Hitsch, G., A. Hortacsu, and D. Ariely (2010): "Matching and Sorting in Online Dating," *American Economic Review forthcoming*.

Lee, S. (2009): "Marriage and Online Mate-Search Services: Evidence From South Korea," mimeo, University of Maryland.

Mézard, M., and A. Montanari (2009): *Information, Physics, and Computation*. Oxford University Press.

Parisi, G. (1988): *Statistical Field Theory*. Perseus Books.

Ruggles, S., M. Sobek, T. Alexander, C. Fitch, R. Goeken, P. Hall, M. King, and C.Ronnander (2008): "Integrated Public Use Microdata Series: Version 4.0," Discussion paper, Minnesota Population Center.

Rüschendorf, L. (1995): "Convergence of the iterative proportional fitting procedure," *Annals of Statistics*, 23, 1160–1174.

Rüschendorf, L., and W. Thomsen (1998): "Closedness of Sum Spaces and the Generalized Schrödinger Problem," *Theory of Probability and its Applications*, 42, 483–494.

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

Siow, A. (2009): "Testing Becker's Theory of Positive Assortative Matching," Discussion paper, University of Toronto.

Villani, C. (2003): *Topics in Optimal Transportation*. American Mathematical Society.

——— (2009): *Optimal Transport, Old and New*. Springer.

A Facts from Convex Analysis

A.1 Basic results

We only sum up here the concepts we actually use in the paper; we refer the reader to Hiriart-Urrut and Lemaréchal (2001) for a thorough exposition of the topic.

Take any set $Y \subset \mathbb{R}^d$; then the *convex hull* of Y is the set of points in $\mathbb{R}^d$ that are convex combinations of points in Y. We usually focus on its closure, the closed convex hull, denoted $cch\,(Y)$.

The *support function* S_Y of Y is defined as

$$S_Y(x) = \sup_{y \in Y} x \cdot y$$

for any x in Y. It is a convex function, and it is homogeneous of degree one. Moreover, $S_Y = S_{\mathrm{cch}(Y)}$ where $\mathrm{cch}\,(Y)$ is the closed convex hull of Y, and $\partial S_Y(0) = \mathrm{cch}\,(Y)$.

A point in Y is an *extreme point* if it does not belong to any open line segment joining two points of Y.

Now let u be a convex, continuous function defined on $\mathbb{R}^d$. Then the gradient ∇u of u is well-defined almost everywhere and locally bounded. If u is differentiable at x, then

$$u\left(x'\right) \geq u(x) + \nabla u(x) \cdot (x' - x)$$

for all $x' \in \mathbb{R}^d$. Moreover, if u is also differentiable at x', then

$$\left(\nabla u(x) - \nabla u\left(x'\right)\right) \cdot \left(x - x'\right) \geq 0.$$

When u is not differentiable in x, it is still *subdifferentiable* in the following sense. We define $\partial u(x)$ as

$$\partial u(x) = \left\{ y \in \mathbb{R}^d : \forall x' \in \mathbb{R}^d, u\left(x'\right) \geq u(x) + y \cdot (x' - x) \right\}.$$

Then $\partial u(x)$ is not empty, and it reduces to a single element if and only if u is differentiable at x; in that case $\partial u(x) = \{\nabla u(x)\}$.

A.2 Generalized Convexity

In order to make the paper self-contained, we present basic results on the theory of *generalized convexity*, sometimes called the theory of c-convex functions. This theory extends many results from convex analysis and, in particular, duality results, to a much more general setting. We refer to Villani (2009), p. 54–57 (or Villani (2003), pp. 86–87) for a detailed account[7].

Let ω be a function from the product of two sets $\mathcal{X} \times \mathcal{Y}$ to $[-\infty, +\infty)$.

