JEFFREY NEWMAN, LAURIE GARROW

COMPUTATIONAL APPROACHES FOR EFFICIENT ESTIMATION OF DISCRETE CHOICE MODELS

1000011001010011 1
0001010110010100 1

BIG
DATA

1100101001110010 1
1000011001010011 1
0001010110010100 1
0100110110112011 1
1101101010110101 0

- Big Data is here, is everywhere, is cheap, is ready, yay! 😁

10000110010100111
00010101100101001
BIG
DATA
11001010011100101
10000110010100111
00010101100101001
01001101101120111
11011010101101010

- Big Data is here, is everywhere, is cheap, is ready, yay! 😁
- Wait, stop, too much!
My model crashed 😭😭😭

10000110010100111
00010101100101001
BIG
DATA
11001010011100101
10000110010100111
00010101100101001
01001101101120111
11011010101101010

- Big Data is here, is everywhere, is cheap, is ready, yay! 😁
- Wait, stop, too much! My model crashed 😭😭😭
- Big data sometimes is best leveraged by small models — but a bunch of them
- Today we'll discuss a couple ways to efficiently stitch together smaller models to get the BIG model we really want

THE UNDERLYING ARCHITECTURE

- Work in Python with standard tools for data processing and analysis: **numpy** + **scipy** + **pandas** + **larch**
- **Larch** complements the general tools to add discrete choice model estimation and application capabilities, and provides all the tools needed to build and estimate connected sets of models together

Market Segmentation via Linked Models

more models, less problems

AIRLINE ITINERARY CHOICE DATA

- We developed models to evaluate the impacts of price endogeneity on airline itinerary choice
- The discrete choice data used for model estimation is **transactional data** from a airline ticket clearinghouse
  - Up to 156 itinerary alternatives in each choice set,
  - over 1 million choice sets,
  - over 3 million directional itineraries, or
  - over 10 million passenger trips.

Multiple similar customers see the same choices

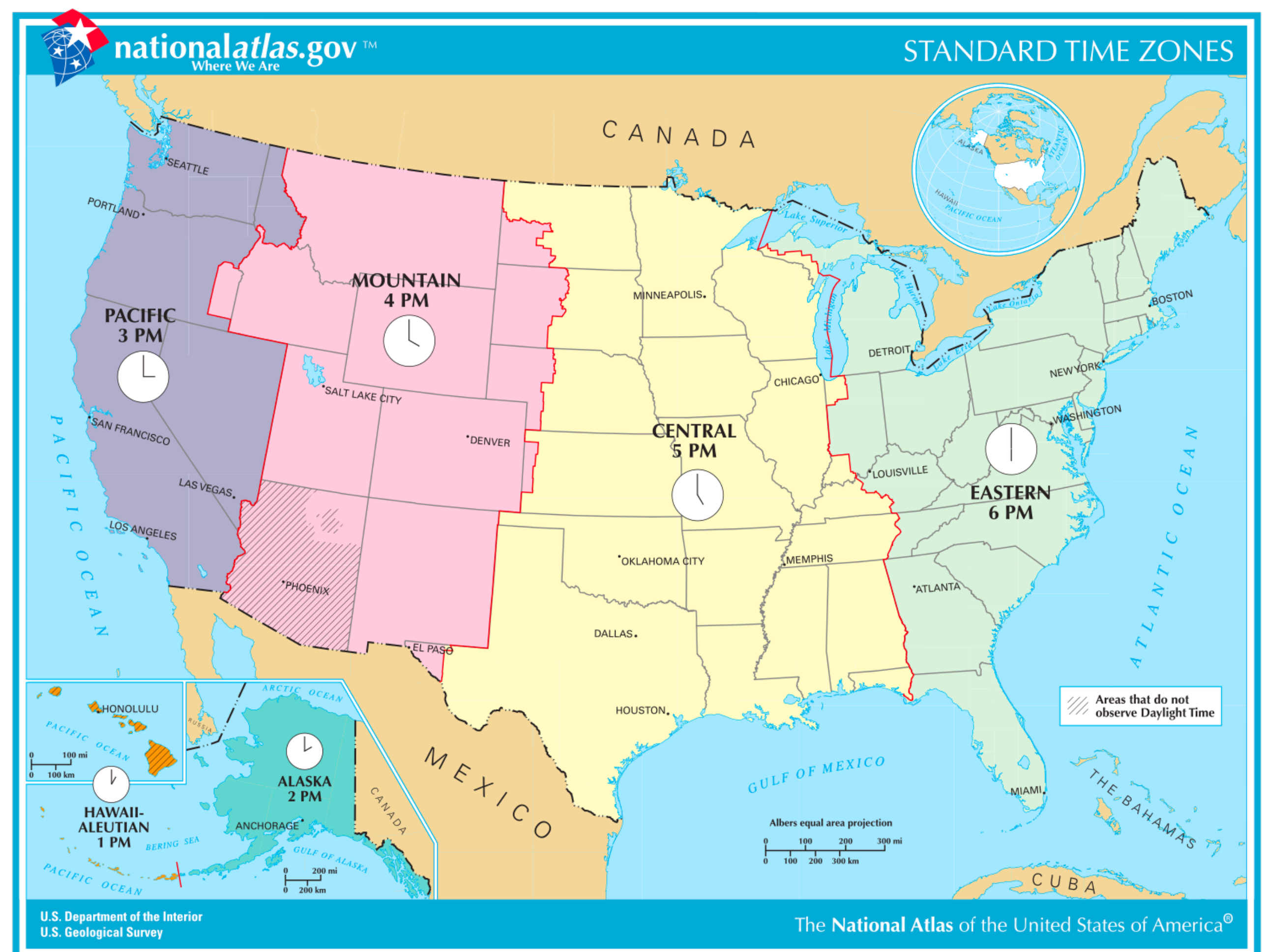
Multiple passengers on a single itinerary

