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

Alfred Galichon (New York University)

Spring 2018
Day 4, January 18 2018: "Multinomial choice"
Block 10. Basics of static discrete choice

Learning objectives: block 10

- Emax operator and generalized entropy of choice
- The Daly-Zachary-Williams theorem
- The GEV class
- Parametric estimation of multinomial choice models

References for block 10

- [OTME], App. E
- 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.
- G and Salanié (2017). "Cupid's invisible hands". Preprint.
- Chiong, G and Shum, "Duality in Discrete Choice Models". *Quantitative Economics*, 2016.
- Greene and Hensher (1997) Multinomial logit and discrete choice models.

Section 1

Theory

- 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

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

Demand map: definition

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

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

  The map σ is called *demand map*, and the vector s is called vector of market shares, or vector of choice probabilities.
- 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.

Demand map: properties

- $\sigma_y(U)$ 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 $\sigma(U + c) = \sigma(U)$.

Normalization

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

- Define the expected indirect utility of consumers by

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

  This is called *Emax operator*, a.k.a. *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 demand map σ is the gradient of the Emax operator G, that is*

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

Example 1: logit

- Assume that **P** is the distribution of i.i.d. *centered type I extreme value* a.k.a. *centered Gumbel* terms, 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 ε_1,...,ε_n are i.i.d. Gumbel distributions, then $\max\{U_y + \varepsilon_y\}$ has the same distribution as $\log\left(\sum_{y=1}^{n} \exp U_y\right) + \epsilon$, where ϵ is also a Gumbel. (Proof of this fact later).

- 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.

Example 1: logit, Emax function and demand map

- The Emax operator associated with the logit model can be given in closed form as

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

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

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

Let $\mathbf{F}$ be a cumulative distribution such that function g defined by

$$g(x_1, ..., x_n) = -\log \mathbf{F}(-\log x_1, ..., -\log x_n) \tag{2}$$

is positive homogeneous of degree 1. (This inverts into $\mathbf{F}(u_1, ..., u_n) = \exp(-g(e^{-u_1}, ..., e^{-u_n}))$). We have by a theorem of McFadden (1978):

THEOREM

Let $(\varepsilon_y)_{1 \leq y \leq n}$ be a random vector with c.d.f. $\mathbf{F}$, and define

$$Z = \max_{y=1,...,n} \{U_y + \varepsilon_y\}.$$

Then Z has the same distribution as $\log g\left(e^{U_1}, ..., e^{U_n}\right) + \gamma + \epsilon$, where ϵ is a standard Gumbel. In particular,

$$\mathbb{E}\left[\max_{y=1,...,n} \{U_y + \varepsilon_y\}\right] = \log g\left(e^{U_1}, ..., e^{U_n}\right) + \gamma$$

where γ is the Euler constant $\gamma \simeq 0.5772$.

PROOF.

Let F_Z be the c.d.f. of $Z = \max_{y=1,...,n} \{U_y + \varepsilon_y\}$. One has

$$\begin{aligned} F_Z(z) &= \Pr\left(\max_{y=1,...,n} \{U_y + \varepsilon_y\} \leq z\right) = \Pr(\forall y : \ \varepsilon_y \leq z - U_y) \\ &= \mathbf{F}(z - U_1, ..., z - U_n) = \exp\left(-g\left(e^{U_1 - z}, ..., e^{U_n - z}\right)\right) \\ &= \exp\left(-e^{-z} g\left(e^{U_1}, ..., e^{U_n}\right)\right) = \varphi\left(z - \log g\left(e^{U_1}, ..., e^{U_n}\right) - \gamma\right) \end{aligned}$$

where $\varphi(z) := \exp\left(-e^{-(z-\gamma)}\right)$ is the cdf of the standard Gumbel distribution. Hence Z has the distribution of $\log g\left(e^{U_1}, ..., e^{U_n}\right) + \gamma + \epsilon$, where ϵ is a standard Gumbel. □

