Faster estimation of discrete choice models via dataset reduction

22nd Swiss Transport Research Conference
18–20 May 2022

Nicola Ortelli$^{1,2}$, Matthieu de Lapparent$^1$, Michel Bierlaire$^2$

$^1$ IIDE, HEIG-VD
$^2$ TRANSP-OR, EPFL
Outline

1. Background

2. LSH-based dataset reduction

3. Case study

4. Conclusion
Outline

1. Background

2. LSH-based dataset reduction

3. Case study

4. Conclusion
Discrete choice models (DCMs)

What are DCMs?

• Suppose $N$ observations, each containing:
  • a vector of explanatory variables $x_n$;
  • the observed choice $i_n$.

• A DCM calculates the choice probabilities as a function of $x_n$ and $\theta$:

$$P(i \mid x_n; \theta),$$

• where $\theta$ is a vector of model parameters.
Background

Estimating DCMs

Maximum likelihood estimation (MLE)

- Find $\theta$ so as to maximize the joint probability of the observed choices:

$$\max_{\theta} \mathcal{L}(\theta) = \sum_{n=1}^{N} P(i_n | x_n; \theta).$$

- Solved using iterative methods — BFGS, Newton, etc.
- Each iteration is $O(N)$.
- MLE is burdensome for large datasets!
Intuition

Factoring out redundancy

• If the data contains groups of **identical observations**:

\[
L(\theta) = \sum_{u=1}^{U} N_u \cdot P(i_u \mid x_u; \theta),
\]

• \(U\) unique observations.
• Each appears \(N_u\) times in the original data.
• Can we extend this “factorization trick” to **nearly identical observations**?
**Faster estimation of DCMs**

### Assumption
- If $i_p = i_q$ and $x_p \approx x_q$, then $P(i_p \mid x_p; \theta) \approx P(i_q \mid x_q; \theta)$.

### Approach
- Aggregate similar observations together.
- Associate weights.
- MLE on the reduced, weighted sample.

### Challenges
- Clustering must be fast $\Rightarrow$ Use locality-sensitive hashing (LSH).
- Minimize degradation of estimation results.
Outline

1. Background

2. LSH-based dataset reduction

3. Case study

4. Conclusion
Procedure

Example

- 2 alternatives.
- 2 explanatory variables.
Procedure

Data rescaling

\[
\frac{x - \min(x)}{\max(x) - \min(x)}
\]
**LSH-based dataset reduction**

**Procedure**

**LSH function**

\[
    h(x) = \left\lfloor \frac{a \cdot x + b}{w} \right\rfloor
\]

- \(a \sim \mathcal{N}(0, 1)\).
- \(b \sim \mathcal{U}(0, w)\).
- \(w\) is the **bucket width**.

**Example**

\[
    a = \left(\frac{1}{2}, \frac{1}{2}\right), \ b = \frac{1}{5}, \ w = \frac{1}{4}.
\]
Procedure

**LSH function**

\[ h(x) = \left\lfloor \frac{a \cdot x + b}{w} \right\rfloor \]

- \( a \) iid \( \sim \mathcal{N}(0, 1) \).
- \( b \sim \mathcal{U}(0, w) \).
- \( w \) is the bucket width.*

**Example**

\( a = (1, -\frac{1}{2}), \ b = -\frac{1}{10}, \ w = \frac{1}{4}. \)
**LSH-based dataset reduction**

**Procedure**

**LSH function**

\[
h(x) = \left\lfloor \frac{a \cdot x + b}{w} \right\rfloor
\]

- \(a \sim \mathcal{N}(0, 1)\).
- \(b \sim \mathcal{U}(0, w)\).
- \(w\) is the **bucket width**.

**Example**

\[a = (0, 1), \ b = \frac{1}{20}, \ w = \frac{1}{4}.\]
Procedure

**AND-construction**

- Combine \( r \) LSH functions.*
- For any pair of points \((x_p, x_q)\):

\[
h'(x_p) = h'(x_q) \iff h_k(x_p) = h_k(x_q) \ \forall k = 1, \ldots, r.
\]
## Procedure

### Instance selection
- In each bucket, **for each class**, pick 1 instance $x_s$ **at random**.

### Weight assignment
- The associated weight $\omega_s$ is equal to the number of instances of alternative $i_s$ in the bucket.
- $\sum_s \omega_s = N$. 
Procedure

**MLE**

- Data are scaled back.
- Log likelihood function:
  \[ \mathcal{L}(\theta) = \sum_s \omega_s P(i_s | x_s, \theta) \]
Outline

1. Background

2. LSH-based dataset reduction

3. Case study

4. Conclusion
Dataset and models

LPMC data

- Mode choice, 4 alternatives: walk, cycle, drive and public transport.
- 81k observations:
  - 55k for estimation;
  - 26k for out-of-sample validation.

Model

- Multinomial logit.
- 11 continuous attributes.
- 15 binary variables.
- **53 parameters.**
Preliminary results — random sample

![Graph showing estimation time vs. percentage of original dataset size]

- **Estimation time [s]**
- **medians**
- **Q25–Q75**
- **min–max**
- **LSH-DR**

N. Ortelli, M. de Lapparent, M. Bierlaire
Preliminary results — $r = 5$, $w = 0.1, \ldots, 1.0$
Preliminary results—random sample
Preliminary results — $r = 5, w = 0.1, \ldots, 1.0$
Preliminary results — random sample

![Graph showing the relationship between percentage of original dataset size and normalized out-of-sample log likelihood. The graph includes curves representing medians, Q25–Q75, min–max, and LSH-DR. The x-axis represents the percentage of original dataset size ranging from 20% to 100%, while the y-axis shows the normalized out-of-sample log likelihood ranging from -0.710 to -0.700.](image-url)
Preliminary results — $r = 5, \ w = 0.1, \ldots, 1.0$
Outline

1. Background

2. LSH-based dataset reduction

3. Case study

4. Conclusion
Conclusion

Summary

- Reducing computational time comes at the cost of deteriorating results.
- This can be mitigated by carefully sampling observations.
- Factor out redundancy, but keep diversity!

Future work

- Knowledge-based LSH.
- Informed sampling from the buckets.
- Embed dataset reduction within MLE.
Faster estimation of discrete choice models via dataset reduction

22nd Swiss Transport Research Conference
18–20 May 2022

Nicola Ortelli\textsuperscript{1,2}, Matthieu de Lapparent\textsuperscript{1}, Michel Bierlaire\textsuperscript{2}

\textsuperscript{1} IIDE, HEIG-VD
\textsuperscript{2} TRANSP-OR, EPFL