The mass transport approach to demand inversion in multinomial choice models

Alfred Galichon (NYU Econ+Math)

Based on joint works with O. Bonnet, K. Chiong, V. Chernozhukov, M. Henry, B. Pass, and B. Salanié.

November 2017

References

- McFadden (1981). "Econometric Models of Probabilistic Choice," in C.F. Manski and D. McFadden (eds.), *Structural analysis of discrete data with econometric applications*, MIT Press.
- McFadden (1989). "A Method of Simulated Moments for Estimation of Discrete Response Models WithoutNumerical Integration". *Econometrica*.
- Berry, Levinsohn, and Pakes (1995). "Automobile Prices in Market Equilibrium," *Econometrica*.
- Train. (2009). *Discrete Choice Methods with Simulation*. 2nd Edition. Cambridge University Press.
- Chiong, G and Shum, "Duality in Discrete Choice Models". *Quantitative Economics*, 2016.
- G and Salanié (2016). "Cupid's invisible hands". Preprint.
- Bonnet, G and Shum (2017). "Yogurts choose consumers? Identification of Random Utility Models via Two-Sided Matching".
- Chernozhukov, G, Henry, and Pass (2017). Single market nonparametric identification of multi-attribute hedonic equilibrium models.

Section 1

Additive discrete choice

- Assume a consumer is facing a number of options $y \in \mathcal{Y}_0 = \mathcal{Y} \cup \{0\}$, where $y = 0$ is a default option. The consumer is drawing a utility shock which is a vector $\varepsilon = (\varepsilon_0, \ldots, \varepsilon_{|\mathcal{Y}|}) \sim \mathbf{P}$ such that the utility of option y is $U_y + \varepsilon_y$, while the outside option yields utility ε_0.
- U is called vector of *systematic utilities*; ε is called vector of *utility shocks*.
- We assume thoughout that $\mathbf{P}$ has a density with respect to the Lebesgue measure, and has full support.
- The preferred option is the one which attains the maximum in

$$\max_{y \in \mathcal{Y}} \left\{ U_y + \varepsilon_y, \varepsilon_0 \right\}.$$

- Let $s_y = \sigma_y(U)$ be the probability of choosing option y, where σ is given by

$$\sigma_y(U) = \Pr(U_y + \varepsilon_y \geq U_z + \varepsilon_z \text{ for all } z \in \mathcal{Y}_0).$$

- Note that if $s = \sigma(U)$, then $s_y > 0$ for all $y \in \mathcal{Y}_0$ and $\sum_{y \in \mathcal{Y}_0} s_y = 1$.
- Note that because the distribution $\mathbf{P}$ of ε is continuous, the probability of being indifferent between two options is zero, and hence we could have indifferently replaced weak preference $\geq$ by strict preference $>$. Without this, choice probabilities may not have been well defined.

- $\sigma_y(U)$ is increasing in U_y.
- $\sigma_y(U)$ is weakly decreasing in $U_{y'}$ for $y' \neq y$.
- If one replaces (U_y) by $(U_y + c)$, for a constant c, one has $\sigma(U + c) = \sigma(U)$.

- Because of the last property, we can normalize the utility of one of the alternatives. We will normalize the utility of the utility associated to $y = 0$, and hence take

$$U_0 = 0.$$

- Thus in the sequel, σ will be seen as a mapping from $\mathbb{R}^{\mathcal{Y}}$ to the set of $(s_y)_{y \in \mathcal{Y}}$ such that $s_y > 0$ and $\sum_{y \in \mathcal{Y}} s_y < 1$, and the choice probability of alternative $y = 0$ is recovered by

$$s_0 = 1 - \sum_{y \in \mathcal{Y}} s_y.$$

- In many settings, the econometrician observes the market shares s_y and wants to deduce the corresponding vector of systematic utilities. That is, we would like to solve:
  Problem. *Given a vector s with positive entries satisfying* $\sum_{y\in\mathcal{Y}} s_y < 1$, *characterize and compute the set*

$$\sigma^{-1}(s) = \left\{ U \in \mathbb{R}^{\mathcal{Y}} : \sigma(U) = s \right\}.$$

- This problem is called "demand inversion," or "conditional choice probability inversion," or "identification problem." It is a central issue in econometrics/industrial organization and will be a key building block for matching models.