MARKET SEGMENTS FOR AIR TRAVEL

Market segments for time of day preferences delineated by:

- 3 directionality types (outbound, return, & one-way),
- 7 days of week, and
- 10 geographic segments:
  - distance,
  - number of time zones traversed
  - direction of travel (e.g., east-to-west).

= 210 market segments



THE DESIRED UTILITY SPECIFICATION

16 generic universal parameters
(e.g. fare, travel time, number of connections, equipment type, etc.)

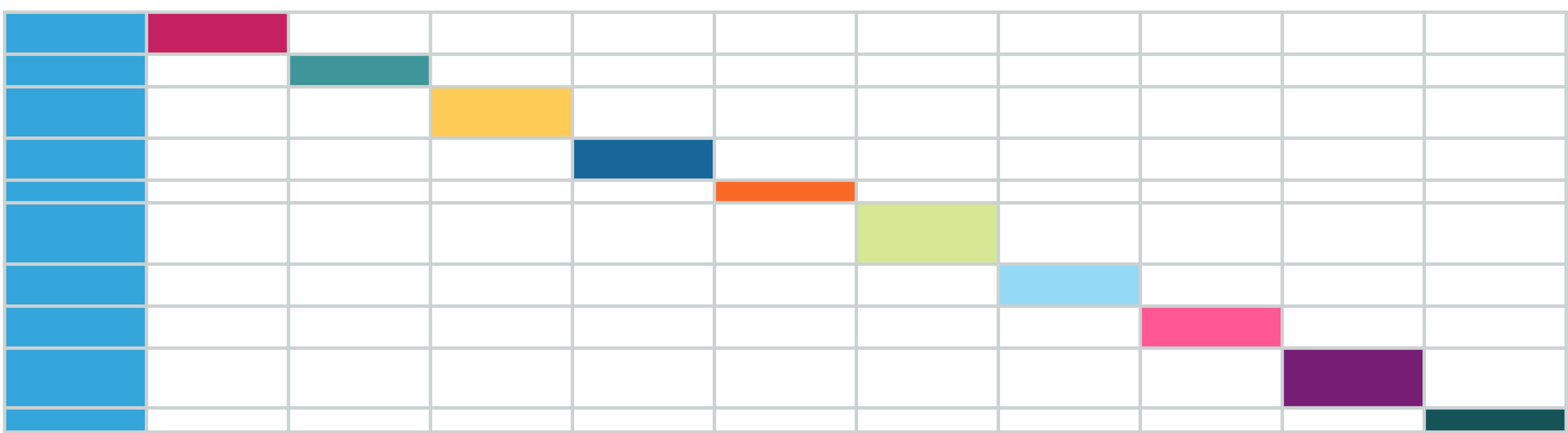
\+ **6** departure time-of-day preference parameters for each market segment