Definition 1 *Consider any function $\psi : \mathcal{X} \to (-\infty, +\infty]$. Its generalized Legendre transform $\psi^{\perp} : \mathcal{X} \to [-\infty, +\infty)$ is defined by*

$$\psi^{\perp}(y) = \inf_{x \in \mathcal{X}} \{\psi(x) - \omega(x, y)\}.$$

Conversely, take any function $\zeta : \mathcal{Y} \to [-\infty, +\infty)$; then its generalized Legendre transform $\zeta^{\top} : \mathcal{X} \to (-\infty, +\infty]$ is defined by

$$\zeta^{\top}(x) = \sup_{y \in \mathcal{Y}} \{\zeta(y) + \omega(x, y)\}.$$

A function ψ is called ω-convex if it is not identically $+\infty$ and if there exists $\zeta : \mathcal{Y} \to [-\infty, +\infty]$ such that

$$\psi = \zeta^{\top}.$$

Recall that the usual Legendre transform is defined as

$$\psi^{*}(y) = \inf_{x \in \mathcal{X}} \{\psi(x) - x \cdot y\};$$

thus it coincides with the generalized Legendre transform when ω is bilinear, and then ω-convexity boils down to standard convexity.

Our analysis rests on the following fundamental result, which generalizes standard convex analysis.

[7] A cautionary remark is in order here: the sign conventions vary in the literature, so our own choices may differ from those of any given author.

Proposition 7 *For every function $\psi : \mathcal{X} \to (-\infty, +\infty]$,*

$$\psi^{\perp\top} \leq \psi$$

with equality if and only if ψ is ω-convex.

Proof Take any $x \in \mathcal{X}$; then

$$\psi^{\perp\top}(x) = \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \left\{ \psi\left(x'\right) - \omega\left(x', y\right) + \omega\left(x, y\right) \right\};$$

taking $x' = x$ shows that $\psi^{\perp\top}(x) \leq \psi(x)$.

Conversely, if $\psi^{\perp\top} = \psi$ then $\psi(x) = \zeta^{\top}(x)$, with $\zeta = \psi^{\perp}$. But for any function ζ, the triple transform $\zeta^{\top\perp\top}$ coincides with $\zeta^{\top}$. To see this, write

$$\zeta^{\top\perp\top}(x) = \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \sup_{y' \in \mathcal{Y}} \left\{ \zeta\left(y'\right) + \omega\left(x', y'\right) - \omega\left(x', y\right) + \omega\left(x, y\right) \right\}.$$

Now for all x and y,

$$\inf_{x' \in \mathcal{X}} \sup_{y' \in \mathcal{Y}} \left\{ \zeta\left(y'\right) + \omega\left(x', y'\right) - \omega\left(x', y\right) \right\} \geq \zeta(y)$$

as is easily seen by taking $y' = y$; therefore

$$\zeta^{\top\perp\top}(x) \geq \sup_{y \in \mathcal{Y}} \left\{ \zeta(y) + \omega(x, y) \right\} = \zeta^{\top}(x).$$

Applying this to the ζ such that $\psi = \zeta^{\top}$ concludes the proof.

QED.

B Proofs

B.1 Proof of Theorem 1

In order to prove Theorem 1, some preparation is needed. For any function $\tilde{u}(\tilde{x})$, fix x and use the theory of generalized convexity briefly recalled in Appendix (A.2) to define

$$\tilde{u}^{\perp}(x, y) = \inf_{\tilde{x} | x} \left\{ \tilde{u}(\tilde{x}) - \chi(\tilde{x}, y) \right\}$$

the *generalized Legendre transform* of $\tilde{u}$ with respect to χ for fixed x—the notation $\tilde{x}|x$ stands for "the set of values of the full type $\tilde{x}$ for which the observable type takes the value x". We define in the same manner

$$\tilde{v}^{\perp}(x,y) = \inf_{\tilde{y}|y} \{\tilde{v}(\tilde{y}) - \xi(\tilde{y},x)\}.$$

Similarly, for any two functions $U(x,y)$ and $V(x,y)$, we define

$$\begin{aligned} U^{\top}(\tilde{x}) &:= \sup_{y} \{U(x,y) + \chi(\tilde{x},y)\} \\ V^{\top}(\tilde{y}) &:= \sup_{x} \{V(x,y) + \xi(\tilde{y},x)\}. \end{aligned}$$

