ECO499 Proposal (if you are short on time, just read the “Personal Statement” section)

Andrew Scutt[1]

[1]Faculty of Economics
University of Toronto

Debrief, August 2025

Table of Contents

1. What I am interested in this opportunity
2. What we know
3. Algorithm setup
4. Research questions
5. First steps
6. Example

Table of Contents

1. What I am interested in this opportunity
2. What we know
3. Algorithm setup
4. Research questions
5. First steps
6. Example

Personal Statement

To begin, I am passionate about applying mathematical and statistical methods to model and predict abstract decision-making processes. In particular, I am deeply invested in developing economic models to evaluate decision-making situations involving trade-offs. Given this set of interests, I would love to participate in the ECO499 course in hopes of contributing to the field of behavioural economics. More specifically, I am interested in developing a population-level utility estimation method for *n*-dimensional items; for more details, please review the deck of slides below. Embarking on this journey would help me refine my preexisting data analysis skills, deepen my understanding of behavioural economics, build my econometric skills, and, ultimately, become a better researcher for future academic endeavors (i.e. PhD). Crucially, this project does involve empirical data analysis, which is outlined in the ”empirical questions” section.

Personal Statement

To collect my data, I was thinking about creating a survey on Qualtrics, then deploying it on Prolific. For funding, I will apply for the USRA, UTEA, and Robert Douglas research grant. To be clear, I have accomplished both aforementioned processes in previous research projects. Finally, I have discussed the possibility of doing this project with Yoram Halevy and Robert Gazzale.

Table of Contents

1 What I am interested in this opportunity

2 What we know

3 Algorithm setup

4 Research questions

5 First steps

6 Example

Knowledge Landscape

In the early 20th century, choice modeling research was restricted to the binomial choice framework: researchers had to limit their inquiries to binary choices between two items. Afterward, they had to infer the overall preference distribution using this set of pairwise comparisons. A prime example would be Thurstone's paper for estimating the importance of certain social values (Thurstone, 1927). Unfortunately, this approach was extremely time-consuming due to the number of comparisons required, being $\binom{n}{2} = \frac{n!}{2!(n-2)!}$, where n is the number of items. In addition, this method allowed intransitive preferences to seep into the overall preference data due to its reliance on raw pairwise comparisons. To account for preferences involving more than two choices, statisticians developed the conditional logistic model, which circumvents intransitivity problems. It allows us to model the choice as a function of item characteristics.

Knowledge Landscape

For the following discussion, let

$$\arg\max_{k \in K} \Pr(Y = k \mid X = x)$$

denote the choice most likely selected by the participant, where K indexes the available bundles. The choice set is

$$S = \{s_1, s_2, \ldots, s_m\},$$

where each

$$s_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,n}) \in \mathbb{R}^n$$

is an n-dimensional bundle. Conditional logit models characterize the log-odds of choosing bundle j over a baseline bundle m as a linear function of attribute differences:

$$\log\left(\frac{P_{i,j}}{P_{i,m}}\right) = \beta_1(x_{i,j,1} - x_{i,m,1}) + \beta_2(x_{i,j,2} - x_{i,m,2}) + \cdots + \beta_p(x_{i,j,n} - x_{i,m,n}),$$

Knowledge Landscape

where $P_{i,j}$ is the probability that participant i selects bundle j. Simply put, this model is trained on a set of options coupled with each participant's choices (yes/no). Then, the user inputs a new set of items to which the algorithm outputs the probability of choosing each item. However, this model assumes a linear relationship across attributes, limiting their ability to capture truly nonlinear effects; our method can overcome this issue. In addition, our statistical approach can extrapolate beyond the observed choice set structure for estimating (preference values)/(choice probabilities), such that $X_{\text{train}} \in \mathbb{R}^{n \times d}, X_{\text{test}} \in \mathbb{R}^{q \times d},$ where $n \neq q$ (number of choices) unlike conditional logistic models.

Table of Contents

1. What I am interested in this opportunity
2. What we know
3. Algorithm setup
4. Research questions
5. First steps
6. Example

Setup

To begin, our research method fundamentally revolves around revealed preference theory; it assumes that consumers' preferences can be revealed through their observed choices. For example, if a consumer picked an apple over an orange, then we say that the apple is revealed as preferred over an orange. To apply this theory to our dataset, three axioms must hold. The weak axiom of revealed preference states that when given bundles *a* and *b* and a budget constraint *B* and constant preferences, no change in *B* can alter the consumer's preferences. The strong axiom of revealed preference declares that transitivity is preserved across preferences for items. For example, if A is directly revealed as preferred over C and C is directly revealed as preferred to B, then A is considered indirectly revealed as preferred to B. Finally, the generalised axiom of revealed preference affirms that if consumption bundle x^i is preferred to x^j, then whenever x^j is selected, the purchase of x^i would have been more expensive, such that $p_j \cdot x^i \geq p_j \cdot x^j$.

Setup

However, this theory comes with limitations in the real world. Firstly, when making explicit choices in the real world, the discarded options are unknown. For example, think of a consumer who picked an orange at a supermarket; the discarded set of goods in preference for purchasing an orange is unknown. Secondly, this theory assumes that the preferences stay internally constant over time. However, various psychology experiments have demonstrated that framing effects alter people's perception of problems, subsequently changing their set of preferences. Finally, this theory assumes that transitivity is preserved across preferences for items, which is not true in practice.

Setup

Let $\mathbb{R}^n$ denote the space of n-dimensional real-valued vectors.

