CEPS INSTEAD

WORKING PAPERS

Canonical correlation and assortative matching: A remark

Arnaud DUPUY[1]
Alfred Galichon[2]

CEPS/INSTEAD, Luxembourg[1]
Sciences Po, France[2]

Working Paper No 2014-10
July 2014

CANONICAL CORRELATION AND ASSORTATIVE MATCHING: A REMARK

ARNAUD DUPUY[§] AND ALFRED GALICHON[†]

Abstract. In the context of the Beckerian theory of marriage, where men and women match on a single-dimensional index that is the weighted sum of their respective multivariate attributes, many papers in the literature have used linear canonical correlation, and related techniques, in order to estimate these weights. We argue that this estimation technique is inconsistent and suggest some solutions.

Keywords: matching, marriage, assignment, assortative matching, canonical correlation.

JEL codes: C78, D61, C13.

[§] *CEPS/INSTEAD, Maastricht School of Management and IZA. Address: CEPS/INSTEAD, 3, avenue de la Fonte - L-4364 Esch-sur-Alzette, Luxembourg. Email: arnaud.dupuy@ceps.lu.* Tel: +352585855551, Fax: +352585855700.

[†] *Sciences Po Paris, Department of Economics, CEPR and IZA. Address: 27 rue Saint-Guillaume, 75007 Paris, France. E-mail: alfred.galichon@sciencespo.fr.* Tel: +33(0)145498582, Fax: +33(0)145497257.

Date: May 23, 2014. The authors thank two anonymous referees, the Editor (Xavier D'Haultfoeuille), as well as Xavier Gabaix, Bernard Salanié and Marko Terviö for helpful comments. Galichon's research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 313699, and from FiME, Laboratoire de Finance des Marchés de l'Energie. Dupuy warmly thanks the Maastricht School of Management where part of this research was performed.

Introduction. Who marries whom and why are questions that have received tremendous attention by scientists from many different fields: economics, sociology, psychology and biology. This literature shows that a correlation between spouses' attributes exists for many attributes, i.e. height, weight, education, earnings, wealth, religion, ethnicity, personality traits to mention just a few. How many and which of these attributes actually matter for the sorting of men and women? Up until recently, by lack of a better methodology, the literature dealt with the first question by simply assuming that men and women match on a single-dimensional index that is the weighted sum of their respective multivariate attributes. The second question was then dealt with using linear Canonical Correlation, and related techniques, in order to estimate the weights of the indices for men and women. This paper argues that this estimation technique is inconsistent and suggest some solutions.

Since Becker's (1973) seminal contribution, the marriage market has been predominantly modeled as a matching market with transferable utility. Men and women are characterized by vectors of attributes denoted respectively $x \in \mathbb{R}^{d_x}$ for men and $y \in \mathbb{R}^{d_y}$ for women. These vectors may incorporate various dimensions such as education, wealth, health, physical attractiveness, personality traits, etc. It is assumed that when a man with attributes x and a woman with attributes y form a pair, they generate a surplus equal to $\Phi(x, y)$. This surplus is shared endogenously between the two partners. Denoting P and Q the respective probability distributions of attributes of married men and women, it follows from the results of Shapley and Shubik (1972) that the stable matching will maximize

$$\mathbb{E}\left[\Phi(X, Y)\right]$$

with respect to all joint distributions of (X, Y) such that $X \sim P$ and $Y \sim Q$. For convenience, we assume that these distributions are centered $\int x dP(x) = \int y dQ(y) = 0$.

Becker went further in the analysis by assuming that sorting occurs on single-dimensional *ability indices* for men and women, say $\bar{x}$ and $\bar{y}$, which are constructed linearly with respect to the original attributes

$$\bar{x} = \alpha' x \text{ and } \bar{y} = \beta' y$$

where $\alpha \in \mathbb{R}^{d_x}$ and $\beta \in \mathbb{R}^{d_y}$ are the weights according to which the various attributes enter the respective indices. Following Becker (1973), assume that the matching surplus of individuals of attributes x and y, denoted $\Phi(x, y)$, only depends on the indices $\bar{x}$ and $\bar{y}$ and takes the form

$$\Phi(x, y) = \phi\left(\alpha' x, \beta' y\right)$$

where ϕ is supermodular, that is $\partial^2_{\bar{x},\bar{y}} \phi(\bar{x}, \bar{y}) \geq 0$. As a result, the solution exhibits positive assortative matching, that is, the equilibrium distribution of the attributes across couples is represented by a joint random vector $(X, Y) \sim \pi$ where $\alpha' X$ and $\beta' Y$ are *comonotone*: the man at percentile t in the distribution of $\alpha' X$ is matched with the woman at percentile t in the distribution of $\beta' Y$. In other words, denoting F_Z the cumulative distribution function of Z, we can state as the main assumption of this note that:

Assumption 1. *There are weights α and β such that the indices $\alpha' X$ and $\beta' Y$ are comonotone, that is*

$$F_{\beta' Y}\left(\beta' Y\right) = F_{\alpha' X}\left(\alpha' X\right).$$

If the cumulative distribution function $F_{\beta' Y}$ is invertible, one may then write

$$\beta' Y = T\left(\alpha' X\right)$$

where $T(z) = F^{-1}_{\beta' Y} \circ F_{\alpha' X}(z)$ is a nondecreasing map; thus the ability index of a woman is a nondecreasing function of that of the man she is matched with.

Given this specification and the observation of $(X, Y) \sim \pi$, one would like to estimate (α, β). To this end, Becker (1973) suggested (p. 834) to use Canonical Correlation Analysis, a technique originally introduced by Hotelling (1936). This method consists in determining the weights α^c and β^c that maximize the correlation between $\alpha' X$ and $\beta' Y$.[1] Formally, introducing the following notations

$$\Sigma_{XY} = \mathbb{E}_\pi\left[XY'\right],\ \Sigma_X = \mathbb{E}_\pi\left[XX'\right],\ \Sigma_Y = \mathbb{E}_\pi\left[YY'\right],$$

[1] Since we are primarily interested about the consistency of Canonical Correlation and related techniques, throughout this paper we assume that the analyst has access to a sample of infinite size.

Canonical Correlation consists in defining α^c and β^c as the maximizers of the correlation of $\alpha' X$ and $\beta' Y$ over all possible vectors of weights α and β. The problem therefore consists in solving the following program

$$\begin{aligned} &\max_{\alpha\in\mathbb{R}^{d_x},\beta\in\mathbb{R}^{d_y}} \alpha'\Sigma_{XY}\beta \\ &s.t.\ \alpha'\Sigma_X\alpha = 1 \text{ and } \beta'\Sigma_Y\beta = 1 \end{aligned} \tag{1}$$

whose value at optimum is in general less or equal to one.

In the applied literature, α and β are frequently estimated by multivariate Ordinary Least Squares (OLS) regression. It is worth remarking that this is closely related, but not quite identical, to Canonical Correlation. Consider the following OLS regression

$$Y_1 = \alpha' X - \beta_{-1}^{'} Y_{-1} + \varepsilon$$

where ε is an error term, Y_1 is the top element of Y, and Y_{-1} the vector of the remaining entries. Let α^o and β^o_{-1} be the coefficients obtained from OLS. Introducing $\beta^o = \left(1\ \beta^{o\prime}_{-1}\right)'$, it is easy to show that (α^o, β^o) solves the program

$$\begin{aligned} &\max_{\alpha\in\mathbb{R}^{d_x},\beta\in\mathbb{R}^{d_y}} \alpha'\Sigma_{XY}\beta \\ &s.t.\ \alpha'\Sigma_X\alpha = A \text{ and } \beta'\Sigma_Y\beta = B \text{ and } \beta_1 = 1. \end{aligned}$$

where $A = \alpha^{o\prime}\Sigma_X\alpha^o$ and $B = \beta^{o\prime}\Sigma_Y\beta^o$. Without the constraint $\beta_1 = 1$, this would yield the same solutions (up to some rescaling of α and β) as the solutions given by Canonical Correlation. In general, the solutions differ due to this constraint. Even though the OLS technique is better known and more immediately accessible to practitioners, it artificially breaks down symmetry between variables by singling out the role of Y_1. Note that in the case where Y is univariate ($d_y = 1$) the constraint $\beta_1 = 1$ has no bite, and the two solutions coincide (again, up to rescaling).

