‘MATH+ECON+CODE’ MASTERCLASS ON MATCHING MODELS, OPTIMAL TRANSPORT AND APPLICATIONS

Alfred Galichon (New York University)

Day 5: “Empirical matching models”
Block 15. Rank constained models

Learning objectives: block 15

- regularized optimal transport
- the gravity equation
- generalized linear models
- pseudo-Poisson maximum likelihood estimation
- affinity matrix
- index models
- rank-constraint models

Section 1

PART I: THE GRAVITY MODEL

- Anderson and van Wincoop (2003). "Gravity with Gravitas: A Solution to the Border Puzzle". *AER*.
- Head and Mayer (2014). "Gravity Equations: Workhorse, Toolkit and Cookbook". *Handbook of international economics.*
- Gourieroux, Trognon, Monfort (1984). "Pseudo Maximum Likelihood Methods: theory" *Econometrica*.
- McCullagh and Nelder (1989). *Generalized Linear Models*. Chapman and Hall/CRC.
- Santos Silva and Tenreyro (2006). "The Log of Gravity". *REStats*.
- Yotov et al. (2011). *An advanced guide to trade policy analysis*. WTO.
- Guimares and Portugal (2012). "Real Wages and the Business Cycle: Accounting for Worker, Firm, and Job Title Heterogeneity". *AEJ: Macro*.
- Dupuy and G (2014), "Personality traits and the marriage market". *JPE*.
- Dupuy, G and Sun (2016), "Estimating matching affinity matrix under low-rank constraints." arxiv 1612.09585.

- The gravity equation is a very useful tool for explaining trade flows by various measures of proximity between countries.
- A number of regressors have been proposed. They include: geographic distance, common official languague, common colonial past, share of common religions, etc.
- The dependent variable is the volume of exports from country i to country n, for each pair of country (i, n).
- Today, we shall see a close connection between gravity models of international trade and separable matching models.

- Consider the optimal transport duality

$$\max_{\pi \in \mathcal{M}(P,Q)} \sum_{xy} \pi_{xy} \Phi_{xy} = \min_{u_x + v_y \geq \Phi_{xy}} \sum_{x \in \mathcal{X}} p_x u_x + \sum_{y \in \mathcal{Y}} q_y v_y$$

- Now let's assume that we are adding an entropy to the primal objective function. For any $\sigma > 0$, we get

$$\begin{aligned} &\max_{\pi \in \mathcal{M}(P,Q)} \sum_{xy} \pi_{xy} \Phi_{xy} - \sigma \sum_{xy} \pi_{xy} \ln \pi_{xy} \\ &= \min_{u,v} \sum_{x \in \mathcal{X}} p_x u_x + \sum_{y \in \mathcal{Y}} q_y v_y + \sigma \sum_{xy} \exp\left(\frac{\Phi_{xy} - u_x - v_y - \sigma}{\sigma} \right) \end{aligned}$$

- The latter problem is an unconstrained convex optimization problem. But the most efficient numerical computation technique is often coordinate descent, i.e. alternate between minimization in u and minimization in v.

- Maximize wrt to u yields

$$e^{-u_x/\sigma} = \frac{p_x}{\sum_y \exp\left(\frac{\Phi_{xy} - v_y - \sigma}{\sigma}\right)}$$

and wrt v yields

$$e^{-v_y/\sigma} = \frac{q_y}{\sum_x \exp\left(\frac{\Phi_{xy} - v_y - \sigma}{\sigma}\right)}$$

- It is called the "iterated projection fitting procedure" (ipfp), aka "matrix scaling", "RAS algorithm", "Sinkhorn-Knopp algorithm", "Kruithof's method", "Furness procedure", "biproportional fitting procedure", "Bregman's procedure". See survey in Idel (2016).
- Maybe the most often reinvented algorithm in applied mathematics. Recently rediscovered in a machine learning context.

- The goal is to estimate the matching surplus Φ_{xy}. For this, take a linear parameterization

$$\Phi_{xy}^{\beta}=\sum_{k=1}^{K}\beta_k\phi_{xy}^{k}.$$

- Following Choo and Siow (2006), G and Salanié (2017) introduce logit heterogeneity in individual preferences and show that the equilibrium now maximizes the *regularized Monge-Kantorovich problem*

$$W\left(\beta\right)=\max_{\pi\in\mathcal{M}(P,Q)}\sum_{xy}\pi_{xy}\Phi_{xy}^{\beta}-\sigma\sum_{xy}\pi_{xy}\ln\pi_{xy}$$

- By duality, $W\left(\beta\right)$ can be expressed

$$W\left(\beta\right)=\min_{u,v}\sum_{x}p_x u_x+\sum_{y}q_y v_y+\sigma\sum_{xy}\exp\left(\frac{\Phi_{xy}^{\beta}-u_x-v_y-\sigma}{\sigma}\right)$$

and w.l.o.g. can set $\sigma=1$ and drop the additive constant $-\sigma$ in the exp.