Lemma 1 *Let A be the set of pairs of functions (U,V) such that*

$$\forall x, y,\ U(x,y) + V(x,y) \geq \Phi(x,y).$$

Then

$$\mathcal{W} = \inf_{(U,V)\in A} \left\{ \int U^{\top}(\tilde{x})\, d\tilde{P}(\tilde{x}) + \int V^{\top}(\tilde{y})\, d\tilde{Q}(\tilde{y}) \right\}.$$

Proof of Lemma 1 By the Kantorovich duality theorem (Villani (2009) Theorem 5.10),

$$\mathcal{W} = \sup_{\tilde{\pi}\in\mathcal{M}(P,Q)} \int \tilde{\Phi}(\tilde{x},\tilde{y})\, d\tilde{\pi}(\tilde{x},\tilde{y}) = \inf_{(\tilde{u},\tilde{v})\in\tilde{A}} \left\{ \int \tilde{u}(\tilde{x})\, d\tilde{P}(\tilde{x}) + \int \tilde{v}(\tilde{y})\, d\tilde{Q}(\tilde{y}) \right\}, \quad \text{(B.1)}$$

where $\tilde{A}$ is the set of pairs of functions $(\tilde{u},\tilde{v})$ such that

$$\forall \tilde{x}, \tilde{y},\ \tilde{u}(\tilde{x}) + \tilde{v}(\tilde{y}) \geq \tilde{\Phi}(\tilde{x},\tilde{y}).$$

Note the following two facts about the right-hand side of this equality:

1. Since

$$\tilde{\Phi}(\tilde{x},\tilde{y}) = \Phi(x,y) + \chi(\tilde{x},y) + \xi(\tilde{y},x),$$

the infimum in (B.1) can be taken over the pair of functions $(\tilde{u},\tilde{v})$ that satisfy

$$\tilde{u}(\tilde{x}) \geq \sup_{y} \left\{ \Phi(x,y) + \chi(\tilde{x},y) + \sup_{\tilde{y}|y} [\xi(\tilde{y},x) - \tilde{v}(\tilde{y})] \right\},$$

or

$$\tilde{u}(\tilde{x}) \geq \sup_{y} \left\{ \Phi(x,y) + \chi(\tilde{x},y) - \tilde{v}^{\perp}(x,y). \right\}$$

At the optimum this must hold with equality. Going back to Definition 1, it follows that $\tilde{u}$ is χ-convex for any fixed x; and using Proposition 7, we can substitute $\tilde{u}$ with $\tilde{u}^{\perp\top}$, that is:

$$\tilde{u}(\tilde{x}) = \sup_{y} \left\{ \tilde{u}^{\perp}(x,y) + \chi(\tilde{x},y) \right\}.$$

Applying a similar argument to $\tilde{v}$, the objective function can be rewritten as

$$\int \sup_{y} \left\{ \tilde{u}^{\perp}(x,y) + \chi(\tilde{x},y) \right\} d\tilde{P}(\tilde{x}) + \int \sup_{x} \left\{ \tilde{v}^{\perp}(x,y) + \xi(\tilde{y},x) \right\} d\tilde{Q}(\tilde{y}).$$

2. Also note that the constraint of the minimization problem in (B.1) is also

$$\forall x, y, \;\; \tilde{u}^{\top}(x,y) + \tilde{v}^{\perp}(x,y) \geq \Phi(x,y)$$

which follows directly from the fact that

$$\forall \tilde{x}|x, \; \tilde{y}|y, \;\; \tilde{u}(\tilde{x}) - \chi(\tilde{x},y) + \tilde{v}(\tilde{y}) - \xi(\tilde{y},x) \geq \Phi(x,y).$$

Now define

$$U(x,y) = \tilde{u}^{\perp}(x,y) \text{ and } V(x,y) = \tilde{v}^{\perp}(x,y);$$

Given points 1. and 2. above, we can rewrite the value $\mathcal{W}$ as

$$\mathcal{W} = \inf_{(U,V)\in A} \left\{ \int U^{\top}(\tilde{x}) \, d\tilde{P}(\tilde{x}) + \int V^{\top}(\tilde{y}) \, d\tilde{Q}(\tilde{y}) \right\}.$$

QED.

