Identification of Discrete Choice Models using Optimal Transport

Thomas Wiemann
University of Chicago

TA Discussion # 6
Econ 31740

February 21, 2022

Outline

Today's discussion on identification of discrete choice models using optimal transport reviews Chiong et al. (2016) as well as Chapter 9.2 of Galichon (2016) and selected results from Galichon and Salanié (2020).

We will consider the conventional discrete choice model where heterogeneous agents choose an alternative among a finite set of options that maximizes their (unobserved) utility. The key question of interest in this setting is whether we can characterize systematic differences between utilities derived from different choices. We will see that point-identification of these systematic differences coincides with uniqueness of the solution to an optimal transport problem.

The optimal transport approach to discrete choice models is also very interesting from a computational perspective as a discretized version of the problem can be expressed as a finite dimensional linear program.

Discrete Choice Model

Consider the classical discrete choice problem where heterogeneous agents choose an alternative $y \in \mathcal{Y} := \{1, \ldots, J\}$ according to

$$y = \underset{y \in \mathcal{Y}}{\arg\max}\ w_y + \varepsilon_y, \tag{1}$$

where w_y is the systematic utility of choice y shared across all agents, and $\varepsilon_y \sim Q$ is the corresponding latent utility shock.

Define

$$Y(w, \varepsilon) := \underset{y \in \mathcal{Y}}{\arg\max}\ w_y + \varepsilon_y, \tag{2}$$

where $w := (w_j)_{j=1}^J$ and $\varepsilon := (\varepsilon_j)_{j=1}^J$, and let

$$p_y := E_Q\left[Y(w, \varepsilon) = y\right] = P_Q(w_y + \varepsilon_y > w_j + \varepsilon_j, \forall j \neq y). \tag{3}$$

Discrete Choice Model (Contd.)

Suppose that the choice probabilities $(p_j)_{j=1}^{J}$ are observed, and that the distribution Q is known to the researcher. We may then pose the question of identification in the discrete choice model as follows:

Given the choice probabilities $(p_j)_{j=1}^{J}$ *and the distribution of utility shocks* Q*, what are the systematic utilities* $(w_j)_{j=1}^{J}$ *that are compatible with the discrete choice problem* (1)*?*

Here, "compatible" means the systematic utilities $(w_j)_{j=1}^{J}$ are such that $p_j = E_Q\left[Y(w,\varepsilon) = j\right], \forall j \in \mathcal{Y}$, where the LHS and Q are known.

Chiong et al. (2016) consider this question in the setting of a dynamic discrete choice problem. Chapter 9.2 of Galichon (2016) considers the simplified setting discussed here.

Convex Analysis for Discrete Choice Models

To make progress, we need a couple of definitions and results from convex analysis. Appendix A of Chiong et al. (2016) gives an overview of results relevant for discrete choice models.

Define the ex-ante expected utility as

$$W(w) = E_Q\left[\max_{y\in\mathcal{Y}} w_y + \varepsilon_y\right]. \tag{4}$$

Assume that $E_Q[|\varepsilon_y|] < \infty$, then W is a convex function.

Define the convex conjugate of W as

$$W^*(p) = \sup_{w\in\mathbb{R}^J} p^\top w - W(w). \tag{5}$$

As W^* is the supremum of affine functions, it is convex. The domain of W^* is the set $\{p \in \mathbb{R}^J | \, W^*(p) < \infty\}$. From Norets and Takahashi (2013), it follows that $W^*(p) < \infty, \forall p \in \{p \in \mathbb{R}^J_+ | \sum_j p_j = 1\}$.

Convex Analysis for Discrete Choice Models (Contd.)

The *subdifferential* of W at w is defined as

$$\partial W(w) := \left\{ p \in \mathbb{R}^J \middle| \, W(w') \geq W(w) + p^\top (w' - w), \, \forall w' \in \mathbb{R}^J \right\}. \quad (6)$$

Similarly, we have

$$\partial W^*(w) := \left\{ w \in \mathbb{R}^J \middle| \, W^*(p') \geq W^*(p) + w^\top (p' - p), \, \forall p' \in \mathbb{R}^J \right\}. \quad (7)$$

Notice that from the definitions (4) and (5), it holds generally that

$$W(w) + W^*(p) \geq p^\top w. \quad (8)$$

Equality holds if and only if $p \in \partial W(w)$. This is known as *Fenchel's equality*.