- We observe the actual matching $\hat{\pi}_{xy}$. Note that $\partial W / \partial \beta^k = \sum_{xy} \pi_{xy} \phi_{xy}^k$, hence β is estimated by running

$$\min_{u,v,\beta} \sum_x p_x u_x + \sum_y q_y v_y + \sum_{xy} \exp\left(\Phi_{xy}^\beta - u_x - v_y\right) - \sum_{xy,k} \hat{\pi}_{xy} \beta_k \phi_{xy}^k \tag{1}$$

which is still a convex optimization problem.

- This is actually the objective function of the log-likelihood in a Poisson regression with x and y fixed effects, where we assume

$$\pi_{xy} | xy \sim \mathit{Poisson}\left(\exp\left(\sum_{k=1}^K \beta_k \phi_{xy}^k - u_x - v_y\right)\right).$$

- Let $\theta = (\beta, u, v)$ and $Z = (\phi, D^x, D^y)$ where $D^x_{x'y'} = 1\{x = x'\}$ and $D^y_{x'y'} = 1\{y = y'\}$ are x-and y-dummies. Let $m_{xy}(Z;\theta) = \exp(\theta^\intercal Z_{xy})$ be the parameter of the Poisson distribution.
- The conditional likelihood of $\hat{\pi}_{xy}$ given Z_{xy} is

$$\begin{aligned} l_{xy}(\hat{\pi}_{xy};\theta) &= \hat{\pi}_{xy}\log m_{xy}(Z;\theta) - m_{xy}(Z;\theta) \\ &= \hat{\pi}_{xy}(\theta^\intercal Z_{xy}) - \exp(\theta^\intercal Z_{xy}) \\ &= \hat{\pi}_{xy}\left(\sum_{k=1}^{K}\beta_k\phi^k_{xy} - u_x - v_y\right) - \exp\left(\sum_{k=1}^{K}\beta_k\phi^k_{xy} - u_x - v_y\right) \end{aligned}$$

- Summing over x and y, the sample log-likelihood is

$$\sum_{xy}\hat{\pi}_{xy}\sum_{k=1}^{K}\beta_k\phi^k_{xy} - \sum_x p_x u_x - \sum_y q_y v_y - \sum_{xy}\exp\left(\sum_{k=1}^{K}\beta_k\phi^k_{xy} - u_x - v_y\right)$$

hence we recover objective function (1).

- If $\pi_{xy}|xy$ is Poisson, then $\mathbb{E}\left[\pi_{xy}\right] = m_{xy}\left(Z_{xy};\theta\right) = \mathbb{V}ar\left(\pi_{xy}\right)$. While it makes sense to assume the former equality, the latter is a rather strong assumption.
- For estimation purposes, $\hat{\theta}$ is obtained by

$$\max_{\theta} \sum_{xy} l\left(\hat{\pi}_{xy};\theta\right) = \sum_{xy}\left(\hat{\pi}_{xy}\left(\theta^{\intercal} Z_{xy}\right) - \exp\left(\theta^{\intercal} Z_{xy}\right)\right)$$

however, for inference purposes, one shall not assume the Poisson distribution. Instead

$$\sqrt{N}\left(\hat{\theta}-\theta\right) \Longrightarrow \left(A_0\right)^{-1} B_0 \left(A_0\right)^{-1}$$

where $N = |\mathcal{X}| \times |\mathcal{Y}|$ and A_0 and B_0 are estimated by

$$\hat{A}_0 = N^{-1}\sum_{xy} D^2_{\theta\theta} l\left(\hat{\pi}_{xy};\hat{\theta}\right) = N^{-1}\sum_{xy}\exp\left(\hat{\theta}^{\intercal} Z_{xy}\right) Z_{xy} Z_{xy}^{\intercal}$$

$$\hat{B}_0 = N^{-1}\sum_{xy}\left(\hat{\pi}_{xy} - \exp\left(\hat{\theta}^{\intercal} Z_{xy}\right)\right)^2 Z_{xy} Z_{xy}^{\intercal}.$$