We are now in a position to prove the theorem.

Proof of Theorem 1 Start by drawing two samples of size N of men and women from their population distributions P and Q; we denote the corresponding values of the observed

characteristics $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$. Call P_n and Q_n the corresponding sample distributions; e.g. P_n assigns a mass

$$p_{k,n} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(x_j = x^k)$$

to the value x^k of observable characteristics of men. The Law of Large Numbers implies that P_n and Q_n converge in distribution to P and Q, the population distributions of the observable types. Now we have for any possible x

$$\int U^\top(\tilde{x})\, d\tilde{P}_n(\tilde{x}|X = x) = \sum_{\substack{i=1,..,n \\ x_i = x}} \sup_{j=1,...,n} \{U(x_i, y_j) + \chi(\tilde{x}_i, y_j)\} + o(1)$$

As N gets large, each of the possible values of observable characteristics of women y^k is included in the sample $\{y_1, ..., y_n\}$; therefore the supremum in the above expression runs over all such possible values $\left\{y^1, \ldots, y^{n_Q}\right\}$. But under (GUI), conditional on X the random variables $\chi\left(\tilde{x}, y^k\right)$ are independent Gumbel random variables with scaling factor σ_1, so we get

$$\frac{1}{\sigma_1} \int U^\top(\tilde{x})\, d\tilde{P}_n(\tilde{x}|X = x) = \log \sum_{k=1}^{n_Q} \exp\left(U\left(x, y^k\right)/\sigma_1\right) + o_P(1)$$

hence, taking the limit and integrating over x,

$$\int U^\top(\tilde{x})\, d\tilde{P}(\tilde{x}) = \sigma_1 E_P \log \sum_y \exp\left(U(X, y)/\sigma_1\right)$$

and similarly

$$\int V^\top(\tilde{y})\, d\tilde{Q}(\tilde{y}) = \sigma_2 E_Q \log \sum_x \exp\left(V(x, Y)/\sigma_2\right).$$

QED.

B.2 Proof of Theorem 2

Proof By theorem 1, we have

$$\mathcal{W}_N = \inf_{U(x,y)+V(x,y) \geq \Phi(x,y)\ \forall x,y} \left\{ \begin{array}{l} \sigma_1 \sum_x p(x) \log\left(\sum_y \exp\left(U(x, y)/\sigma_1\right)\right) \\ +\sigma_2 \sum_y q(y) \log\left(\sum_x \exp\left(V(x, y)/\sigma_2\right)\right) \end{array} \right\}$$

for which we form the Lagrangian

$$\begin{aligned}\mathcal{W}_N &= \inf_{U(x,y),V(x,y)} \sup_{\pi(x,y)\geq 0} \left\{ \begin{array}{c} \sigma_1 \sum_x p(x) \log\left(\sum_y \exp\left(U(x,y)/\sigma_1\right)\right) \\ +\sigma_2 \sum_y q(y) \log\left(\sum_x \exp\left(V(x,y)/\sigma_2\right)\right) \\ +\sum_{x,y} \pi(x,y)\left(\Phi(x,y) - U(x,y) - V(x,y)\right) \end{array} \right\} \\ &= \sup_{\pi(x,y)\geq 0} \left\{ \sum_{xy} \pi(x,y)\Phi(x,y) + \inf_{U(.,.)} F(U) + \inf_{V(.,.)} G(V) \right\}\end{aligned}$$

where

$$\begin{aligned}F(U) &= \sigma_1 \sum_x p(x) \log\left(\sum_y \exp\left(U(x,y)/\sigma_1\right)\right) - \sum_{xy} \pi(x,y) U(x,y) \\ G(V) &= \sigma_2 \sum_y q(y) \log\left(\sum_x \exp\left(V(x,y)/\sigma_2\right)\right) - \sum_{xy} \pi(x,y) V(x,y).\end{aligned}$$