Identification of Discrete Choice Models

Fenchel's equality is straightforwardly results in a useful equivalence statement:

Theorem 1 (Chiong et al., 2016): $p \in \partial W(w) \Leftrightarrow w \in \partial W^*(p)$.

Sketch of the proof: By Fenchel's equality, $p \in \partial W(w)$ is equivalent to

$$W(w) + W^*(p) = p^\top w. \tag{9}$$

By symmetry, this is equivalent to $w \in \partial W^*(p)$.

Identification of Discrete Choice Models (Contd.)

It follows from Theorem 1 that identification of $(w_j)_{j=1}^J$ given $(p_j)_{j=1}^J$ and Q is equivalent to showing that $\partial W^*(p)$ is a singleton.

Unsurprisingly, this is not the case in general. In particular, if $w \in \partial W^*(p)$, then $(w - k\mathbf{1}_J) \in W^*(p)$ for some $k \in \mathbb{R}$. This the familiar result that choice probabilities are only affected by the difference in levels offered by various alternatives. We thus consider the normalization

$$W(w^0) = 0. \tag{10}$$

Theorem 3 of Chiong et al. (2016), stated on the next slide, shows that this is indeed a normalization and not a model restriction. (In particular, Equation (10) defines a reference point from which all $w \in \partial W^*(p)$ can be represented.)

Identification of Discrete Choice Models (Contd.)

Theorem 2 (Chiong et al., 2016): Assume the distribution Q of the utility shocks ε is such that the distribution of the vector $(\varepsilon_y - \varepsilon_1)_{y \neq 1}$ has full support. Let $p \in \{p \in \mathbb{R}_+^J | \sum_j p_j = 1\}$. Then, for a given Q, there exists a unique $w^0 \in \partial W^*(p)$ such that $W(w^0) = 0$.

The full support assumption on the latent utility shocks is common in the literature. It states that all $y \in \mathcal{Y}$ have positive probability in all choice sets.

Theorem 3 (Chiong et al., 2016): Let $k \in \mathbb{R}$ and maintain the full support assumption. The set of conditions

$$w \in \partial W^*(p) \quad \text{and} \quad W(w) = k, \tag{11}$$

is equivalent to

$$w_y = w_y^0 + k, \, \forall y \in \mathcal{Y}. \tag{12}$$

Identification of Discrete Choice Models (Contd.)

Sketch of the proof (Theorem 2):

1. Choose $\tilde{w} \in \partial W^*(p)$ and let $w_y = \tilde{w}_y - W(\tilde{w})$. Note that $W(w) = E\left[\max_y \tilde{w}_y - W(\tilde{w}) + \varepsilon_y\right] = W(\tilde{w}) - W(\tilde{w}) = 0$, and $w \in \partial W^*(p)$. Then $p = \partial W(w)$ by Theorem 1.
2. To show uniqueness, suppose that $\exists w \neq w' : \; W(w) = W(w') = 0$ and $p \in \partial W(w)$ and $p \in \partial W(w')$. Then $\exists y_0 \neq y_1 : \; w_{y_0} - w_{y_1} \neq w'_{y_0} - w'_{y_1}$.
3. WLOG, consider $w_{y_0} - w_{y_1} > w'_{y_0} - w'_{y_1}$. Define

$$\Gamma := \left\{ \varepsilon \in \operatorname{supp} Q \;\middle|\; \begin{array}{c} w_{y_0} - w_{y_1} > \varepsilon_{y_1} - \varepsilon_{y_0} > w'_{y_0} - w'_{y_1} \\ w_{y_0} + \varepsilon_{y_0} > \max\limits_{y \neq y_0, y_1} w_y + \varepsilon_y \\ w'_{y_1} + \varepsilon_{y_1} > \max\limits_{y \neq y_0, y_1} w'_y + \varepsilon_y \end{array} \right\}. \tag{13}$$

Identification of Discrete Choice Models (Contd.)