- Define the expected indirect utility of consumers by

$$G(U) = \mathbb{E}\left[\max_{y \in \mathcal{Y}}(U_y + \varepsilon_y, \varepsilon_0)\right]$$

In the discrete choice literature, this is called *McFadden's surplus function*.

- As the expectation of the maximum of terms which are linear in U, G is convex function in U (strictly convex in fact), and

$$\frac{\partial G}{\partial U_y}(U) = \Pr(U_y + \varepsilon_y \geq U_z + \varepsilon_z \text{ for all } z \in \mathcal{Y}_0).$$

But the right-hand side is simply the probability s_y of chosing option y; therefore, we get:
Theorem (Daly-Zachary-Williams). *The map σ coincides with the gradient of G, that is*

$$\sigma\left(U\right) = \nabla G(U). \tag{1}$$

- Assume that $\mathbf{P}$ is the distributions of i.i.d. standard type I extreme value random variables, a.k.a. standard Gumbel distributions, which has c.d.f.

$$F(z) = \exp(-\exp(-x+\gamma))$$

  where $\gamma = 0.5772...$ (Euler's constant). The mean of this distribution is zero.
- Basic fact from extreme value theory: if $\varepsilon_1,...,\varepsilon_n$ are i.i.d. Gumbel distributions, then $\max\{u_i + \varepsilon_i\}$ has the same distribution as $\log\left(\sum_{i=1}^n \exp u_i\right) + \epsilon$, where ϵ is also a Gumbel.
- Notes:
  - This distribution is sometimes called the "Gumbel max" distribution, to contrast it with the distribution of its opposite, which is then called "Gumbel min".
  - The literature usually calls "standard Gumbel" the distribution with c.d.f. $\exp(-\exp(-x))$; but that distribution has mean γ, which is why we slightly depart from the convention.

- Then

$$G(U) = \log\left(1 + \sum_{y\in\mathcal{Y}} \exp(U_y)\right)$$

where $s_0 = 1 - \sum_{y\in\mathcal{Y}} s_y$. This is called a *log-partition function*.

- As a result, the choice probability of alternative y is proportional to the exponential of the systematic utility associated with U, that is

$$\sigma_y(U) = \frac{\exp U_y}{1 + \sum_{y'\in\mathcal{Y}} \exp(U_{y'})}$$

which is called a *Gibbs distribution*.

- Assume that the random utility shock is scaled by a factor T. Then

$$\sigma_y(U) = \frac{\exp(U_y/T)}{1 + \sum_{y'\in\mathcal{Y}} \exp(U_{y'}/T)}$$

which is sometimes called the *soft-max operator*, and converges as $T \to 0$ toward

$$\max_{y\in\mathcal{Y}} \{U_y, 0\}.$$

- We can invert gradient of convex functions by the convex conjugate: if G is strictly convex and C^1, then

$$\sigma^{-1}(s) = \nabla G^{-1}(s) = \nabla G^*(s).$$

- G^* is the Legendre-Fenchel transform of G; G and Salanié call it the *entropy of choice*, defined by

$$G^*(s) = \max_U \left\{ \sum_{y \in \mathcal{Y}} s_y U_y - G(U) \right\}. \tag{2}$$

- Hence, $\sigma^{-1}(s)$ is the vector U such that

$$U \in \arg\max_U \left\{ \sum_{y \in \mathcal{Y}} s_y U_y - G(U) \right\}.$$

- Convex duality implies that if s and U are related by $s \in \partial G(U)$, then

$$G(U) = \sum_{y \in \mathcal{Y}} s_y U_y - G_x^*(s). \tag{3}$$

- But letting $Y = \arg\max_y \{U_y + \varepsilon_y\}$, $G(U) = \mathbb{E}[U_Y + \varepsilon_Y]$ implies

$$G(U) = \sum_{y \in \mathcal{Y}} s_y U_y + \mathbb{E}[\varepsilon_Y],$$

thus one has

$$G^*(s) = -\mathbb{E}[\varepsilon_Y]. \tag{4}$$

Hence, the entropy of choice $G^*(s)$ is interpreted as minus the expected amount of heterogeneity needed to rationalize the choice probabilities s.