× **210** distinct market segments

= **1,272** parameters

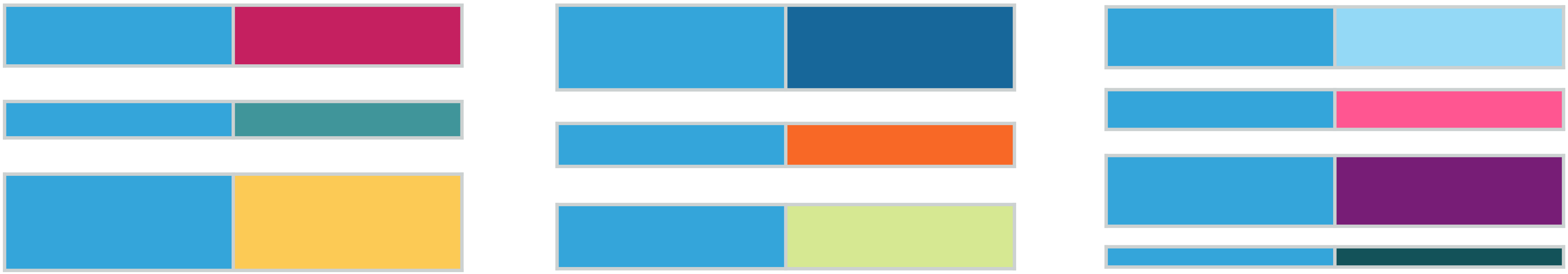
BRUTE FORCE IS TOO BRUTE

- A naïve approach: one model with market segmented variables
  - 1276 variables × 156 alternatives × 1M choice sets ≈ 1.6 terabytes of raw exogenous data
- But, this "raw" data is extraordinarily sparse



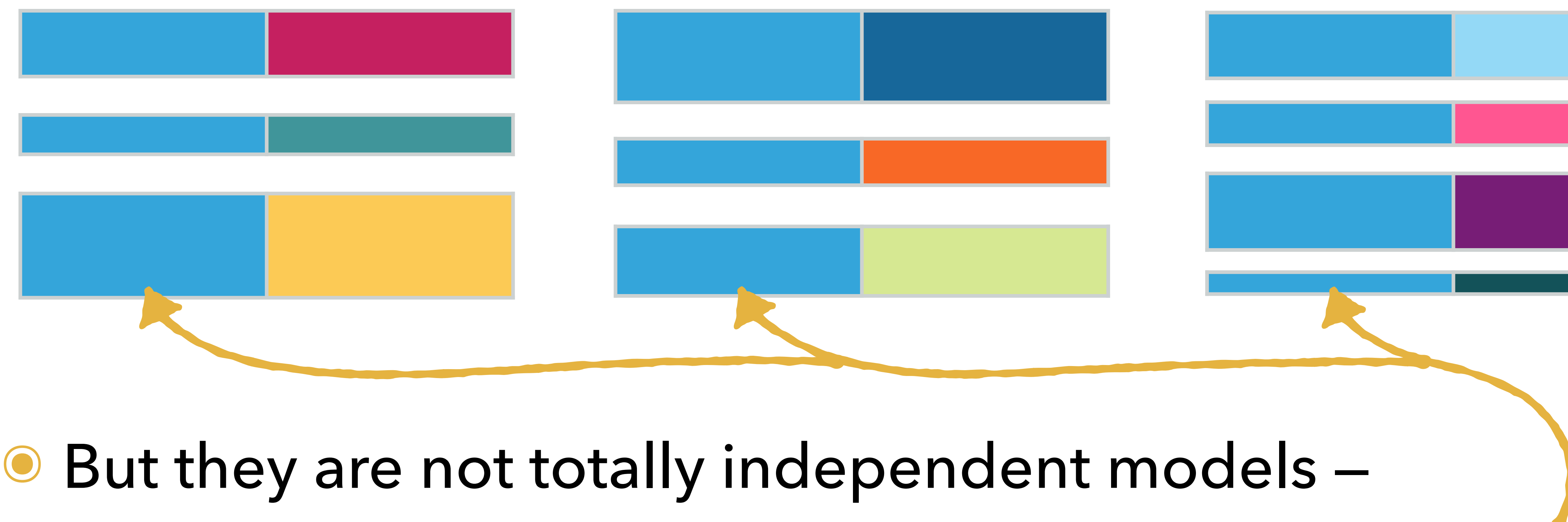
SEPARATE MODELS ESTIMATED TOGETHER

- Instead create a set of small models, one for each market segment
  - Each model only includes the relevant data and parameters



