September 15, 2018
The 17th Summer Course for Behavior Modeling in Transportation Networks
@The University of Tokyo

Application of AI for travel behavior modelling in urban networks

Makoto Chikaraishi
Hiroshima University

A major concern when using machine learning (ML) methods for modelling travel behavior

- **What is a *good* behavior model?**
  1. High predictability
  2. High interpretability
     - Particularly solid microeconomic foundations
- **General Features of machine learning (ML)**
  - Very high predictability for short-run forecasting
    - Not really sure for long-run forecasting
  - **Little theoretical foundation**
    - Even it is difficult to identify the factors affecting the outcome when using deep learning techniques.
- **Today's contents**
  - Review some recent studies which attempt to solve the shortcomings of ML methods in the context of modeling discrete choice behavior.

Papers reviewed:

- **“Replacement” rather than “integration”** (may not be very interesting for behavioral modelers)
  - Acuna-Agost, R., Delahaye, T., Lheritier, A., Bocamazo, M., 2017. Airline Itinerary Choice Modelling Using Machine Learning. International Choice Modelling Conference 2017.
  - Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications 78, 273-282.
  - Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention 108, 27-36.
  - Yang, J., Shebalov, S., Klabjan, D., 2017. Semi-supervised learning for discrete choice models. arXiv preprint arXiv:1702.05137.
- **Integration (1): Discrete choice with decision trees**
  - Brathwaite, T., Vij, A., Walker, J.L., 2017. Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice. arXiv preprint arXiv:1711.04826.
- **Integration (2): Discrete choice with neural network**
  - Sifringer, B., Lurkin, V., Alahi, A., 2018. Enhancing Discrete Choice Models with Neural Networks. 18th Swiss Transport Research Conference, Monte Verità, May 16–18.

Brathwaite, T., Vij, A., Walker, J.L., 2017. Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice. arXiv preprint arXiv:1711.04826.

DISCRETE CHOICE WITH DECISION TREES

Background and objective

- **Background**
  - The logistic regression model from statistics and the binary probit model from psychology were **linked with random utility theory**.
  - Recently, the fields of statistics, computer science, and machine learning have created numerous methods for modeling discrete choices, while these newer methods have not been derived from or linked with economic theories of human decision making.
- **Objective**
  - Bridging the gap by providing a **microeconomic framework for decision trees**

Contributions

- **Major contributions of the paper**
  1. Connect decision trees to economic theory, where decision trees correspond to a **non-compensatory**, microeconomic decision protocol known as "disjunctions-of-conjunctions"
  2. Advance the state of the art in the modeling of **semi-compensatory** decision making by combining decision trees with traditional discrete choice models.
  3. Demonstrate the performance of the proposed method (focus: mode choice)

Non-compensatory decision protocols

- **Compensatory decision protocols**
  - High levels of satisfaction with one attribute compensate for low levels of satisfaction with other attributes.
- **Non-compensatory decision protocols**
  - Not always allow positive attributes of a given alternative to compensate for negative attributes of that same alternative.
  - Not typically require the evaluation of all attributes of all alternatives. They better capture the limited cognitive resources of decision makers.

Basic idea

- Manski's (1977) two-stage characterization of the choice process:

Decision trees

$$P(i) = \sum_{C \subseteq \Delta(M)} \boxed{P(i|C)} \boxed{Q(C)}$$

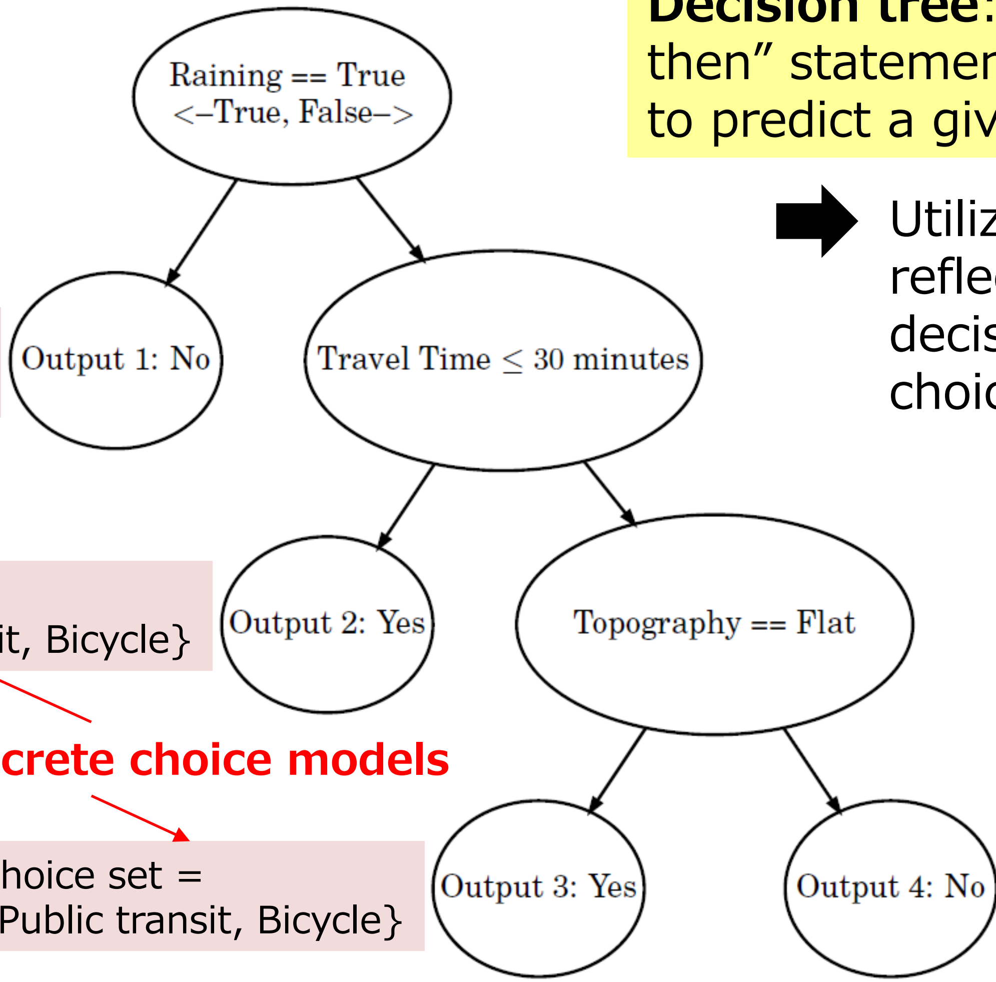
Conventional discrete choice such as MNL

$P(i)$	: Probability of choosing alternative *i*
C	: A choice set in the set of subsets of *M*, $\Delta(M)$
$P(i\|C)$	: Conditional probability of choice given set *C*
$Q(C)$	: The probability that *C* is the true choice set

Basic idea

Decision tree: A set of “if-then” statements that are used to predict a given quantity.

➡ Utilizing decision trees to reflect no-compensatory decision protocol of choice set formation



Choice set = {Public transit}

Choice set = {Public transit, Bicycle}

Applying discrete choice models

Choice set = {Public transit, Bicycle}

Choice set = {Public transit}

Figure 1: Example decision tree for bicycle consideration

Different non-compensatory decision rules

- Dominance (Cascetta and Papola, 2009)
- Lexicography (Kohli and Jedidi, 2007)
- Elimination-by-aspects (Tversky, 1972)
- Satisficing (Stuttgen et al., 2012)
- Conjunctive rules
- Disjunctive rules
- Subset-conjunctive rules
- Disjunctions-of-conjunctions

Related to this paper

Different non-compensatory decision rules

- **Conjunctive rules**
  - An individual only considers alternatives that meet all of a given number of requirements.
- **Disjunctive rules**
  - An individuals only considers alternatives that meet at least one of a given set of requirements.
- **Subset-conjunctive rules**
  - A generalization of both conjunctive rules and disjunctive rules.
  - An individual only considers alternatives that meet a certain number of requirements.
- **Disjunctions-of-conjunctions**
  - A generalization of conjunctive, disjunctive, and subset-conjunctive decision rules.
  - An individual considers any alternative that meets at least one of a given set of conjunctive conditions.
  - **Highly flexible non-compensatory decision protocols.**

Linking decision trees with disjunctions-of-conjunctions

- **Conjunctive rule:**

  $$\text{if } (\prod_{i=1}^{B} b_i) == 1 \text{ then } y$$

  – If all requirements b_i (noted as p_i) are met, then y.

- **Disjunctive rule:**

  $$\text{if } (\sum_{i=1}^{B} b_i) \geq 1 \text{ then } y$$

  – If at least one (i.e., if any) of the requirements b_i are met, then y.

- **Disjunctions-of-conjunctions rule:**

  $$\text{if } (\sum_{i=1}^{D} \prod_{j=1}^{|p_i|} b_j^i) \geq 1 \text{ then } y$$

  – If at least one of some set of conjunctive conditions, p, is met, then y.