Clearly, $U(.,.)$ and $V(.,.)$ in the inner minimization problems satisfy

$$\pi(x,y) = \frac{p(x)\exp\left(U(x,y)/\sigma_1\right)}{\sum_y \exp\left(U(x,y)/\sigma_1\right)} = \frac{q(y)\exp\left(V(x,y)/\sigma_2\right)}{\sum_x \exp\left(V(x,y)/\sigma_2\right)}; \tag{B.2}$$

note that these equations imply that $\sum_y \pi(x,y) = p(x)$ and $\sum_x \pi(x,y) = q(y)$, so that $\pi \in \mathcal{M}(P,Q)$. Rearranging terms,

$$\mathcal{W}_N = \sup_{\pi\in\mathcal{M}(P,Q)} \left\{ \begin{array}{c} \sum_{xy} \pi(x,y)\Phi(x,y) - (\sigma_1+\sigma_2)\sum_{xy} \pi(x,y)\log\pi(x,y) \\ +\sigma_1 \sum_x p(x)\log p(x) + \sigma_2 \sum_y q(y)\log q(y) \end{array} \right\}$$

and noticing that $\sum_{xy} \pi(x,y)\log\pi(x,y) = -S(\pi) = I(\pi) - S(P) - S(Q)$ gives the desired result.

B.3 Proof of Corollary 1

Proof The result follows directly from the Kantorovich duality theorem (cf. Villani (2009), Ch. 2); it can also be obtained by letting σ_1 and σ_2 tend to zero in Theorem 1 and by noting that as $\sigma_1, \sigma_2 \to 0$,

$$\begin{aligned}\sigma_1 E_P\left[\log \sum_y \left[\exp\left(U(X,y)/\sigma_1\right)\right]\right] &\to E_P\left[\max_y U(X,y)\right], \\ \sigma_2 E_Q\left[\log \sum_x \left[\exp\left(V(x,Y)/\sigma_2\right)\right]\right] &\to E_Q\left[\max_x U(x,Y)\right].\end{aligned}$$

B.4 Proof of Proposition 1

Proof Non-emptiness is obvious. To see that $\mathcal{F}_c$ is convex, let $\hat{C}$ and $\tilde{C}$ be two feasible cross-product matrices in $\mathcal{F}_c$. Pick any $\alpha \in [0,1]$ and consider $\alpha\hat{C}+(1-\alpha)\tilde{C}$. By definition of $\mathcal{F}_c$, there exist $\hat{\pi}$ and $\tilde{\pi}$ in $\mathcal{M}(P,Q)$ such that $\hat{C}_{ij} = E_{\hat{\pi}}[X_{ij}Y_{ij}]$ and $\tilde{C}_{ij} = E_{\tilde{\pi}}[X_{ij}Y_{ij}]$. Let $\bar{\pi} = \alpha\hat{\pi} + (1-\alpha)\tilde{\pi}$. Then $\alpha\hat{C}_{ij} + (1-\alpha)\tilde{C}_{ij} = E_{\bar{\pi}}[X_{ij}Y_{ij}]$, and $\bar{\pi} \in \mathcal{M}(P,Q)$, thus $\alpha\hat{C} + (1-\alpha)\tilde{C} \in \mathcal{F}_c$.

Now we prove that $\mathcal{F}_c$ is closed: Let C_n be a sequence in $\mathcal{F}_c$ converging to $C \in \mathbb{R}^{rs}$, and let π_n be the associated matching. By Theorem 11.5.4 in Dudley (2002), as $\mathcal{M}(P,Q)$ is uniformly tight, π_n has a weakly converging subsequence in $\mathcal{M}(P,Q)$; call π its limit. Then C is the cross-product associated to π, so that $C \in \mathcal{F}_c$.

Finally, $\mathcal{F}$ is a closed convex set as it is the upper graph of the function $I_r(C)$ defined in Eq. (5.2).