- Dupuy and G (2014) focus on cross-dimensional interactions

$$\phi^A_{xy} = \sum_{p,q} A_{pq} \xi^p_x \xi^q_y$$

  and estimate "affinity matrix" A on a dataset of married individuals where the "big 5" personality traits are measured.
- A is estimated by

$$\min_{s_i, m_n} \min_{A} \left\{ \begin{array}{c} \sum_x p_x u_x + \sum_y q_y v_y \\ + \sum_{xy} \exp\left(\sum_{p,q} A_{pq} \xi^p_x \xi^q_y - u_x - v_y \right) \\ - \sum_{x,y,p,q} \hat{\pi}_{xy} A_{pq} \xi^p_x \xi^q_y \end{array} \right\}.$$

- Dupuy, G and Sun (2016) consider the case when the space of characteristics is high-dimensional. More on this this afternoon.

TABLE: Affinity matrix. Source: Dupuy and G (2014).

Wives Husbands	Education	Height.	BMI	Health	Consc.	Extra.	Agree.	Emotio.	Auto.	Risk
Education	**0.46**	0.00	**-0.06**	0.01	-0.02	0.03	-0.01	-0.03	0.04	0.01
Height	0.04	**0.21**	0.04	0.03	**-0.06**	0.03	0.02	0.00	-0.01	0.02
BMI	-0.03	0.03	**0.21**	0.01	0.03	0.00	**-0.05**	0.02	0.01	-0.02
Health	-0.02	0.02	- 0.04	**0.17**	- 0.04	0.02	- 0.01	0.01	-0.00	0.03
Conscienciousness	**-0.07**	-0.01	**0.07**	-0.00	**0.16**	0.05	0.04	0.06	0.01	0.01
Extraversion	0.00	-0.01	0.00	0.01	**-0.06**	**0.08**	-0.04	-0.01	0.02	**-0.06**
Agreeableness	0.01	0.01	-0.06	0.02	**0.10**	**-0.11**	0.00	0.07	-0.07	-0.05
Emotional	0.03	-0.01	0.04	**0.06**	**0.19**	0.04	0.01	-0.04	**0.08**	**0.05**
Autonomy	0.03	0.02	0.01	0.02	**-0.09**	**0.09**	-0.04	0.02	**-0.10**	0.03
Risk	0.03	-0.01	-0.03	-0.01	0.00	-0.02	-0.03	-0.03	**0.08**	**0.14**

Note: Bold coefficients are significant at the 5 percent level.

- "Structural gravity equation" (Anderson and van Wincoop, 2003) as reviewed in Head and Mayer (2014) handbook chapter:

$$X_{ni} = \underbrace{\frac{Y_i}{\Omega_i}}_{S_i} \underbrace{\frac{X_n}{\Psi_n}}_{M_n} \Phi_{ni}$$

  where n=importer, i=exporter, X_{ni}=trade flow from i to n, $Y_i = \sum_n X_{ni}$ is value of production, $X_n = \sum_i X_{ni}$ is importers' expenditures, and ϕ_{ni}=bilateral accessibility of n to i.

- Ω_i and Ψ_n are "multilateral resistances", satisfying the set of implicit equations

$$\Psi_n = \sum_i \frac{\Phi_{ni} Y_i}{\Omega_i} \text{ and } \Omega_i = \sum_n \frac{\Phi_{ni} X_n}{\Psi_n}$$

- These are exactly the same equations as those of the regularized OT.

- Parameterize $\Phi_{ni} = \exp\left(\sum_{k=1}^{K} \beta_k D_{ni}^k\right)$, where the D_{ni}^k are K pairwise measures of distance between n and i. We have

$$X_{ni} = \exp\left(\sum_{k=1}^{K} \beta_k D_{ni}^k - s_i - m_n\right)$$

where fixed effects $s_i = -\ln S_i$ and $m_n = -\ln M_n$ are adjusted by

$$\sum_i X_{ni} = Y_i \text{ and } \sum_n X_{ni} = X_n.$$

- Standard choices of D_{ni}^k's:
  - logarithm of bilateral distance between n and i
  - indicator of contiguous borders; of common official language; of colonial ties
  - trade policy variables: presence of a regional trade agreement; tariffs
  - could include many other measures of proximity, e.g. measure of genetic/cultural distance, intensity of communications, etc.