B : Total number of requirements in the rule

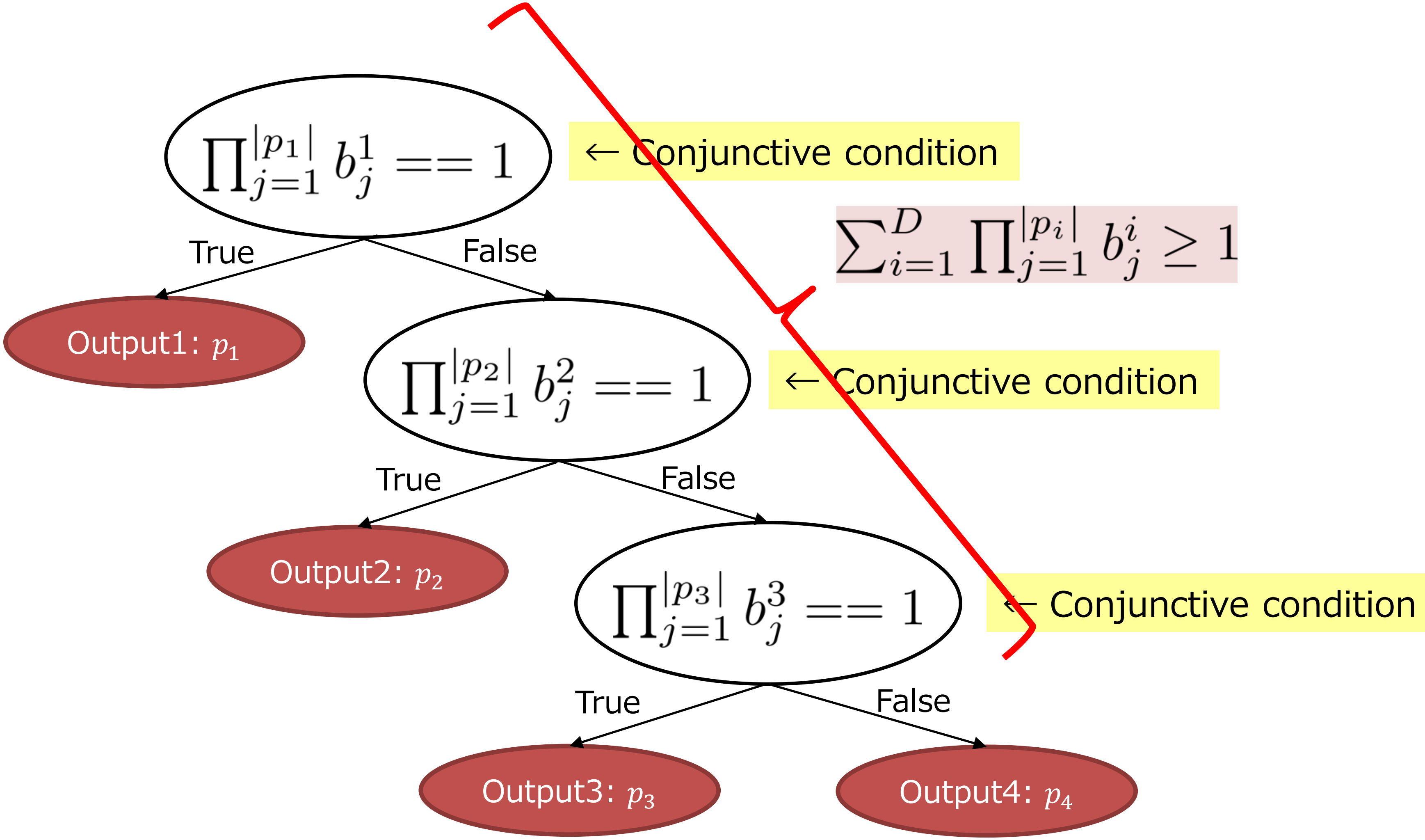
b_i : A primitive Boolean statement (1: True; 0: False)

y : Outcome

D : Total number of conjunctive conditions

$|p_i|$: The number of requirements that make up conjunctive condition p_i

Linking decision trees with disjunctions-of-conjunctions



Enumeration of conjunction conditions can be done by using the FP-growth algorithm (Letham et al., 2015). ZDD-growth algorithm (Minato, 2006) may also be able to use.

Decision tree variants

- Decision tree models can be extended to:
    1. make **probabilistic** predictions
    2. represent **heterogeneity** in a population's non-compensatory rules
    3. represent estimation **uncertainty**
    4. represent **context-dependent** preference heterogeneity
    5. satisfy **monotonicity** constraints
- These extensions are not new, but can be econometrically explained!!