THEOREM (G AND SALANIÉ)

Consider a solution $(u(\varepsilon), v_y)$ *to the dual Monge-Kantorovich problem with cost* $\Phi(\varepsilon, y) = \varepsilon_y$*, that is:*

$$\min_{u,v} \int u(\varepsilon)\, d\mathbf{P}(\varepsilon) + \sum_{y \in \mathcal{Y}_0} v_y s_y \tag{5}$$

$$s.t.\ u(\varepsilon) + v_y \geq \Phi(\varepsilon, y)$$

Then:
(i) $U = \sigma^{-1}(s)$ *is given by* $U_y = v_0 - v_y$.
(ii) The value of Problem (5) is $-G^*(s)$.

The approach is generalized to the case of continuous choice in Chernozhukov, G, Henry and Pass (2017).

PROOF.

$\sigma^{-1}(s) = \arg\max_{U:U_0=0} \left\{ \sum_{y\in\mathcal{Y}} s_y U_y - G(U) \right\}$, thus, letting $v = -U$, v is the solution ot

$$\min_{v:v_0=0} \left\{ \sum_{y\in\mathcal{Y}_0} s_y v_y + G(-v) \right\}$$

which is exactly Problem (5). □

- In a number of cases, one cannot compute the choice probabilities $\sigma(U)$ using a closed-form expression. In this case, one simulates N points $\varepsilon^i \sim P$ and compute the *accept-reject simulator*

$$\sigma_y^N(U) = N^{-1} \sum_{i=1}^{N} 1\left\{ U_y + \varepsilon_y^i \geq U_z + \varepsilon_z^i \ \forall z \in \mathcal{Y} \right\}$$

- McFadden's smoothed accept-reject simulator (SARS) consists in sampling $\varepsilon \sim P$: $\varepsilon^1, ..., \varepsilon^N$, and replacing the max by the smooth-max

$$\sigma_y^{N,T}(U) = \sum_{i=1}^{N} \frac{1}{N} \frac{\exp\left(\left(U_y + \varepsilon_y^i \right) / T \right)}{\sum_z \exp\left(\left(U_z + \varepsilon_z^i \right) / T \right)}$$

- One seeks U so that the induced choice probabilities are s, that is

$$s_y = \sum_{i=1}^{N} \frac{1}{N} \frac{\exp\left(\left(U_y + \varepsilon_y^i \right) / T \right)}{\sum_z \exp\left(\left(U_z + \varepsilon_z^i \right) / T \right)}.$$

- Let $u_i = T \log \sum_z \exp\left((U_z + \varepsilon_z^i)/T\right)$. One has

$$\begin{cases} s_y = \sum_{i=1}^N \frac{1}{N} \exp\left((U_y - u_i + \varepsilon_y^i)/T\right) \\ \frac{1}{N} = \sum_y \frac{1}{N} \exp\left((U_y - u_i + \varepsilon_y^i)/T\right) \end{cases}.$$

- As a result, (u_i, U_y) are the solution of the regularized OT problem

$$\min_{u,U} \sum_{i=1}^N \frac{1}{N} u_i - \sum s_y U_y + \sum_{i,y} \frac{1}{N} \exp\left((U_y - u_i + \varepsilon_y^i)/T\right).$$

- Consider the IPFP algorithm for solving the latter problem:

$$\begin{cases} \exp\left(u_i^{k+1}/T\right) = \sum_z \exp\left(\left(U_z^k + \varepsilon_z^i\right)/T\right) \\ \exp U_y^{k+1}/T = \frac{N s_y}{\sum_{i=1}^N \exp\left(\left(-u_i^{k+1} + \varepsilon_y^i\right)/T\right)} \end{cases}$$

- This rewrites as

$$\exp U_y^{k+1}/T = \frac{N s_y}{\sum_{i=1}^N \frac{\exp\left(\varepsilon_y^i/T\right)}{\sum_z \exp\left(\left(U_z^k + \varepsilon_z^i\right)/T\right)}}, \text{ i.e.}$$

$$U_y^{k+1} = T \log s_y - T \log \sum_{i=1}^N \frac{1}{N} \frac{\exp\left(\varepsilon_y^i/T\right)}{\sum_z \exp\left(\left(U_z^k + \varepsilon_z^i\right)/T\right)}$$

which is exactly the contraction mapping algorithm of Berry, Levinsohn and Pakes (1995, appendix 1).