For each $i \in \{1, \ldots, m\}$, let $s_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n}) \in \mathbb{R}^n$ represent an individual n-dimensional item.

Let $S = \{s_1, s_2, \ldots, s_m\} \subseteq \mathbb{R}^n$ be the finite set of all such items.

Let there be N participants, indexed by $p \in \{1, \ldots, N\}$, each of whom provides a strict preference ranking for all s_i in S (i.e. complete and transitive ordering of S).

Setup

- For each participant p, construct a pairwise comparison matrix $M^p \in \mathbb{R}^{m \times m}$, where:

$$M^p_{i,j} = \begin{cases} 1 & \text{if } s_i \succ_p s_j \\ 0 & \text{if } s_j \succ_p s_i \\ \text{Undefined} & \text{if } i = j \end{cases}$$

  Here, $s_i \succ_p s_j$ indicates that participant p prefers item s_i over s_j, and x denotes an undefined diagonal value.

- Aggregate the individual matrices by summing over all participants:

$$A = \sum_{p=1}^{N} M^p$$

Setup

- Normalize the aggregate matrix by dividing each entry by the number of participants:

$$\bar{A} = \frac{1}{N} A$$

- Convert each non-diagonal proportion $\bar{A}_{i,j}$ (i: row j: column) into a z-score by applying the inverse of the standard normal cumulative distribution function (CDF), denoted Φ^{-1}:

$$Z_{i,j} = \Phi^{-1}(\bar{A}_{i,j})$$

- For each item s_j, compute its average z-score:

$$z_j = \frac{1}{m-1} \sum_{\substack{i=1 \\ i \neq j}}^{m} Z_{i,j}$$

Setup

- Interpret the vector $z = (z_1, z_2, \ldots, z_m)$ as the estimated utility values of the items $(s_1, s_2, \ldots, s_m)$, where higher z_j indicates stronger preference under the assumption of normally distributed utilities.
- Afterward, we can represent each item i in the following manner: $(z_i, x_{i,1}, x_{i,2}, \ldots, x_{i,n})$
- Using this vectorial representation coupled with machine learning techniques, such as symbolic regression, KNN, regression trees, multiple linear regression, etc, we can determine how each dimension $x_{i,s}$ contributes to z_i where $s \in \{1, 2, \ldots, n\}$.
- Please note that this algorithm operates under contrived strict preference relations.

Table of Contents

1. What I am interested in this opportunity
2. What we know
3. Algorithm setup
4. Research questions
5. First steps
6. Example

Theoretical Questions

- Does this algorithm output cardinal utility for each item?

Empirical questions

- How accurate is our *n*-dimensional item utility estimation method relative to self-report measures of utility?
- Which machine learning algorithm provides us with the most accurate insights regarding how each element in s_i contributes to z_i?
- Which machine learning algorithm provides us with the most accurate insights regarding how a person values out-of-set items?

Table of Contents

1. What I am interested in this opportunity
2. What we know
3. Algorithm setup
4. Research questions
5. First steps
6. Example

We need to make sure that our algorithm satisfies the following axioms to ensure that it outputs cardinal utility for a population of participants:

Completeness: $\forall x, y \in X, \quad x \succeq y$ or $y \succeq x$

Transitivity: $\forall x, y, z \in X, \quad (x \succeq y \wedge y \succeq z) \implies x \succeq z$

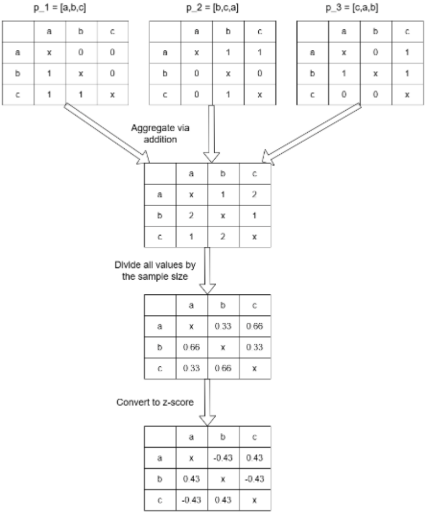
Independence:
$\forall x, y, z \in X, \forall \alpha \in (0,1), \quad x \succeq y \iff \alpha x + (1-\alpha) z \succeq \alpha y + (1-\alpha) z$

Continuity:
$\forall x, y, z \in X, \quad x \succ y \succ z \implies \exists \alpha \in (0,1) : y \sim \alpha x + (1-\alpha) z$

Generate an experimental design to answer our second, third, and fourth questions

Table of Contents

1. What I am interested in this opportunity
2. What we know
3. Algorithm setup
4. Research questions
5. First steps
6. Example



p_1 = [a,b,c]

	a	b	c
a	x	0	0
b	1	x	0
c	1	1	x

p_2 = [b,c,a]

	a	b	c
a	x	1	1
b	0	x	0
c	0	1	x

p_3 = [c,a,b]

	a	b	c
a	x	0	1
b	1	x	1
c	0	0	x

Aggregate via addition

	a	b	c
a	x	1	2
b	2	x	1
c	1	2	x

Divide all values by the sample size

	a	b	c
a	x	0.33	0.66
b	0.66	x	0.33
c	0.33	0.66	x

Convert to z-score

	a	b	c
a	x	-0.43	0.43
b	0.43	x	-0.43
c	-0.43	0.43	x

As a result, z_a = z_b = z_c = 0