Many papers have used Canonical Correlation or OLS techniques to estimate α and β. Notable examples of the application of Canonical Correlation on the marriage market are Suen and Lui (1999), Gautier et al. (2005) and Taubman (2006). Many papers have applied OLS techniques to study assortative mating when faced with multiple dimensions, see Kalmijn (1998) for a survey of this literature. A notable example of such applications of OLS is the extensive literature on the effect of a wife's education on her husband's earnings: see among others Benham (1974), Scully (1979), Wong (1986), Lam and Schoeni (1993, 1994), and Jepsen (2005).

The consistency problem. A crucial question is whether the Canonical Correlation method is consistent, namely whether $(\alpha^c, \beta^c) = (\alpha, \beta)$. It turns out that the answer is yes in the case of Gaussian marginal distributions P and Q, but no in more general cases as we shall now explain. We now state our result. The main statement, part (ii) of the theorem, is proven using a counterexample.

Theorem 1 ((In-)Consistency of Canonical Correlation)**.** *The following holds:*

(i) If P and Q are Gaussian distributions, then the Canonical Correlation is consistent in the sense that

$$(\alpha^c, \beta^c) = (\alpha, \beta)\,.$$

(ii) In general, Canonical Correlation is not consistent.

Proof. (i) When $P = N(0, \Sigma_X)$ and $Q = N(0, \Sigma_Y)$, with $\alpha, \beta \neq 0$ two vectors of weights, then

$$\max_{X \sim P, Y \sim Q} \mathbb{E}\left[\alpha' X Y' \beta\right] = \sqrt{\alpha' \Sigma_X \alpha}\sqrt{\beta' \Sigma_Y \beta},$$

where the optimization is over the set of random vectors (X, Y) with fixed marginal distributions P and Q. Thus, for $(X\,, Y)$ solution of the above problem, the correlation between $\alpha' X$ and $\beta' Y$ is one. Indeed, the optimal (X, Y) is such that

$$\beta' Y = \sqrt{\frac{\beta' \Sigma_Y \beta}{\alpha' \Sigma_X \alpha}} \alpha' X.$$

The result is immediate: for the optimal (X, Y), the correlation between $\alpha' X$ and $\beta' Y$ is one and since this is the maximal value of Program (1), it follows that $(\alpha, \beta) = (\alpha^c, \beta^c)$.

(ii) However, when P and Q fail to be Gaussian, the Canonical Correlation estimator (α^c, β^c) differs from the true parameters (α, β) in general. Consider the following example. Let P be the distribution of (X_1, X_2) where X_1 is independent of X_2 and $V(X_1) = V(X_2) = 1$. Let Q be the distribution of Y. Provided that the surplus function satisfies $\Phi(x, y) = \phi(\alpha' x, \beta' y)$ such that sorting is unidimensional, optimal matching yields: $Y = \frac{T(\alpha_1 X_1 + \alpha_2 X_2)}{\beta}$ where $T := F^{-1}_{\alpha' X}(F_{\beta Y}(.))$ and $F_{\gamma' Z}$ denotes the c.d.f. of $\gamma' Z$. Note that the mapping function T depends on P, Q, α and β. In this setting, the Canonical Correlation estimator (α_1^c, α_2^c) of (α_1, α_2) solves

$$\max_{\alpha_1, \alpha_2} \alpha_1 cov(X_1, Y) + \alpha_2 cov(X_2, Y)$$
$$s.t. \ \alpha_1^2 + \alpha_2^2 = 1$$

whose solution is

$$\frac{\alpha_1^c}{\alpha_2^c} = \frac{cov(X_1 Y)}{cov(X_2 Y)}. \tag{2}$$

In such an economy, data on "couples" are such that $Y = \frac{T(\alpha_1 X_1 + \alpha_2 X_2)}{\beta}$ for all X. Replacing Y by its expression in terms of X in the right hand side of eq. 2 yields

$$\frac{\alpha_1^c}{\alpha_2^c} = \frac{cov(X_1 T(\alpha_1 X_1 + \alpha_2 X_2))}{cov(X_2 T(\alpha_1 X_1 + \alpha_2 X_2))}.$$