Section 2

Part II: rank-constrained models

- Becker (1973). A Theory of Marriage: Part I. *JPE*.
- [COQ] Chiappori, Oreffice and Quintana-Domeque (2012). "Fatter Attraction: Anthropometric and Socioeconomic Matching on the Marriage Market," *Journal of Political Economy*.

- Chiappori, Oreffice and Quintana-Domeque [COQ] assume individuals match on a scalar "index of attractiveness" subsuming BMI, salary, education. Then the surplus function is

$$\Phi(x, y) = \left(\sum_k \zeta_k x_k\right) \left(\sum_l \nu_l y_l\right)$$

  $\zeta_k / \zeta_{k'}$ and $\nu_l / \nu_{l'}$ are marginal rates of substitution: how much richer do men/women need to be in order to compensate an increase in Body Mass Index?
- This problem can be solves by looking for the vectors of weights ζ and ν such that the rank correlation of $\zeta^\intercal x$ and $\nu^\intercal y$ is maximal.

- [COQ] look at the characteristics of married couples, in particular body mass index, wages, and education.
- According to [COQ] (*Journal of Political Economy*, 2012): "Men may compensate 1.3 additional units of BMI with a 1%-increase in wages, while women may compensate two BMI units with one year of education."

- Recall

$$\mathcal{W}(A) = \max_{\pi \in \mathcal{M}(P,Q)} \int x' A y d\pi(x,y) - \sigma \int \pi(x,y) \, d\pi(x,y).$$

and note that

$$\frac{\partial \mathcal{W}(A)}{\partial A_{ij}} = C_{ij}^{A}$$

- We are therefore looking for the estimator $\hat{A}$ of the true A such that $\partial \mathcal{W}(A) / \partial A_{ij} = \hat{C}_{ij}$.
- Thus we shall estimate A by $\hat{A}$ the solution of

$$\min_{A} \left\{ \mathcal{W}(A) - \sum_{ij} A_{ij} \hat{C}_{ij} \right\}$$

which is a nice and smooth convex minimization problem.

- Several proposal to estimate ζ and ν:
  1. Becker (1973): use Hotelling's canonical correlation analysis
  $$\max_{\zeta,\nu} \mathbb{E}\left[\zeta^\intercal X Y^\intercal \nu\right],$$
  which is unbiased if (X, Y) is Gaussian. Can be biased outside that case, cf. Dupuy-Galichon (AES, 2015).
  2. Chiappori, Oreffice and Quintana-Domeque (JPE 2013): when Y is 1-dimensional, regress Y on X.
  3. Terviö (AER 2007): maximize Spearman's rank correlation
  $$\max_{\zeta,\nu} \mathbb{E}\left[F_{\zeta^\intercal X}\left(\zeta^\intercal X\right) F_{\nu^\intercal Y}\left(\nu^\intercal Y\right)\right],$$
  where $F_{\zeta^\intercal X}$ and $F_{\nu^\intercal Y}$ are the cdfs of $\zeta^\intercal X$ and $\nu^\intercal Y$ respectively.
  4. In the spirit of Han (JE 1987), maximize
  $$\sum_{ij}\left(1\left\{\zeta^\intercal X_i > \zeta^\intercal X_j\right\} 1\left\{\nu^\intercal Y_i > \nu^\intercal Y_j\right\} + 1\left\{\zeta^\intercal X_i < \zeta^\intercal X_j\right\} 1\left\{\nu^\intercal Y_i < \nu^\intercal Y_j\right\}\right)$$
  5. Dupuy-Galichon-Sun (2017): perform rank-constrained estimation of $\Phi(x, y) = x' A y$ using nuclear norm regularization.

- Recall that any $d \times d$ matrix A has a singular value decomposition

$$A = U \Lambda V^{\intercal}$$

where U and V are orthogonal matrices, and $\Lambda = diag\left(\lambda_1, ..., \lambda_d\right)$ is diagonal with positive entries ordered in descending order, i.e. $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq 0$.
- Note:
  - Λ are the eigenvalues of $AA^{\intercal}$, and also of $A^{\intercal}A$.
  - If A is symmetric positive, then Λ are the eigenvalues of A
  - The rank of A is the number of nonzero entries of λ.
- The nuclear norm of A, denoted $|A|_*$, is simply the L1 norm of λ, that is

$$|A|_* = \sum_{i=1}^{d} \lambda_i.$$

  - Controlling for nuclear norm is a good proxy for controlling for rank.
  - Further, the nuclear norm is convex.