Sketch of the proof (Theorem 2, contd.):

4. Note that $\forall \varepsilon \in \Gamma$, it holds that $Y(w, \varepsilon) = y_0$ and $Y(w', \varepsilon) = y_1$. Because of the full support assumption, $P_Q(\varepsilon \in S) > 0$.
5. Let $\bar{w} = \frac{w+w'}{2}$. Because $p \in \partial W^{(}w)$ and $p \in \partial W^{(}w')$, W is linear on $[w, w']$; combining with $W(w) = W(w') = 0$ we have $W(\bar{w}) = 0$.
6. Thus

$$\begin{aligned} 0 &= E\left[\bar{w}_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}\right] \\ &= \frac{1}{2}E\left[w_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}\right] + \frac{1}{2}E\left[w'_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}\right] \\ &\leq \frac{1}{2}E\left[w_{Y(w,\varepsilon)} + \varepsilon_{Y(w,\varepsilon)}\right] + \frac{1}{2}E\left[w'_{Y(w',\varepsilon)} + \varepsilon_{Y(w',\varepsilon)}\right] \\ &= \frac{1}{2}\left(W(w) + W(w')\right) = 0, \end{aligned} \tag{14}$$

where we used $w_{Y(w,\varepsilon)} + \varepsilon_{Y(w,\varepsilon)} \geq w_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}$.

Identification of Discrete Choice Models (Contd.)

Sketch of the proof (Theorem 2, contd.):

7. It follows from $w^l_{Y(w^l,\varepsilon)} + \varepsilon_{Y(w^l,\varepsilon)} \geq w^l_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}$, $\forall w^l \in \{w, w'\}$ and (14) that the equality holds term by term:

$$\begin{aligned} w_{Y(w,\varepsilon)} + \varepsilon_{Y(w,\varepsilon)} &\geq w_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}, \\ w'_{Y(w',\varepsilon)} + \varepsilon_{Y(w',\varepsilon)} &\geq w'_{Y(\bar{w},\varepsilon)} + \varepsilon_{Y(\bar{w},\varepsilon)}. \end{aligned} \tag{15}$$

8. Take $\varepsilon \in \Gamma$. Then $y_0 = Y(w,\varepsilon) = Y(\bar{w},\varepsilon) = Y(w',\varepsilon) = y_1$, which is the desired contradiction.
9. Hence $w = w'$ and uniqueness follows.

Sketch of the proof (Theorem 3): Recall $W(w^0) = 0$ by definition and note that $\partial W\left(w - W(w)\right) = \partial W(w)$. Then, by uniqueness of w^0 in Theorem 2, it follows that

$$w^0 = w - W(w). \tag{16}$$

Identification using Optimal Transport

Thus far, we have related identification of the systematic utilities w in the discrete choice model of (1) with the sub-gradient of the convex conjugate of the ex-ante expected utility $\partial W^*(p)$. This allowed for insights into the identification requirements (e.g., why the full support assumption is needed), but it may not yet be obvious how the results can be leveraged for practical applications / computation.

Proposition 2 of Galichon (2016) shows that the identification of the discrete choice model can be formulated as a mass transport problem (see also Galichon and Salanié, 2020). In particular, that the identification of the systematic utilities w is equivalent to uniqueness of the solution to an optimal transport problem. (The proposition was first stated in an earlier version of the working paper Galichon and Salanié, 2020).

Identification using Optimal Transport (Contd.)

Proposition 2 (Chiong et al., 2016): Under the full support assumption, we have

$$W^*(p) = \sup_{w,z:\, w_y + z(\varepsilon) \leq c(y,\varepsilon)} E_p[w_Y] + E_Q[z(\varepsilon)], \tag{17}$$

where $c(y,\varepsilon) = -\varepsilon_y$, $w \in \mathbb{R}^J$, and $z(\cdot)$ is a Q-measurable random variable. By Monge-Kantorovich duality, (17) coincides with its dual

$$W^*(p) = \min_{\pi:\, Y \sim p, \varepsilon \sim Q} E_\pi[c(Y,\varepsilon)]. \tag{18}$$

Further, $w \in \partial W^*(p)$ if and only if there exists z such that (w,z) solves (17). Finally, $w^0 \in \partial W^*(p)$ and $W(w^0) = 0$ if and only if there exists z such that (w^0, z) solves (17) and z is such that $E_Q[z(\varepsilon)] = 0$

Identification using Optimal Transport (Contd.)

Sketch of the proof (Proposition 2):

1. By definition, we have

$$\begin{aligned} W^*(p) &= \sup_{w \in \mathbb{R}^J} \sum_{y \in \mathcal{Y}} p_y w_y - E_Q \left[\max_{y \in \mathcal{Y}} \{ w_y + e_y \} \right] \\ &= \sup_{w \in \mathbb{R}^J} \sum_{y \in \mathcal{Y}} p_y w_y + E_Q \left[z(\varepsilon) \right], \end{aligned} \tag{19}$$

   where $z(\varepsilon) := -\min_{y \in \mathcal{Y}} \{-w_y - e_y\}$.