SEPARATE MODELS ESTIMATED TOGETHER

- Instead create a set of small models, one for each market segment
  - Each model only includes the relevant data and parameters



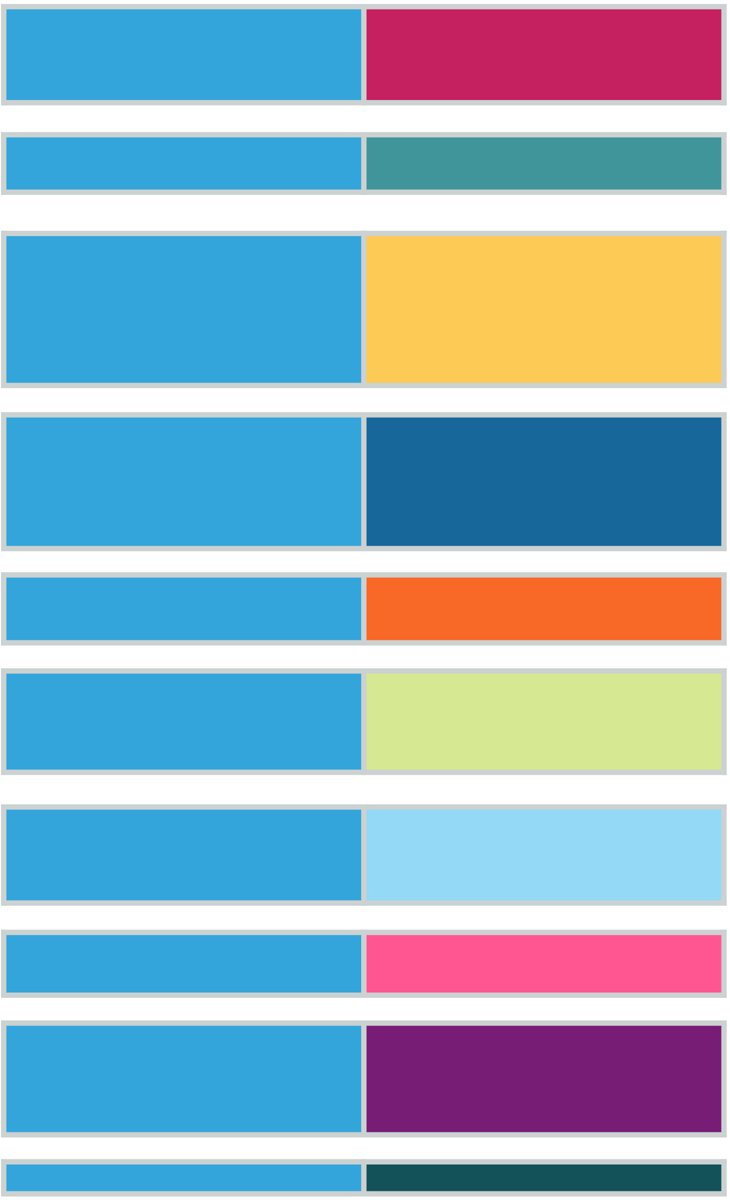
- But they are not totally independent models –
  they share some parameters
  – so they must still be estimated jointly

GOOD NEWS: ESTIMATING TOGETHER IS EASY

- Log likelihood of the joint meta-model is just the sum of the log likelihoods of the parts

$$LL(\beta) = \sum_{i \in {\color{red}\blacksquare}{\color{green}\blacksquare}{\color{blue}\blacksquare}} LL_i(\beta)$$

- The derivative of the log likelihood w.r.t. any parameter is also just the sum of the parts
- Or: the sum over the common parameters plus concatenation of the segment-unique parameters