1. probabilistic predictions

- A conventional decision tree involves deterministic outputs through “if-then” rules.
  - However, decisions may not be deterministic in many contexts.
- We can make it probabilistic, for example:
  - The probability of a given alternative is predicted to be the fraction of observations in that output node who chose the alternative in question (Arentze and Timmermans, 2004)

2. Heterogeneous non-compensatory rule

- Individuals may use different non-compensatory rules. There are number of methods reflecting the heterogeneous rules.
- **Local heterogeneity:**
  - <u>Heterogeneity within a certain node</u>
    - Soft decision trees / fuzzy decision trees
- **Global heterogeneity:**
  - <u>Heterogeneity in the structure of decision tree</u>
    - “Ensembles” of decision trees (considering latent classes for decision trees).
    - Similar with random forests, but the classes may need to be behaviorally understandable.

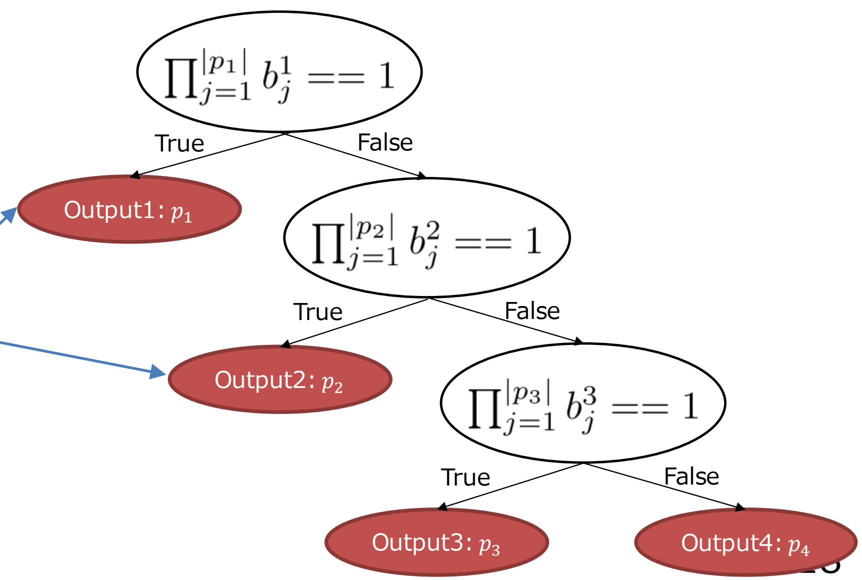
3. estimation uncertainty

- Quantification of inferential uncertainty is important.
- Ensemble methods such as Bayesian decision trees and bagging can be used to obtain the “approximate” measure of uncertainty.

4. context-dependent preference

- The context in which a decision is made is an important determinant of outcomes (Swait et al., 2002).
- Model trees
  - decision trees where the output at a given output node is a statistical model (in this paper, discrete choice model)

Set different parameters to take into account context-dependent preference heterogeneity.



5. monotonicity constraints

- Constraints are often needed to economically understand the model:
  - As the travel cost increases, the probability of choosing the alternative should decrease (monotonicity constraint)
    - We could reflect it by using monotonic decision trees, where the desired monotonicity constraints are not violated.

Empirical model specification

$$\hat{P} = \sum_{m=1}^{M} P_{Post}(Y|X,m)P_{Post}(m)$$

$$= \sum_{m=1}^{M} \left[\frac{1}{S_m}\sum_{s=1}^{S_m} P(Y|X,\gamma_s^m,m)\right] P_{Post}(m)$$

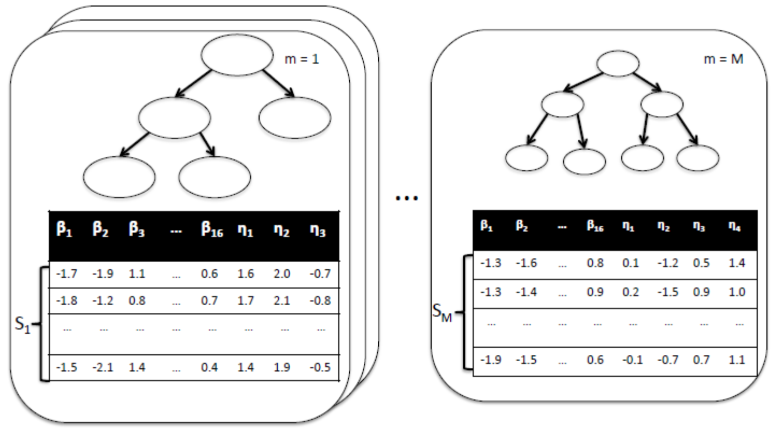


Figure 3: Procedural diagram of bayesian model trees

$\hat{P}$: the predicted probability of outcome *Y*

M : the total number of unique trees in one's sample

$P_{Post}(Y|X,\gamma_s^m,m)$: the choice model probability of *Y* given X,γ_s^m,m