B.5 Proof of Proposition 2

Proof $\mathcal{R}$ is the union of the subgradients of $\mathcal{W}$ which was seen in Prop. 1 to be the support function of $\mathcal{F}$: hence $\mathcal{R}$ is the frontier of $\mathcal{F}$.

B.6 Proof of Proposition 3

Proof a) Positive homogeneity and convexity of degree one follow from the fact that $\mathcal{W}$ is the support function of $\mathcal{F}$. Strict convexity for $\sigma > 0$ follows from the strict convexity of $I(\pi)$. Part b) follows directly from the envelope theorem. Part c) results of $I_r(C)$ being the Legendre transform of $\mathcal{W}(\lambda,1)$; since the latter is strictly convex, I_r is convex on $\mathcal{F}_c$, it is C^1 on its interior, and by the envelope theorem, $\frac{\partial I_r}{\partial C^k} = \lambda_k$.

B.7 Proof of Proposition 4

Proof a) The sets $\mathcal{R}_c(I)$ are extreme points of the sets $I_r^{-1}([0,I])$ which are closed convex sets. One has $I_r^{-1}(\{0\}) = \{C_\infty\}$ which corresponds to $\Pi = P \otimes Q$, and $I_r^{-1}([0, S(P) + S(Q)]) = \mathcal{F}_c$.

b) Clearly, $\hat{C} \in \mathcal{R}_c\left(I_r\left(\hat{C}\right)\right)$.

c) The differential form $\sum_k \Lambda_k dC^k - \sigma dI$ vanishes on all of $\mathcal{F}$; but since $dI = 0$ on $\mathcal{R}_c(I)$, the form $\sum_k \Lambda_k dC^k$ vanishes there too.

d) was proved in the text and in c) above.

e) follows from the definition of $\bar{C}$.

B.8 Proof of Theorem 3

Proof 1. For $\sigma > 0$, the map $\pi \to \sum_{x,y} \pi(x,y)\Phi(x,y) - \sigma I(\pi)$ is strictly concave and finite-valued on the convex domain $\mathcal{M}(P,Q)$; thus there exists a unique $\pi \in \mathcal{M}(P,Q)$ maximizing (2.2).

2. Let B be the set of pairs of functions $(u(x), v(y))$ such that $\sum_x u(x) p(x) = \sum_y v(y) q(y) = 0$, and for $(u,v) \in B$, let Z be the function defined by

$$Z(u,v) := \sum_{x,y} p(x) q(y) \exp\left(\frac{\Phi(x,y) - u(x) - v(y)}{\sigma}\right).$$

Introduce

$$p_{u,v}(x) \quad := \quad = \frac{\partial \log Z(u,v)}{\partial u(x)} = \frac{\sum_y p(x) q(y) \exp\left(\frac{\Phi(x,y)-u(x)-v(y)}{\sigma}\right)}{\sum_{x,y} p(x) q(y) \exp\left(\frac{\Phi(x,y)-u(x)-v(y)}{\sigma}\right)}$$

$$q_{u,v}(y) \quad := \quad = \frac{\partial \log Z(u,v)}{\partial v(y)} = \frac{\sum_x p(x) q(y) \exp\left(\frac{\Phi(x,y)-u(x)-v(y)}{\sigma}\right)}{\sum_{x,y} p(x) q(y) \exp\left(\frac{\Phi(x,y)-u(x)-v(y)}{\sigma}\right)}$$

as a result $p_{u,v}$ and $q_{u,v}$ are probability vectors. It is easy to see that Z is strictly log-convex;

thus there exists a unique pair of functions $(u, v) \in B$ such that

$$\begin{aligned} p &= p_{u,v} \\ q &= q_{u,v}. \end{aligned}$$

Take this pair of functions and define $c = \sigma \log Z(u, v)$; then the function

$$\pi(x, y) = p(x)\, q(y) \exp\left(\frac{\Phi(x, y) - u(x) - v(y) - c}{\sigma}\right)$$

belongs to $\mathcal{M}(P, Q)$.