Section 2

Nonadditive discrete choice

- Consider now a (nonadditive) discrete choice problem where an agent ("consumer") draws a utility shock $\varepsilon \sim P$, and faces a choice between alternatives ("yoghurts") $j \in \mathcal{J}_0 = \mathcal{J} \cup \{0\}$, where $j = 0$ is the default option.
- Alternative $j \in \mathcal{J}_0$ brings the agent utility $\mathcal{U}_{\varepsilon j}(\delta_j)$, where δ_j is a systematic utility-shifter, and $\mathcal{U}_{\varepsilon j}(.)$ is continuous and increasing.
- The utility associated with alternative 0 is normalized to 0: $\delta_0 = 0$.
- The agent's problem is

$$\max_{j \in \mathcal{J}_0} \mathcal{U}_{\varepsilon j}(\delta_j), \tag{6}$$

and the agent chooses random variable $\tilde{j}(\varepsilon) \in \arg\max_{j \in \mathcal{J}_0} \mathcal{U}_{\varepsilon j}(\delta_j)$.

- In additive random utility models (ARUMs), $\mathcal{U}_{\varepsilon j}\left(\delta_j\right)$ is quasilinear in δ_j.
- If $\varepsilon \in \mathbb{R}^{\mathcal{J}}$, and δ_j=systematic utility, ARUMs have

$$\mathcal{U}_{\varepsilon j}\left(\delta_j\right) = \delta_j + \varepsilon_j$$

- Specifications include:
  - logit model: $\left(\varepsilon_j\right)$ are i.i.d. Gumbel
  - pure characteristics: $\varepsilon_j = \epsilon^{\intercal}\xi_j$ with ϵ a consumer-specific random taste vector of $\mathbb{R}^d$ and $\xi_j \in \mathbb{R}^d$ is the vector of characteristics of alternative j.

- In nonadditive random utility models (NARUMs), $\mathcal{U}_{\varepsilon j}\left(\delta_j\right)$ is increasing and continuous, but not quasilinear in δ_j.
- Investments with taxes: $\varepsilon \in \{0,1\} \times \mathbb{R}^{\mathcal{J}}$; $\varepsilon^1 = 1$ if tax-exempt individual, $\varepsilon^1 = 0$ if tax-liable; $\delta_j + \varepsilon_j^2 =$ project j's pre-tax earnings, tax rate $\tau \in (0,1)$, so

$$\mathcal{U}_{\varepsilon j}\left(\delta_j\right) = \varepsilon^1 \left(\delta_j + \varepsilon_j^2\right) + \left(1 - \varepsilon^1\right)(1-\tau)\left(\delta_j + \varepsilon_j^2\right).$$

- Product choice with waiting lines: $\delta_j = T - T_j$ is the time spared in line waiting for product j; price is p_j; $\varepsilon \in \mathbb{R} \times \mathbb{R}^{\mathcal{J}}$, where ε^1 is the valuation of time and ε_j^2 some value shock, so

$$\mathcal{U}_{\varepsilon j}\left(\delta_j\right) = \varepsilon^1 \delta_j - p_j + \varepsilon_j^2.$$

- Other examples include: risk aversion; cap on compensation; etc.

- Assume for now no indifference:

$$\Pr\left(\mathcal{U}_{\varepsilon j}\left(\delta_j\right) = \mathcal{U}_{\varepsilon j'}\left(\delta_{j'}\right)\right) = 0 \text{ for every } \delta \text{ and } j \neq j'$$

then the *demand map* $\tilde{\sigma} : \mathbb{R}^{\mathcal{J}} \to \mathbb{R}^{\mathcal{J}}$ associates the market share (=vector of choice probabilities) to vector of systematic utilities δ, i.e.

$$s = \tilde{\sigma}\left(\delta\right) \iff s_j = \Pr\left(\mathcal{U}_{\varepsilon j}\left(\delta_j\right) \geq \max_{j' \in \mathcal{J}_0} \mathcal{U}_{\varepsilon j'}\left(\delta_{j'}\right)\right)$$

- Our focus is *inverse demand*, namely

$$\tilde{\sigma}^{-1}\left(s\right) = \left\{\delta \in \mathbb{R}^{\mathcal{J}} : s = \tilde{\sigma}\left(\delta\right)\right\}$$

a.k.a. "CCP inversion".
- In the logit case (additive random utility with Gumbel distribution), done in closed form, but in general, need to resort to various numerical methods.