$P_{Post}(m)$: the posterior probability of a given tree $(P_{Post}(m) = \frac{S_m}{\sum_l S_l})$

S_m : the number of sampled elements containing tree *m*

γ_s^m : a set of parameters at node *m*

Computationally very expensive, requiring an efficient estimation method

Empirical analysis

- Data set
  - California Household Travel Survey data
  - 1,015 observations
  - Choice context: Mode choice with 8 alternatives

$$
\begin{aligned}
V_{DA} &= \beta_{\text{travel-time-auto}} \text{TravelTime}_{\text{DA}} + \beta_{\text{autos-per-driver}} \text{AutosPerDriver} \\
V_{SR2} &= \text{ASC}_{\text{shared-ride-2}} + \beta_{\text{travel-time-auto}} \text{TravelTime}_{\text{SR2}} + \beta_{\text{autos-per-driver}} \text{AutosPerDriver} \\
&\quad + \beta_{\text{cross-bay}} \text{CrossBay} + \beta_{\text{num-kids}} \text{NumberKids} + \beta_{\text{household-size}} \text{HouseholdSize} \\
V_{SR3} &= \text{ASC}_{\text{shared-ride-3}} + \beta_{\text{travel-time-auto}} \text{TravelTime}_{\text{SR3}} + \beta_{\text{autos-per-driver}} \text{AutosPerDriver} \\
&\quad + \beta_{\text{cross-bay}} \text{CrossBay} + \beta_{\text{num-kids}} \text{NumberKids} + \beta_{\text{household-size}} \text{HouseholdSize} \\
V_{WTW} &= \text{ASC}_{\text{walk-transit-walk}} + \beta_{\text{travel-time-transit}} \text{TravelTime}_{\text{WTW}} + \beta_{\text{travel-cost-transit}} \text{TravelCost}_{\text{WTW}} \\
V_{WTD} &= \text{ASC}_{\text{walk-transit-drive}} + \beta_{\text{travel-time-transit}} \text{TravelTime}_{\text{WTD}} + \beta_{\text{travel-cost-transit}} \text{TravelCost}_{\text{WTD}} \\
V_{DTW} &= \text{ASC}_{\text{drive-transit-walk}} + \beta_{\text{travel-time-transit}} \text{TravelTime}_{\text{DTW}} + \beta_{\text{travel-cost-transit}} \text{TravelCost}_{\text{DTW}} \\
V_{walk} &= \text{ASC}_{\text{walk}} + \beta_{\text{distance-walk}} \text{TravelDistance}_{\text{walk}} \\
V_{bike} &= \boxed{\text{ASC}_{\text{bike}}} + \beta_{\text{distance-walk}} \text{TravelDistance}_{\text{bike}}
\end{aligned}
$$

Different across discrete choice models

Empirical analysis

- Variables for decision trees (requirements):

- Number of Kids: $[0,1]$, $[2]$, and $[3,\infty)$
- Minimum distance (miles): $[0,1.17]$, $(1.17,1.92]$, $(1.92,3.00]$, $(3.00,4.37]$, $(4.37,\infty)$
- Average Speed Limit (miles per hour): $[23.01,25.15]$, $(25.15,25.78]$, $(25.78,\infty)$
- Median Slope (meters per foot): $[0,0.01]$, $(0.01,0.02]$, $(0.02,0.03]$, $(0.03,0.04]$, $(0.04,\infty)$
- Proportion of roadway miles along one's shortest path with speed limits < 25 miles per hour: $[0,0.66]$, $(0.66,0.83]$, $(0.83,0.95]$, $(0.95,0.9984]$, $(0.9984,0.9986]$, $(0.9986,\infty)$
- Proportion of roadway miles with bicycle lanes: $[0,0.04]$, $(0.04,0.11]$, $(0.11,\infty)$
- Proportion of roadway miles with "share the road" markings: $[0,0.08]$, $(0.08,0.14]$, $(0.14,\infty)$

Notes:
1. Discretization was done to construct the decision tree into as many equal sized groups as possible.
2. The maximum number of requirements in a conjunctive condition is 2.
3. Only consider the conjunctive conditions that apply to (1) 10% or more of those who bicycle or (2) 10% or more of those who did no bicycle.
4. Enumeration of conjunctive conditions were done by using the FP-growth algorithm.