3. Let $\pi \in \mathcal{M}(P, Q)$ be the solution of (2.2). From Expression (B.2) in the proof of Theorem 1,

$$\begin{aligned} \sigma_1 \log \pi(x, y) &= U(x, y) + \sigma_1 \log p(x) - \sigma_1 \log\left(\sum_y \exp\left(U(x, y)/\sigma_1\right)\right) \\ \sigma_2 \log \pi(x, y) &= V(x, y) + \sigma_2 \log q(y) - \sigma_2 \log\left(\sum_x \exp\left(V(x, y)/\sigma_2\right)\right). \end{aligned}$$

Thus, summing up:

$$\sigma \log \frac{\pi(x, y)}{p(x)\, q(y)} = \Phi(x, y) - \bar{u}(x) - \bar{v}(y)$$

where

$$\begin{aligned} \bar{u}(x) &= \sigma_2 \log p(x) + \sigma_1 \log\left(\sum_y \exp\left(U(x, y)/\sigma_1\right)\right) \\ \bar{v}(y) &= \sigma_1 \log q(y) + \sigma_2 \log\left(\sum_x \exp\left(V(x, y)/\sigma_2\right)\right). \end{aligned}$$

Now take $c_1 = \sum_x p(x)\bar{u}(x)$ and $c_2 = \sum_y q(y)\bar{v}(y)$; and define

$$u(x) \equiv \bar{u}(x) - c_1, \quad v(y) \equiv \bar{v}(y) - c_2.$$

By construction, $(u, v) \in B$. Hence π is solution of equation (3.1). It follows immediately that $c = \mathcal{W}$.

4. The consequence is obvious; now take another Φ' that satisfies the cross-difference equation for some $\sigma' > 0$, and functions u', v' and a constant c' such that

$$u'(x) + v'(y) + c' \equiv \frac{\sigma'}{\sigma}(\Phi(x,y) - u(x) - v(y) - c) - \Phi'(x,y).$$

Since $\sigma'\Delta_2\Phi' \equiv \sigma\Delta_2\Phi$, the right-hand side has zero cross-differences, and so this functional equation has an infinity of solutions. We just need to normalize them by $E_P u'(X) \equiv E_Q v'(Y) \equiv 0$.

5. The fact that if $p(x)$ and $q(y)$ are positive then so is $\pi(x,y)$ follows directly from formula 3.1.

B.9 Proof of Proposition 5

Proof $\mathcal{M}(p,q) = \left\{\pi(x,y) : \sum_y \pi(x,y) = p(x),\ \sum_x \pi(x,y) = q(y),\ \pi(x,y) \geq 0\right\}$. Now, $\lambda = \Lambda/\sigma \in \mathcal{D}$ implies that the solution π_θ belongs to the strict interior of $\mathcal{M}(p,q)$, so that none of the non-negativity constraints are binding. For a given θ, the vector π_θ is defined as the maximizer of a C^∞ function of θ and π with respect to π on $\mathcal{M}(p,q)$, thus by the implicit function theorem, $\theta \to \pi_\theta(x,y)$ is C^∞ on $\mathcal{D}$. By equation (3.1), we have

$$\log \frac{\pi_\theta(x,y)}{p(x)\,q(y)} = \frac{\sum_k \Lambda^k \phi^k(x,y) - u_\theta(x) - v_\theta(y) - c_\theta}{\sigma}$$

hence $\sigma \frac{\partial \log \pi_\theta}{\partial \Lambda_k}(x,y) = \phi^k(x,y) - \frac{\partial u_\theta(x)}{\partial \Lambda_k}(x) - \frac{\partial v_\theta(y)}{\partial \Lambda_k} - \frac{\partial c_\theta}{\partial \Lambda_k}$. But

$$\sum_x \frac{\partial \log \pi_\theta}{\partial \Lambda_k}(x,y)\,\pi_\theta(x,y) = \sum_x \frac{\partial \pi_\theta}{\partial \Lambda_k}(x,y) = \frac{\partial}{\partial \Lambda_k}\sum_x \pi_\theta(x,y) = \frac{\partial q(y)}{\partial \Lambda_k} = 0,$$