- Demand inversion: literature started by Hotz and Miller (1993), Berry (1994). See among others Aguirregabiria and Mira (2002), Arcidiacono and Miller (2011), Kristensen, Nesheim, and de Paula (2014)...
- Fixed-point approach: Berry (1994) and Berry, Levinsohn, and Pakes (1995).
- MCMC approach: Rossi, Allenby and McCulloch (2005).
- MPEC approach: Dube, Fox and Su (2012).
- Optimal transport/linear programming approach (additive models): G and Salanié (2014), Chiong, G and Shum (2016), Chernozhukov, G, Henry and Pass (2014).

The Equivalence theorem (Bonnet, G and Shum, 2017)

- Recall:
  - Demand computation (σ): **fix** $\varepsilon \sim P$ and $\delta \in \mathbb{R}^{\mathcal{J}}$, **compute** $\tilde{j} \sim s$.
  - Demand inversion (σ^{-1}): **fix** $\varepsilon \sim P$ and $\tilde{j} \sim s$, **compute** $\delta \in \mathbb{R}^{\mathcal{J}}$.
- Our result shows equivalence between the latter and:
  - Matching equilibrium: **fix** $\varepsilon \sim P$ and $\tilde{j} \sim s$, **compute** $\delta \in \mathbb{R}^{\mathcal{J}}$.
  - Here, $\varepsilon \sim P$=distribution of workers' types; $j \sim s$=distribution of firms' types; δ=wage offered by firm j.
- 2 sets of implications:
  - Computational: a new class of algorithms (coming from matching theory) for performing demand inversion.
  - Theoretical: (1) lattice structure; (2) inverse isotonicity; (3) desirable properties of the identified set (connectedness, point-identication, stability)

- Consider a market where workers characteristics is $\varepsilon \in \mathcal{X}$ (continuous) and firms' types are $j \in \mathcal{J}_0$ (discrete). Assume $\varepsilon \sim P$, and $j \sim s$, i.e. workers are distributed as P and the mass of firm's type j is s_j. One has

$$\int dP(\varepsilon) = 1 \text{ and } \sum_{j \in \mathcal{J}_0} s_j = 1,$$

so that there is the same mass (normalized to one) of workers and firms.

- Assume that if worker ε works with firm j at wage u, then the firm's profit is $\Phi_{\varepsilon j}(u)$, decreasing and continuous in u.
- An outcome is the specification of $(\pi(\varepsilon, j), u(\varepsilon), v_j)$, where
  - $\pi(\varepsilon, j)$ is a matching: it is the probability of drawing a matched pair of worker ε and firm j. It should have margins P and s, i.e.

$$\pi \in M(P, s) \Leftrightarrow_{def} \begin{cases} \int \pi(\varepsilon, j)\, d\varepsilon = s_j \\ \sum_{j \in \mathcal{J}_0} \pi(\varepsilon, j) = P(\varepsilon) \end{cases}.$$

  - $(u(\varepsilon), v_j)$ are payoffs: $u(\varepsilon)$ is the wage of worker ε, and v_j is the profit of a firm of type j.

- Cf. Roth-Sotomayor, ch. 9. Outcome (π, u, v) if *pairwise stable* when:
  - *no blocking pair*: for all ε and j,

$$v_j \geq \Phi_{\varepsilon j}(u(\varepsilon))$$

    otherwise the pair would be "blocking" i.e. it would be possible for firm j to hire ε at her current wage and get more than v_j.
  - *feasibility*: if ε and j are matched, i.e. if ,

$$(\varepsilon, j) \in Supp(\pi) \implies v_j = \Phi_{\varepsilon j}(u(\varepsilon))$$

    otherwise the payoffs $u(\varepsilon)$ and δ_j would not be sustainable for matched pair (ε, j).
- Walrasian interpretation of pairwise stability: if $(u(\varepsilon), v_j)$ are stable payoffs, then no blocking pair+feasibility imply that

$$v_j = \max_{\varepsilon \in \mathcal{X}} \Phi_{\varepsilon j}(u(\varepsilon)) \text{ and } u(\varepsilon) = \max_{j \in \mathcal{J}_0} \Phi_{\varepsilon j}^{-1}(v_j). \quad (7)$$

  - This is expresses that the problem of workers and the problem of firms are dual of one another.
  - Equivalence result stems from identifying $\Phi_{\varepsilon j}^{-1}(v_j)$ with $\mathcal{U}_{\varepsilon j}(\delta_j)$.