Some results

- Accuracy:
  – The proposed Bayesian model tree is more than 1000-times more likely to be closer to true data-generating process than the MNL model.
    - The cost of model complexity is somehow already taken into account in the model estimation process.
- Forecasts:

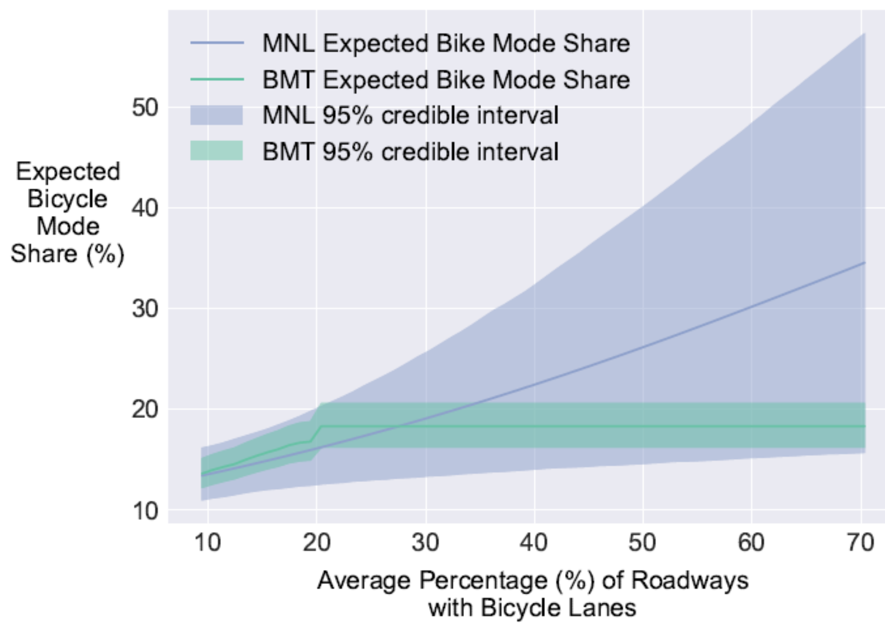


Figure 4: Expected Bicycle Mode Share versus Mean Percentage of Bicycle Lanes

Sifringer, B., Lurkin, V., Alahi, A., 2018. Enhancing Discrete Choice Models with Neural Networks. 18th Swiss Transport Research Conference, Monte Verità, May 16–18.

DISCRETE CHOICE WITH NEURAL NETWORK

Background and objective

- **RUM model vs neural network**
  - Advantage of RUM model
    - Interpretability of the results.
  - Advantage of neural network
    - Better goodness-of-fit
- **Objective**
  - Bringing the predictive strength of Neural Networks, a powerful machine learning-based technique, to the field of Discrete Choice Models (DCM) without compromising interpretability of these choice models.

RUM: random utility maximization

RUM model and neural network (NN)

Discrete choice model as a Random Utility Maximization (RUM) model

Utility function: $$U_{in} = \beta_1 x_{1in} + \ldots + \beta_d x_{din} + \varepsilon_{in} \qquad \forall i \in C_n$$
$$= V_{in} + \varepsilon_{in}$$

Choice probability: $P(i|C_n) = P(U_{in} > \max_{j \neq i}(U_{jn})) = \frac{\exp(V_{in})}{\sum_{j \in C_n} \exp(V_{jn})}$

(negative) log-likelihood: $LL = -\sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \log[P(i|C_n)]$

A discrete choice model from the perspective of neural network

Softmax activation function: $(\sigma(\mathbf{V}_n))_i = \frac{\exp(V_{in})}{\sum_{j \in C_n} \exp(V_{jn})}$

Cross-entropy: $H_n(\sigma, \mathbf{y}_n) = -\sum_{i \in C_n} y_{in} \log[(\sigma(\mathbf{V}_n))_i]$

The conventional MNL can be seen as a neural network model with a simple network structure.

Discrete Choice Model with NN

Utility function with non-linear component:

$$\mathbf{U}_n = \boldsymbol{\beta}\boldsymbol{\chi}^T + \mathbf{u}_n + \boldsymbol{\varepsilon}_n$$

Linear-in-parameters component

Non-linear component (via NN)