It follows that the Canonical Correlation estimator is consistent if and only if

$$\frac{\alpha_1^c}{\alpha_2^c} = \frac{\alpha_1}{\alpha_2}$$

that is if and only if

$$\frac{\alpha_1}{\alpha_2} = \frac{cov(X_1 T(\alpha_1 X_1 + \alpha_2 X_2))}{cov(X_2 T(\alpha_1 X_1 + \alpha_2 X_2))}. \tag{3}$$

It is easy to see that this condition will be satisfied whenever T is linear (with constant a and slope b), a case that arises for instance when P and Q are Gaussian as in i), for then one has

$$\frac{cov(X_1 (a + b\alpha_1 X_1 + b\alpha_2 X_2))}{cov(X_2 (a + b\alpha_1 X_1 + b\alpha_2 X_2))} = \frac{\alpha_1}{\alpha_2}.$$

However, as soon as T is nonlinear, there are no reasons to expect that T will satisfy condition 3. For instance, let P be the distribution of (X_1, X_2) where X_1 takes value 1 with probability 1/2 and -1 with probability 1/2, and X_2 is exponentially distributed with parameter 1 and independent of X_1. Let G be the c.d.f. of X_2, so that $G(z) = 1 - \exp(-z)$.

Let $Q = \mathcal{U}([0,1])$. Set $\alpha_1 = \alpha_2 = 1/\sqrt{2}$, so that $\hat{X} = \frac{X_1+X_2}{\sqrt{2}}$. Hence the optimal coupling $\left(\hat{X}, \hat{Y}\right)$ is such that $\hat{Y} = F_{\hat{X}}\left(\hat{X}\right)$ where $F_{\hat{X}}(.)$ is the c.d.f. of $\hat{X}$, which is expressed as

$$F_{\hat{X}}(x) = \frac{1}{2}\left(G\left(x\sqrt{2}+1\right) + G\left(x\sqrt{2}-1\right)\right).$$

Clearly, in this example $T := F_{\hat{X}}$ is not linear in this case such that one should not expect Canonical Correlation to be consistent. And, indeed, one has

$$\hat{Y} = \begin{cases} \frac{1}{2}\left(G(X_2) + G(X_2 - 2)\right) & \text{if } X_1 = -1 \\ \frac{1}{2}\left(G(X_2+2) + G(X_2)\right) & \text{if } X_1 = 1, \end{cases}$$

and a calculation shows that

$$cov\left(X_1, \hat{Y}\right) = \frac{\mathbb{E}G(X_2+2) - \mathbb{E}G(X_2-2)}{4}$$

and as $\mathbb{E}G(X_2+2) = 1 - e^{-2}/2$ and $\mathbb{E}G(X_2-2) = e^{-2}/2$, we get

$$cov\left(X_1, \hat{Y}\right) = \frac{1}{4}\left(1 - e^{-2}\right). \tag{4}$$

Similarly,

$$\mathbb{E}\left[X_2\hat{Y}\right] = \frac{1}{4}\mathbb{E}\left[X_2 G(X_2-2)\right] + \frac{1}{4}\mathbb{E}\left[X_2 G(X_2+2)\right] + \frac{1}{2}\mathbb{E}\left[X_2 G(X_2)\right]$$

and using the fact that $\mathbb{E}\left[X_2 G(X_2-2)\right] = 7e^{-2}/4$, that $\mathbb{E}\left[X_2 G(X_2+2)\right] = 1 - e^{-2}/4$, and that $\mathbb{E}\left[X_2 G(X_2)\right] = 3/4$, we get $\mathbb{E}\left[X_2\hat{Y}\right] = \left(3e^{-2}+5\right)/8$, hence, as $\mathbb{E}\left[X_2\right]\mathbb{E}\left[\hat{Y}\right] = 1/2$, one obtains

$$cov\left(X_2, \hat{Y}\right) = \frac{3e^{-2}+1}{8}. \tag{5}$$

Using (4) and (5), this becomes

$$\frac{\alpha_2^c}{\alpha_1^c} = \frac{3+e^2}{2e^2-2} \neq \frac{\alpha_2}{\alpha_1} = 1.$$

Therefore the Canonical Correlation estimator is not consistent in this example. □

Note that the example in part (ii) of the proof also shows that OLS is inconsistent. In this example the dimension of Y is one, so that OLS and Canonical Correlation yield the same estimators of α and β. The above example has nothing pathological and implies that estimators of (α, β) based on Canonical Correlation face the risk of being biased as soon as the marginal distributions are not Gaussian or such that the mapping function $T := F_{\alpha' X}^{-1}\left(F_{\beta' Y}(.)\right)$ is not linear.