thus for all x and y,

$$E\left(\frac{\partial \log \pi_\theta(X,Y)}{\partial \Lambda_k} | X = x\right) = 0 \text{ and } E\left(\frac{\partial \log \pi_\theta(X,Y)}{\partial \Lambda_k} | Y = y\right) = 0;$$

hence the two-way ANOVA decomposition of ϕ^k for π_θ is

$$\phi^k(x,y) = \sigma \frac{\partial \log \pi_\theta}{\partial \Lambda_k}(x,y) + \frac{\partial u_\theta(x)}{\partial \Lambda_k} + \frac{\partial v_\theta(y)}{\partial \Lambda_k} + E_{\pi_\theta}\phi^k(X,Y),$$

and the ANOVA residue of ϕ^k for π_θ is

$$\varepsilon(x,y;\phi^k,\pi_\theta) = \sigma \frac{\partial \log \pi_\theta}{\partial \Lambda_k}(x,y).$$

B.10 Proof of Proposition 6

Proof We have

$$\frac{\partial \mathcal{W}}{\partial \Lambda_l}(\Lambda, \sigma) = E_{\pi_\theta} \phi^l(X, Y);$$

hence

$$\frac{\partial^2 \mathcal{W}}{\partial \Lambda_k \partial \Lambda_l}(\Lambda, \sigma) = E_{\pi_\theta}\left[\phi^l(X, Y) \frac{\partial \log \pi_\theta}{\partial \Lambda_k}(X, Y)\right] = \sigma E_{\pi_\theta}\left(\frac{\partial \log \pi_\theta}{\partial \Lambda_k}(X, Y)\frac{\partial \log \pi_\theta}{\partial \Lambda_l}(X, Y)\right) = \sigma \mathcal{I}^{kl}.$$

Moreover, since

$$\varepsilon(x, y; \phi^k, \pi_\theta) = \sigma \frac{\partial \log \pi_\theta}{\partial \Lambda_k}(x, y),$$

we get

$$\sigma^2 \mathcal{I}^{kl} = E\left(\varepsilon(x, y; \phi^k, \pi_\theta)\varepsilon(x, y; \phi^l, \pi_\theta)\right).$$

B.11 Proof of Theorem 4

Proof We have $\hat{\lambda}_N = \frac{\partial I_r}{\partial C}$, hence at first order $\hat{\lambda}_N - \lambda = D^2 I_r.\left(\hat{C}_N - C\right) + o_P\left(1/\sqrt{N}\right)$. But I_r is the Legendre transform of $\mathcal{W}(\cdot, 1)$, and therefore $D^2 I_r = \left(D^2 \mathcal{W}(\cdot, 1)\right)^{-1} = \mathcal{I}^{-1}$ by Proposition 6.

C Connections to Statistical physics

There is a very close parallel between our theory and statistical physics and thermodynamics. We refer to Parisi (1988) for more on statistical physics, and to Mézard and Montanari (2009) for the connection with information theory. Let us just mention here that the social welfare $\mathcal{W}$ is the analog of a *total energy*; the term $\sum \lambda_k C^k$ is the analog of an *internal energy*; $I(\pi)$ is the analog of an *entropy*; the parameter σ is the analog of a *temperature*. A pure matching is the equivalent of a *solid state*; the points of nondifferentiability of $\mathcal{W}$ are analog to *critical points*.

Equation 3.1 is known in the mathematical physics literature as the Schrödinger-Bernstein equation, cf. Rüschendorf and Thomsen (1998) and references therein. It was first studied

by Erwin Schrödinger as part of his research program in time irreversibility in statistical physics. Interestingly, it also bears some connections with his better-known "Schrödinger equation" in quantum mechanics. In fact, as discovered by Zambrini, a dynamic formulation of this equation is the Euclidian Schrödinger equation which arises in Ed Nelson's formulation of "stochastic mechanics," a Euclidian analog of quantum mechanics. For more on this topic, see Parisi (1988), Chap. 19.