$\mathbf{U}_n$: $\{U_{1n}, U_{2n}, ..., U_{In}\}$

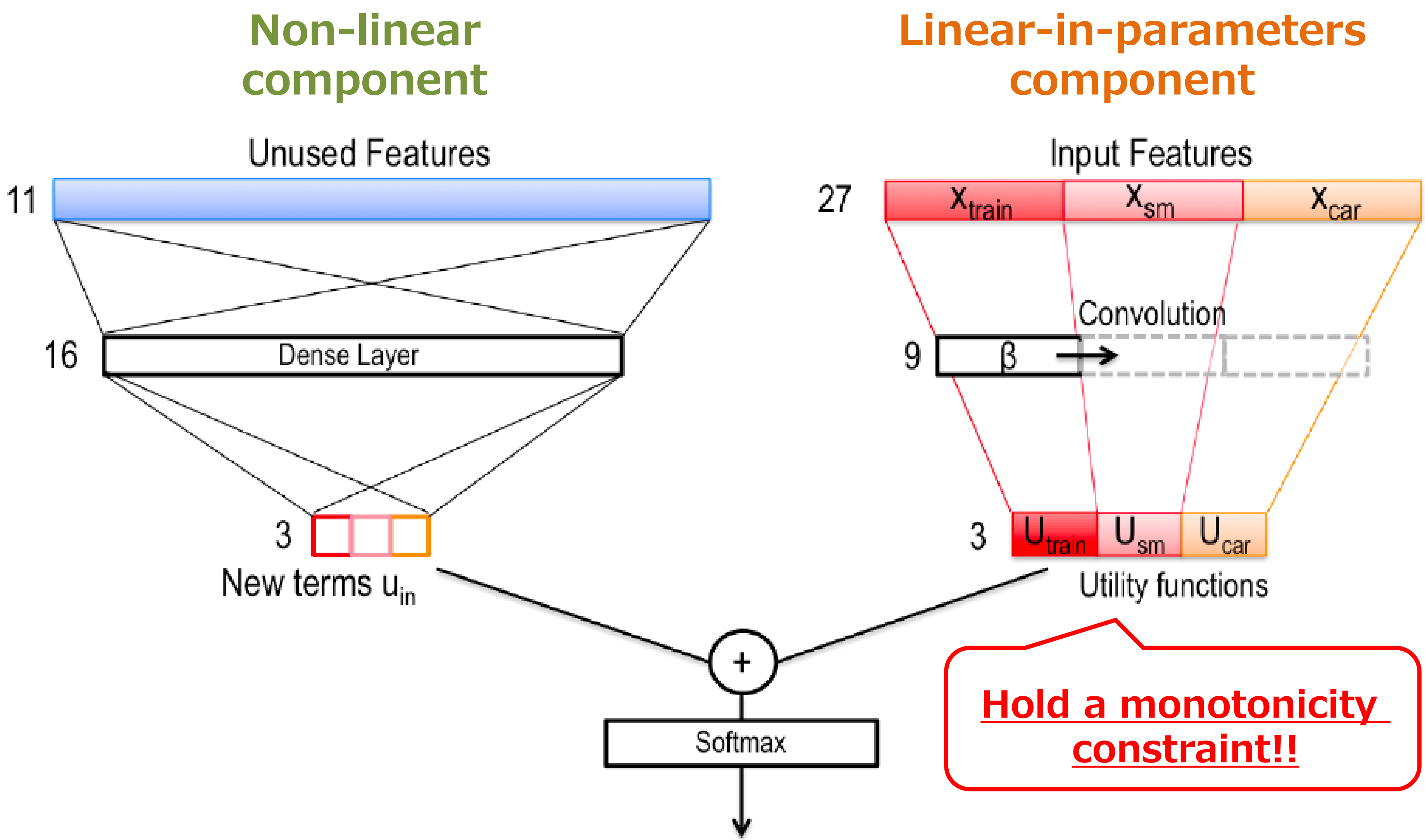
$\boldsymbol{\beta}$: A vector of parameters ($1 \times d$)

$\boldsymbol{\chi}$: A set of explanatory variables ($I \times d$)

$$\mathbf{u}_n = \psi(\mathbf{Q})$$

where $\mathbf{Q}$ is the ensemble of input features, and, ψ is the function defined by multiple neural network layers and their corresponding activation functions.

Discrete Choice Model with NN



Empirical analysis

- Dataset
  - Swissmetro dataset (Bierlaire et al., 2001)
  - A stated preference data on mode choice
  - 10700 entries from 1190 individuals
- Linear-in-parameters component:

Variable		Alternative		
		Car	Train	Swissmetro
ASC	Constant	Car-Const		SM-Const
TT	Travel Time	B-Time	B-Time	B-Time
Cost	Travel Cost	B-Cost	B-Cost	B-Cost
Freq	Frequency		B-Freq	B-Freq
GA	Annual Pass		B-GA	B-GA
Age	Age in classes		B-Age	
Luggage	Pieces of luggage	B-Luggage		
Seats	Airline seating			B-Seats