2. Defining $c(y, \varepsilon) := -\varepsilon_y$, we can introduce the constraint $w_y + z(\varepsilon) \leq c(y, \varepsilon)$, which is equivalent to $w_y + \varepsilon_y \leq \max_{y \in \mathcal{Y}} \{w_y + \varepsilon_y\}$. Then (19) is equivalent to

$$W^*(p) = \sup_{w_y + z(\varepsilon) \leq c(y,\varepsilon)} E_p[w_Y] + E_Q \left[z(\varepsilon) \right]. \tag{20}$$

Identification using Optimal Transport (Contd.)

Sketch of the proof (Proposition 2, contd.):

3. By Monge-Kantorovich, (20) is the dual to the optimal transport problem given by

$$W^*(p) = \min_{\pi:\, Y \sim p,\, \varepsilon \sim Q} E_\pi \left[c(Y, \varepsilon) \right]. \tag{21}$$

4. Comparing (19) with (20), we see that $w \in \partial W^*(p)$ if z such that (w, z) solve (20).

Computation using Optimal Transport

Let $\hat{Q}$ be a discrete approximation to Q constructed as a uniform distribution over S randomly sampled iid draws from Q. Note that as $S \to \infty, \hat{Q} \to Q$ by Glivenko-Cantelli. Using $\hat{Q}$, we can consider a discretized analogue to (17):

$$\begin{aligned} \min_{\lambda \in \mathbb{R}^J, z \in \mathbb{R}^S} \quad & \sum_{y \in \mathcal{Y}} p_y \lambda_y + \frac{1}{S} \sum_{s=1}^{S} z_s \\ \text{s.t.} \quad & \lambda_y + z_s \geq \varepsilon_y^s, \qquad \forall y \in \mathcal{Y}, s \in \{1, \ldots, S\}, \end{aligned} \tag{22}$$

where $\{p_y\}_{y \in \mathcal{Y}}$ is known (or estimated) and $\{\varepsilon_y^s\}_{y \in \mathcal{Y}, s \in \{1, \ldots, S\}}$ are iid draws from Q.

From Proposition 2, we see that the minimand λ of (22) is an estimate of $w \in \partial W^*(p)$.

Computation using Optimal Transport (Contd.)

Chiong et al. (2016) note that $\{\lambda_y\}$ are the Lagrange multipliers to the first set of constraints of the primal problem given by

$$\begin{aligned} \max_{\pi \in \mathbb{R}_+^{J \times S}} \quad & \sum_{y,s} \pi_{ys} \varepsilon_y^s \\ \text{s.t.} \quad & \sum_{s=1}^{S} \pi_{ys} = p_y, & \forall y \in \mathcal{Y}, \\ & \sum_{y \in \mathcal{Y}} \pi_{ys} = \frac{1}{S}, & \forall s \in \{1, \ldots, S\}. \end{aligned} \tag{23}$$

The optimal $\{\lambda_y\}_{y \in \mathcal{Y}}$ can thus be obtained by computing either (22) or (23). Chiong et al. (2016) note that they implement and consider the primal problem (23) for computation.

Computation using Optimal Transport (Contd.)

Note that there is a disconnect between Theorem 2 and the linear programming formulations: Theorem 2 crucially required full support of the latent utility shocks, which cannot hold for the discretized distribution $\hat{Q}$ (for finite S).

Define the identified set of w given p and (a possibly discrete distribution) Q as

$$\mathcal{I}(p) := \left\{ w \in \mathbb{R}^J \middle| \, P_Q \left(Y(w, \varepsilon) = y \right) = p_y, \forall y \in \mathcal{Y} \right\}. \tag{24}$$

Theorem 4 (Galichon, 2016): $\mathcal{I}(p)$ is the set of w such that there exists a z such that (w, z) is a solution to (17). Therefore,

$$\mathcal{I}(p) = \left\{ w \in \mathbb{R}^J \middle| \, \exists z, \, w_y + z_\varepsilon \leq c(y, \varepsilon), \, E_p[w_Y] + E_Q[z_\varepsilon] = W^*(p) \right\},$$

and

$$\mathcal{I}_0(p) = \left\{ w \in \mathbb{R}^J \middle| \, \exists z, \, w_y + z_\varepsilon \leq c(y, \varepsilon), \, E_p[w_Y] = W^*(p), \, E_Q[z_\varepsilon] = 0 \right\}.$$

Computation using Optimal Transport (Contd.)