Final remarks. The problem discussed in this paper obviously raises the question: how can we replace Canonical Correlation by a technique that is consistent? One first proposal, as suggested in Terviö (2003, p. 83), is to look for α and β that maximize Spearman's rank correlation between $\alpha' X$ and $\beta' Y$. In other words, look for

$$\begin{aligned}&\max_{\alpha\in\mathbb{R}^{d_x},\beta\in\mathbb{R}^{d_y}} \mathbb{E}\left[F_{\alpha' X}\left(\alpha' X\right) F_{\beta' Y}\left(\beta' Y\right)\right]\\ &s.t.\ \alpha'\Sigma_X\alpha = 1 \text{ and } \beta'\Sigma\beta = 1.\end{aligned}$$

where we recall that $F_{\alpha' X}$ stands for the c.d.f. of $\alpha' X$. The value of this program cannot exceed 1/3 and, when the distributions of X and Y are continuous, it is equal to 1/3 when $\alpha' X$ and $\beta' Y$ are comonotone. However the objective function, which can be rewritten as

$$\int \Pr\left(\max\left(\alpha'\left(x-X\right), \beta'\left(y-Y\right)\right) \leq 0\right) dF_X\left(x\right) dF_Y\left(y\right),$$

has no reason to be convex with respect to α and β, so global optimization techniques may be needed. Also, this technique, just as Canonical Correlation, does not deal with any kind of unobserved heterogeneity. To remedy this drawback, two solutions have very recently been proposed.

The first solution is justified if one is willing to assume that sorting occurs on a single index of attractiveness. This strategy, developed by Chiappori et al. (2012), consists in estimating the conditional expectations $\mathbb{E}\left[Y_k|X=x\right]$, which, if the sorting actually occurs on a single-index, should be a deterministic function of $\alpha' X$. Hence the weight vector α is identified up to a constant by the marginal rates of substitutions

$$\frac{\alpha_i}{\alpha_j} = \frac{\partial\mathbb{E}\left[Y_k|X=x\right]/\partial x_i}{\partial\mathbb{E}\left[Y_k|X=x\right]/\partial x_j}.$$

Moving outside of single-dimensional indices, Dupuy and Galichon (2014) have introduced a technique they call "saliency analysis", which allows to infer the *number of dimensions* on which sorting occurs, and estimate the corresponding (possibly multiple) *indices of attractiveness* that determine this sorting.

Saliency analysis requires first to obtain a consistent estimate of A in the quadratic specification for the surplus function

$$\Phi(x, y) = x' A y.$$

This is done by applying for instance the procedure depicted in Dupuy and Galichon (2014). Dupuy and Galichon (2014) show that this estimate of A is asymptotically normal and provide an expression for its asymptotic variance-covariance matrix.

To understand the intuition for the second step of saliency analysis, it is important to note that when A is of rank 1, one has $A = \alpha\beta'$ such that sorting occurs on single-dimensional indices $\alpha' x$ and $\beta' y$ and, the procedure depicted in Dupuy and Galichon (2014) yields consistent estimates of α and β. Nevertheless, there is generally no reason to expect A to be of rank 1. In fact, one would like to test the rank of A as this test would be informative about whether sorting is unidimensional or not. Drawing insights from the literature on matrix rank tests, the second step of saliency analysis consists in performing the singular value decomposition of A, which yields[2]

$$A = U' \Lambda V,$$

where the diagonal matrix Λ has nonincreasing elements $(\lambda_1, ..., \lambda_d)$ called singular values, $d = \min(d_x, d_y)$, on its diagonal. By construction, U and V are orthogonal matrices. The (estimates of) matrices U, V and Λ are obtained from applying standard numerical methods performing the singular value decomposition of (the estimate of) A and pre-programmed in many statistical packages such as MATLAB or STATA.