- The nuclear norm can be expressed as

$$|A|_* = \max_{U,V \in O_d} Tr\left(U^\intercal AV\right)$$

  from which its gradient may be inferred (from the envelope theorem).
- In general, one can use the nuclear norm for problems of the type

$$\min_A W\left(A\right) + \gamma \left|A\right|_*$$

  which will drive low-rank solutions.

- Usual gradient descent step:

$$x_{t+1} = x_t - \epsilon \nabla h \left(x_t \right).$$

- Proximal gradient descent step: look for x such that

$$x_{t+1} = x_t - \epsilon \nabla h \left(x_{t+1} \right).$$

- Note that the first expression cannot be recast as a minimization problem, while the second one does. Indeed, the second expression can be expressed as

$$x_{t+1} \in prox_{\epsilon h} \left(x_t \right)$$

where for a convex function f, the proximal operator is defined as

$$prox_f \left(x \right) = \arg \min_z \left\{ f \left(z \right) + \frac{1}{2} \left\| z - x \right\|^2 \right\}.$$

- The ability to recast the descent step as a minimization problem is very useful because it applies also when f is nonsmooth.

- This works when h is nonsmooth too. When $f(z) = \epsilon |z|$, we have a closed-form expression for

$$prox_{\epsilon|.|}(x) = \arg\min_z \left\{ \epsilon |z| + \frac{1}{2} \|z - x\|^2 \right\}$$

indeed, by first order conditions,

$$\begin{cases} z_k > 0 \implies \epsilon + z_k - x_k = 0 \\ z_k < 0 \implies -\epsilon + z_k - x_k = 0 \\ z_k = 0 \implies z_k - x_k \in [-\epsilon, \epsilon] \end{cases}$$

thus, we get the "soft-threshoding" operator

$$\begin{cases} x_k \in [-\epsilon, \epsilon] \implies z_k = 0 \\ x_k > \epsilon \implies z_k = x_k - \epsilon \\ x_k < -\epsilon \implies z_k = x_k + \epsilon \end{cases}$$

- When the argument of f is a square matrix M, and $f(M) = |M|_*$ where $|M|_*$ is the sum of the singular values of M. Recall, if $M = U\Lambda V^\top$, where $\Lambda = diag(\lambda_1, \lambda_2, ...\lambda_d)$ with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d \geq 0$, then $|M|_* = Tr(\Lambda) = \lambda_1 + \lambda_2 + ... + \lambda_d$.
- When $f(M) = \epsilon |M|_*$, we have a closed-form expression for

$$prox_{\epsilon|.|_*}(A) = \arg\min_M \left\{ \epsilon |M|_* + \frac{1}{2} \|M - A\|^2 \right\}$$

indeed, by first order conditions,

$$\begin{cases} z_k > 0 \implies \epsilon + z_k - x_k = 0 \\ z_k < 0 \implies -\epsilon + z_k - x_k = 0 \\ z_k = 0 \implies z_k - x_k \in [-\epsilon, \epsilon] \end{cases}$$

thus, we get the "soft-threshoding" operator

$$\begin{cases} x_k \in [-\epsilon, \epsilon] \implies z_k = 0 \\ x_k > \epsilon \implies z_k = x_k - \epsilon \\ x_k < -\epsilon \implies z_k = x_k + \epsilon \end{cases}$$

- Consider $\min g(x) + h(x)$, where g is convex and differentiable, and h is convex and possibly nonsmooth.
- Standard gradient descent: $x_{t+1} = x_t - \epsilon \nabla g(x_t) - \epsilon \nabla h(x_t)$
- Proximal gradient descent: $x_{t+1} = x_t - \epsilon \nabla g(x_t) - \epsilon \nabla h(x_{t+t})$, that is $x_{t+1} + \epsilon \nabla h(x_{t+1}) = x_t - \epsilon \nabla g(x_t)$, or in other words

$$x_{t+1} = prox_{\epsilon h}\left(x_t - \epsilon \nabla g(x_t)\right)$$

- We would like to interpret $x_{t+1} = prox_{\epsilon h}\left(x_t - \epsilon \nabla g(x_t)\right)$ as $x_{t+1} = x_t - \epsilon G_\epsilon(x_t)$. For this, define

$$G_\epsilon(x) = \frac{x - prox_{\epsilon h}\left(x - \epsilon \nabla g(x)\right)}{\epsilon}$$

- **And now, let's code**!