Non-Normalized Nested Logit

Hello, old friend

MUCH MALIGNED
OFTEN IGNORED
YET SURPRISINGLY USEFUL

- The **non-normalized nested logit** (**NNNL**) is typically not preferred
- Most applications focus on the MNL model, because it is **easy** and **fast**
- Travel demand forecasters using a nested model usually stick to the **utility maximizing nest logit** (**UMNL**) because it is theoretically "correct"
  - *With one exception: ALOGIT users*

Georgia Tech

UTILITY MAXIMIZING NESTED LOGIT IS TANTALIZINGLY CLOSE TO MULTINOMIAL LOGIT

$$LL(\beta,\mu) = \sum_{i \in \mathbf{C}} \delta_i \log\left(P_{i|\mathbf{n}_i}(\beta,\mu) P_{\mathbf{n}_i}(\beta,\mu)\right)$$

$$= \sum_{i \in \mathbf{C}} \delta_i \log\left(P_{i|\mathbf{n}_i}(\beta,\mu)\right) + \sum_{i \in \mathbf{C}} \delta_i \log\left(P_{\mathbf{n}_i}(\beta,\mu)\right)$$

You can split it into parts that look so familiar

$$\frac{\exp\left(\frac{V_i(\beta)}{\mu_{\mathbf{n}_i}}\right)}{\sum_{j \in \mathbf{n}_i} \exp\left(\frac{V_j(\beta)}{\mu_{\mathbf{n}_i}}\right)}$$

$$\frac{\exp\left(\mu_{\mathbf{n}_i} \Gamma_{\mathbf{n_i}}(\beta,\mu)\right)}{\sum_{\mathbf{m} \in \mathbf{N}} \exp\left(\mu_{\mathbf{m}} \Gamma_{\mathbf{m}}(\beta,\mu)\right)}$$

$$\log \sum_{k \in \mathbf{m}} \exp\left(\frac{V_k(\beta)}{\mu_{\mathbf{m}}}\right)$$

But none of these parts is quite exactly a plain old MNL model

NON NORMALIZED NESTED LOGIT IS BASICALLY JUST A BUNCH OF MULTINOMIAL LOGIT MODELS

$$LL(\beta,\mu) = \sum_{i\in\mathbf{C}} \delta_i \log\left(P_{i|\mathbf{n}_i}(\beta,\mu) P_{\mathbf{n}_i}(\beta,\mu)\right)$$
$$= \sum_{i\in\mathbf{C}} \delta_i \log\left(P_{i|\mathbf{n}_i}(\beta,\mu)\right) + \sum_{i\in\mathbf{C}} \delta_i \log\left(P_{\mathbf{n}_i}(\beta,\mu)\right)$$

This part is exactly a regular MNL model

$$\frac{\exp\left(V_i(\beta)\right)}{\sum_{j\in\mathbf{n}_i} \exp\left(V_j(\beta)\right)}$$

$$\frac{\exp\left(\mu_{\mathbf{n}_i}\Gamma_{\mathbf{n_i}}(\beta,\mu)\right)}{\sum_{\mathbf{m}\in\mathbf{N}} \exp\left(\mu_{\mathbf{m}}\Gamma_{\mathbf{m}}(\beta,\mu)\right)}$$

$$\log \sum_{k\in\mathbf{m}} \exp\left(V_k(\beta)\right)$$

This part is close enough to make it easy to use the same methods

WHAT DO YOU GET

- ⊕ Big computational speed gains for certain model structures
- ⊕ Best improvements when there are a lot of alternatives and few nests

WHAT DO YOU LOSE

- ⊖ Need to impose constraints on parameters that you probably were going to use anyhow
- ⊖ Need to scale parameters back to UMNL form if you want to have consistency

WHAT KIND OF SPEED BOOST ARE WE TALKING ABOUT HERE?

- An example: we estimated a usual workplace destination choice model on about 16,000 observations
- Approximately 5,000 zonal alternatives nested together
- One “work at home” alternative by itself (not nested with others)
- NNNL model estimation completes in 28 minutes
- UMNL model estimation abandoned after several hours

None of these tricks really matter unless your data or model (most likely both) is at least kind of

BIG

If your model is **SMALL** and you are not frustrated by long estimation time or unwieldy memory requirements, then don't try to fix it

http://larch.readthedocs.io

jpn@gatech.edu