The singular value decomposition of A not only provides the building blocks to test for the number of sorting dimensions, i.e. the number of non zero singular values, it allows the

[2]Note that the units of the parameters of A reflect the units in which X and Y are measured. For our method to be robust to changes in measurement units, we need to normalize the attributes in X and Y. By performing the singular value decomposition of the matrix A associated with the normalized attributes, we ensure that the loadings of the indices of mutual attractiveness are independent of the choice of measurement units. For the sake of notation and compactness we herewith simply assume that X and Y have been rescaled such that all attributes are of variance 1.

construction of *indices of mutual attractiveness* for men and women. To see this, define vectors of *indices of mutual attractiveness* as

$$\tilde{X} = UX \text{ and } \tilde{Y} = VY,$$

where each index is a weighted sum of the attributes in X and Y respectively. Dupuy and Galichon (2014) have shown that

$$\Phi(x, y) = \sum_{i=1}^{d_x} \sum_{j=1}^{d_y} A_{ij} x_i y_j = \sum_{i=1}^{d} \lambda_i \tilde{x}_i \tilde{y}_i.$$

The weights of each index of mutual attractiveness constructed by saliency analysis can be read on the associated row of U for men and V for women whereas the share of the matching utility of couples explained by the i^{th} pair of indices is given by $\lambda_i / (\sum_i \lambda_i)$. Therefore, saliency analysis answers the two important questions mentioned in the introduction: how many and which attributes matter for the sorting of men and women on the marriage market. Intuitively, the number of non zero singular values indicates the number of indices that matter for the sorting problem and the parameters of U and V indicate which attributes matter in each index of men and each index of women. If there is only one non zero singular value, then sorting occurs on a single index whose weights are given by the first row of U for men and V for women and correspond to α and β respectively.

References

[1] Becker, Gary (1973). "A theory of marriage, part I," *Journal of Political Economy*, 81, pp. 813-846.

[2] Benham, Lee (1974). "Benefits of womens education within marriage," *Journal of Political Economy*, 82(2), pp. S57–S71.

[3] Chiappori, Pierre-Andre., Oreffice, Sonia and Quintana-Domeque, Clement (2012). "Fatter attraction: anthropometric and socioeconomic matching on the marriage market," *Journal of Political Economy*, 120(4), pp. 659–695.

[4] Dupuy, Arnaud, and Alfred Galichon (2014). "Personality traits and the marriage market," to appear in the *Journal of Political Economy.*

[5] Gautier, Pieter, Michael Svarer, and Coen Teulings (2010). "Marriage and the city: Search frictions and sorting of singles," *Journal of Urban Economics* 67(2), pp. 206–218.

[6] Hotelling, Harold (1936). "Relations between two sets of variates," *Biometrika* 28, pp. 321–329.

[7] Jepsen, Lisa (2005). "The relationship between wifes education and husbands earnings: Evidence from 1960–2000," *Review of Economics of the Household* 3, pp. 197–214.

[8] Kalmijn, Matthijs (1998). "Intermarriage and Homogamy: Causes, Patterns, Trends," *Annual Review of Sociology* 24, pp. 395–421.

[9] Lam, David, and Robert Schoeni (1993). "Effects of family background on earnings and returns to schoolings: Evidence from Brazil," *Journal of Political Economy* 101 (4), pp. 710–740.

[10] Lam, David, and Robert Schoeni (1994). "Family ties and labour markets in the United States and Brazil," *Journal of Human Resources* 29, pp. 1235–1258.

[11] Scully, Gerald (1979). "Mullahs, Muslims and marital sorting," *Journal of Political Economy* 87, pp. 1139–1143.

[12] Shapley, Lloyd, and Martin Shubik (1972). "The Assignment Game I: The Core," *International Journal of Game Theory* 1, pp. 111–130.

[13] Suen, Wing, and Hon-Kwong Lui (1999). "A direct test of efficient marriage market hypothesis," *Economic Inquiry* 37 (I), pp. 29–46.

[14] Taubman, Orit (2006). "Couple similarity for driving style," *Transportation Research Part F* 9, pp. 185–193.

[15] Terviö, Marko. (2003). "Studies of Talent Markets," MIT PhD. dissertation.

[16] Wong, Yue-Chim. (1986). "Entrepreneurship, Marriage, and Earnings," *Review of Economics and Statistics* 31 1-23, 693–99.

CEPS
INSTEAD

3, avenue de la Fonte
L-4364 Esch-sur-Alzette
Tél.: +352 58.58.55-801
www.ceps.lu