Theorem 4 suggests a straightforward way of computing upper and lower bounds for each w_y^0 :

1. Construct the discrete approximation $\hat{Q}$ to Q by randomly sampling S iid draws from Q.
2. Solve the linear program in (22) (or (23)) to obtain the objective value $W^*(p)$.
3. Compute a lower bound for w_y^0 via the linear program

$$\begin{aligned} \min_{\lambda \in \mathbb{R}^J, z \in \mathbb{R}^S} \quad & \lambda_y \\ \text{s.t.} \quad & \lambda_y + z_s \geq \varepsilon_y^s, \qquad \forall y \in \mathcal{Y}, s \in \{1, \dots, S\}, \\ & \sum_{y \in \mathcal{Y}} p_y \lambda_y = W^*(p), \\ & \frac{1}{S} \sum_{s=1}^{S} z_s = 0. \end{aligned} \tag{25}$$

An upper bound may be calculated analogously using max instead.

Computation using Optimal Transport (Contd.)

Notice that $(w_y)_{y\in\mathcal{Y}}$ correspond to intercepts in conventional discrete choice models. Often, we are interested instead in parameterized functions, e.g., $w_y(x) = \alpha_y + x_y^\top \beta$, where $x_y \in \mathbb{R}^{d_x}$ is a vector of product characteristics, and $\alpha_y \in \mathbb{R}$ and $\beta \in \mathbb{R}^{d_x}$ are unknown and fixed utility coefficients.

Suppose we observe data $\{(p_{yt}, x_{yt})\}_{y\in\mathcal{Y}, t=1,\ldots,T}$ where T denotes the total number of markets. The linear program formulation in (22) can conveniently be restated to accommodate for such extensions as long as $X \perp\!\!\!\perp \varepsilon$. In particular,

$$\begin{aligned}
\min_{\lambda,z,\alpha,\beta} \quad & \sum_{t=1}^{T}\left[\sum_{y\in\mathcal{Y}} p_{yt}\lambda_{yt} + \frac{1}{S}\sum_{s=1}^{S} z_{st}\right] \\
\text{s.t.} \quad & \lambda_{yt} + z_{st} \geq \varepsilon_{yt}^{s}, \forall y \in \mathcal{Y}, s \in \{1,\ldots,S\}, t \in \{1,\ldots,T\}, \\
& \lambda_{yt} = \alpha_y + x_{yt}^\top \beta, \forall y \in \mathcal{Y}, t \in \{1,\ldots,T\}, \\
& \lambda, \alpha \in \mathbb{R}^J, \beta \in \mathbb{R}^{d_x}, z \in \mathbb{R}^{T\times S}.
\end{aligned} \tag{26}$$

Implementation in Julia

I implement the extended linear programming problem in (26) in Julia. As before all code is readily available via on GitHub (link to `MyMethods.jl`).

A simple illustrative example in a setting with $T = 100$ markets, $J = 3$ products, and product-market characteristics $x_{jt} \in \mathbb{R}^2$ is implemented here.

The implementations are almost surely sub-optimal and inefficient. Possibly as a consequence of poor coding, the mass transport approach even in the simple setting of $T = 100$ and $J = 3$ is *slow*. Using a very crude approximation of Q with $S = 100$ takes approximately 73 seconds (on a 2016 laptop).

The problem also appears to scale poorly in S and T. For example, an even cruder approximation with $S = 10$ only takes 0.5 seconds.

References

Chiong, K. X., Galichon, A., and Shum, M. (2016). Duality in dynamic discrete-choice models. *Quantitative Economics*, 7(1):83–115.

Galichon, A. (2016). *Optimal transport methods in economics*. Princeton University Press.

Galichon, A. and Salanié, B. (2020). Cupid's invisible hand: Social surplus and identification in matching models. *SSRN working paper No 1804623*.

Norets, A. and Takahashi, S. (2013). On the surjectivity of the mapping between utilities and choice probabilities. *Quantitative Economics*, 4(1):149–155.