- As a result, the choice probability of alternative y is

$$\sigma_y(U) = \frac{\frac{\partial g}{\partial x_y}\left(e^{U_1}, ..., e^{U_n}\right)}{g\left(e^{U_1}, ..., e^{U_n}\right)} e^{U_y}.$$

- The GEV framework has several commonly used examples: logit, nested logit, mixture of logit...
- We just saw the logit model, in which $g(x_1, ..., x_n) = e^{-\gamma} \sum_{y=1}^{n} x_y$. In this case, the distribution of

$$Z = \max_{y=1,...,n} \{U_y + \varepsilon_y\}$$

is $\log \sum_{y=1}^{n} e^{U_y} + \epsilon$, where ϵ is a standard Gumbel.

Example 3: nested logit model

- The nested logit model is an instance of GEV model where alternatives can be grouped in nests. Eg, people choose their means of transportation (nest), and within this nest, a particular operator.
- Let $\mathcal{X}$ be the set of nests and assume that for each nest x, there is a set $\mathcal{Y}_x$ alternatives. Let U_{xy} be utility from alternative y in nest x, and $\lambda_x \in [0,1]$ and

$$g(U_{xy}) = e^{-\gamma} \sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}_x} U_{xy}^{1/\lambda_x} \right)^{\lambda_x}.$$

- In this case,

$$G(U) = \mathbb{E}\left[\max_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}_x} \{U_{xy} + \varepsilon_{xy}\} \right] = \log \sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}_x} e^{U_{xy}/\lambda_x} \right)^{\lambda_x}$$

$$\sigma_{xy}(U) = \frac{\left(\sum_{y \in \mathcal{Y}_x} e^{U_{xy}/\lambda_x} \right)^{\lambda_x}}{\sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}_x} e^{U_{xy}/\lambda_x} \right)^{\lambda_x}} \frac{e^{U_{xy}/\lambda_x}}{\left(\sum_{y \in \mathcal{Y}_x} e^{U_{xy}/\lambda_x} \right)}$$

so the demand map has an interesting interpretation as "choice of nest then choice of alternative".

- 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.

- We saw in Lecture 3 how to invert gradient of convex functions: 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; we call it the *entropy of choice*, defined by

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

- 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

$$G(U) = \sum_{y \in \mathcal{Y}} s_y U_y - G_x^*(s). \tag{4}$$

- 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

$$G^*(s) = -\mathbb{E}[\varepsilon_Y]. \tag{5}$$

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

- Then
$$G^*(s) = s_0 \log(s_0) + \sum_{y \in \mathcal{Y}} s_y \log s_y$$
where $s_0 = 1 - \sum_{y \in \mathcal{Y}} s_y$. Hence, G^* is a bona fide entropy function when $\mathbf{P}$ is Gumbel–hence the name of *entropy of choice* in general.
- As a result,
$$\sigma_y^{-1}(s) = \log \frac{s_y}{s_0}$$
which is the celebrated *log-odds ratio formula*: the log of the odds of alternatives y and 0 identify the difference between the systematic utilities of these alternatives.

Example: entropy of choice and identification, nested logit model

- The entropy of choice G^* in the nested logit model is given by

$$G^*(s) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}_x} \lambda_x s_{xy} \ln s_{xy} + \sum_{x \in \mathcal{X}} (1 - \lambda_x) \left(\sum_{z \in \mathcal{Y}_x} s_{xz} \right) \ln \left(\sum_{z \in \mathcal{Y}_x} s_{xz} \right) \tag{6}$$

  if $s_{xy} \geq 0$ and $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}_x} s_{xy} = 1$, $G^*(s) = +\infty$ otherwise.
- Identification in the nested logit model: with normalization $\sum_{x \in \mathcal{X}} \left(\sum_{y \in \mathcal{Y}_x} e^{U_{xy}/\lambda_x} \right)^{\lambda_x} = 1$, one has $s_{xy} = \left(\sum_{y \in \mathcal{Y}_x} e^{U_{xy}/\lambda_x} \right)^{\lambda_x - 1} e^{U_{xy}/\lambda_x}$, thus $\sum_{y \in \mathcal{Y}_x} e^{U_{xy}/\lambda_x} = \left(\sum_{y \in \mathcal{Y}_x} s_{xy} \right)^{1/\lambda_x}$, therefore

$$U_{xy} = \lambda_x \log s_{xy} - (\lambda_x - 1) \log \sum_{y \in \mathcal{Y}_x} s_{xy}.$$