- The following result holds:

THEOREM (BONNET, G AND SHUM)

One has

$$\delta \in \tilde{\sigma}^{-1}(s)$$

if and only if (u, v) *is a stable outcome in the matching problem, where*

$$\delta_j = -v_j \text{ and } \mathcal{U}_{\varepsilon j}(\delta_j) = \Phi_{\varepsilon j}^{-1}(-\delta_j)$$

and

$$u(\varepsilon) = \max_{j \in \mathcal{J}_0} \{\mathcal{U}_{\varepsilon j}(\delta_j)\}.$$

- Note that Φ is obtained from $\mathcal{U}$ by

$$\Phi_{\varepsilon j}(u) = -\mathcal{U}_{\varepsilon j}^{-1}(u).$$

- Consider the pure characteristics model:

$$\mathcal{U}_{\varepsilon j}\left(\delta_j\right) = \delta_j + \epsilon^{\intercal}\xi_j$$

(w.l.o.g. the argument extends to any ARUM). Then $\mathcal{U}_{\varepsilon j}^{-1}\left(u\right) = u - \epsilon^{\intercal}\xi_j$, so

$$\Phi_{\epsilon\xi_j}\left(u\right) = -u + \epsilon^{\intercal}\xi_j.$$

- $\left(\pi, u, v\right)$ is a stable outcome in the matching problem if for all ε and j,

$$u\left(\epsilon\right) + \delta_j \geq \epsilon^{\intercal}\xi_j$$

with equality if $\left(\epsilon, j\right) \in Supp\left(\pi\right)$.

- Outside of ARUMs, BLP's contraction mapping is no longer necessarily a contraction mapping, and fails to converge in a number of instances.
- McFadden's smooth accept-reject simulator (SARS) consists in sampling $\varepsilon \sim P$: $\varepsilon^1, ..., \varepsilon^N$, and replacing the max

$$u_i = \max_{j \in \mathcal{J}_0} \mathcal{U}_{\varepsilon_i j}\left(\delta_j\right)$$

by the smooth-max

$$u_i = T \log \sum_{j \in \mathcal{J}_0} \exp\left(\mathcal{U}_{\varepsilon_i j}\left(\delta_j\right) / T\right)$$

- One seeks δ so that the induced choice probabilities are s, that is

$$s_j = \sum_{i=1}^{N} \frac{1}{N} \frac{\exp\left(\mathcal{U}_{\varepsilon_i j}\left(\delta_j\right) / T\right)}{\exp\left(u_i / T\right)}.$$

- Letting

$$\pi_{ij} = \frac{1}{N} \exp\left(\frac{\mathcal{U}_{\varepsilon_i j}\left(\delta_j\right) - u_i}{T} \right),$$

  one has

$$\begin{cases} \sum_{j \in \mathcal{J}_0} \pi_{ij} = \frac{1}{N} \\ \sum_{1 \leq i \leq N} \pi_{ij} = s_j \end{cases}$$

- The generalized IPFP algorithm proposed in G, Kominers and Weber (2016) allows for fast and scalable computation of this problem by iteratively fitting these equations. Namely, given any starting point δ^0, solve

$$\begin{cases} u_i^{k+1} = T \log \sum_{j \in \mathcal{J}_0} \exp\left(\mathcal{U}_{\varepsilon_i j}\left(\delta_j^0\right) / T \right) \\ \delta_j^{k+1} \text{ such that } \frac{1}{N} \sum_{1 \leq i \leq N} \exp\left(\frac{\mathcal{U}_{\varepsilon_i j}\left(\delta_j^{k+1}\right) - u_i}{T} \right) = s_j \end{cases}$$

- Can be implemented in parallel. Converges in cases where BLP's contraction mapping algorithm fails.