- Assume the utilities are parameterized as follows: $U = X\beta$ where $\beta \in \mathbb{R}^p$ is a parameter, and X is a $|\mathcal{Y}| \times p$ matrix.
- The log-likelihood function is given by

$$I(\beta) = N \sum_y \hat{s}_y \log \sigma_y (X\beta)$$

- A common estimation method of β is by maximum likelihood

$$\max_\beta I(\beta).$$

MLE is statistically efficient; the problem is that the problem is not guaranteed to be convex, so there may be computational difficulties (e.g. local optima).

MLE, LOGIT CASE

- In the logit case,

$$I(\beta) = N\left\{\hat{s}^{\mathsf{T}} X\beta - \log \sum_y \exp(X\beta)_y\right\}$$

so that the max-likehood amounts to

$$\max_\beta \left\{\hat{s}^{\mathsf{T}} X\beta - G(X\beta)_y\right\}$$

whose value is the Legendre-Fenchel transform of $\beta \to G(X\beta)$ evaluated at $X^{\mathsf{T}}\hat{s}$.

- Note that the vector $X^{\mathsf{T}}\hat{s}$ is the vector of empirical moments, which is a sufficient statistics in the logit model.
- As a result, in the logit case, the MLE is a convex optimization problem, and it is therefore both statistically efficient and computationally efficient.

- The previous remark will inspire an alternative procedure based on the moments statistics $X^{\mathsf{T}}\hat{s}$.
- The social welfare is given in general by $W(\beta) = G(X\beta)$. One has $\partial_{\beta^i} W(\beta) = \sum_y \sigma_y(X\beta) X_{yi}$, that is

$$\nabla W(\beta) = X^{\mathsf{T}} \sigma(X\beta),$$

  which is the vector of predicted moments.
- Therefore the program

$$\max_{\beta} \left\{ \hat{s}^{\mathsf{T}} X\beta - G(X\beta)_y \right\}$$

  picks up the parameter β which matches the empirical moments $X^{\mathsf{T}}\hat{s}$ with the predicted ones $\nabla W(\beta)$. This procedure is not statistically efficient, but is computationally efficient becauses it arises from a convex optimization problem.

Section 2

CODING

Today

- The data ('10-appli-travelmode') is taken from Greene and Hensher (1997). 210 individuals are surveyed about their choice of travel mode between Sydney, Canberra and Melbourne, and the various costs (time and money) associated with each alternative. Therefore there are 840 = 4 x 210 observations, which we can stack into 'travelmodedataset' a 3 dimensional array whose dimensions are mode,individual,dummy for choice+covariates.
- First, we compute the unconditional market shares:
  `s = apply(X = travelmodedataset[,,3],FUN = mean, MARGIN = 1)`
  which yields:

```
 air       train      bus        car
 0.2761905 0.3000000  0.1428571  0.2809524
```

- Define "car" as the default alternative. The utilities in the logit model are obtained by the log-odds ratio formula:
  `Ulogit = log(s[1:4]/s[4])`
  which yields

```
  air           train        bus           car
-0.01709443   0.06559728   -0.67634006   0.00000000
```

- Now compute these utilities using a nested logit model with two nests, "noncar" and "car", and taking $\lambda = 0.5$ in both nests. Do:

```
Unocar = lambda[1]*log(s[1:3])+(1-lambda[1]) *
log(sum(s[1:3]))
Ucar = lambda[2]*log(s[4])+(1-lambda[2]) * log(sum(s[4]))
Unested = c(Unocar,Ucar ) - Ucar
```

  which yields

```
  air          train       bus          car
0.4613240   0.5026698   0.1317012   0.00000000
```

- We see how correlation whithin nests impacts the estimation of the systematic utilities. Why?