Empirical analysis

- Non-linear component:
  1. **Travel purpose**: Discrete value between 1 to 9 (Business, leisure, travel,...)
  2. **First class**: 0 for no or 1 for yes if passenger is a first class traveler in public transport
  3. **Ticket**: Discrete value between 0 to 10 for the ticket type (One-way, half-day, ...)
  4. **Who**: Discrete value between 0 to 3 for who pays the travel (self, employer, ...)
  5. **Male**: Traveler’s gender, 0 for female and 1 for male
  6. **Income**: Discrete value between 0 to 4 concerning the traveler’s income per year
  7. **Origin**: Discrete value defining the canton in which the travel begins
  8. **Dest**: Discrete value defining the canton in which the travel ends

Multinomial Logit as Benchmark

Table 2: MNL parameter values

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_{Car}	1.20	0.183	6.58	0.00
2	ASC_{SM}	1.19	0.182	6.53	0.00
3	β_{age}	0.175	0.0512	3.41	0.00
4	β_{cost}	-0.00690	0.000577	-11.97	0.00
5	β_{freq}	-0.00704	0.00116	-6.09	0.00
6	β_{GA}	1.54	0.168	9.17	0.00
7	$\beta_{luggage}$	-0.113	0.0479	-2.36	0.02
8	β_{seats}	0.432	0.115	3.76	0.00
9	β_{time}	-0.0129	0.000842	-15.34	0.00

Number of observations = 7234

$$\mathcal{L}(\hat{\beta}) = -5766.705$$

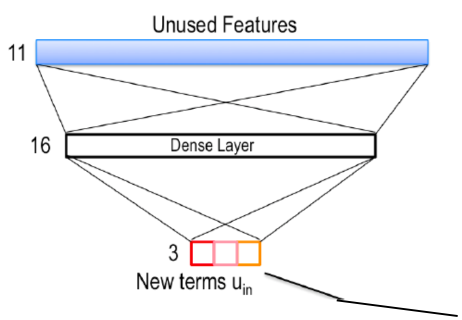
Hybrid model (1)

Table 3: Hybrid Model parameter values

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_{Car}	0.0652	0.179	0.37	0.71
2	ASC_{SM}.	0.327	0.171	1.92	0.06
3	β_{age}	0.376	0.0464	8.12	0.00
4	β_{cost}	-0.0141	0.000595	-23.63	0.00
5	β_{freq}	-0.00807	0.00123	-6.55	0.00
6	β_{GA}	0.130	0.181	0.72	0.47
7	$\beta_{luggage}$	0.0153	0.0505	0.30	0.76
8	β_{seats}	0.207	0.106	1.95	0.05
9	β_{time}	-0.0157	0.000952	-16.53	0.00
10	β_{NN}	1.24	0.0524	23.74	0.00

Number of observations = 7234

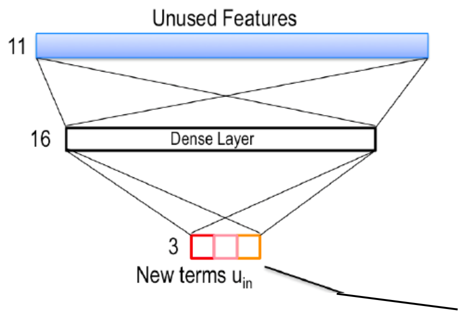
$\mathcal{L}(\hat{\beta}) = -5008.996$



Note: Statistical properties of the parameters are obtained through Biogeme (Bierlaire, 2009)

Simplified hybrid model (2)

Table 4: Hybrid model containing only values of greater interest



Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_{Car}	0.966	0.0977	9.89	0.00
2	ASC_{SM}	1.13	0.0941	11.97	0.00
3	β_{cost}	-0.0165	0.000666	-24.71	0.00
4	β_{freq}	-0.00820	0.00129	-6.38	0.00
5	β_{time}	-0.0171	0.000853	-20.05	0.00
6	β_{NN}	1.25	0.0854	14.65	0.00

Number of observations = 7234

$\mathcal{L}(\hat{\beta}) \quad = \quad -4894.539$

All remaining variables are used here

Conclusions & future works

Conclusions:

- Combining the advantage of linear-in-parameters RUM model and the advantage of neural network where highly non-linear impacts of explanatory variables

Future works:

- The selection of hyper parameters (it would change the results)
- Possibility of using the model for long-term demand forecasting (cross-validation may not be enough)
- Possibility of using different NN components (e.g., convolutional NN, recurrent NN, etc.)

Comparison of key parameters

Table 6: Parameter ratio comparison

Parameter	MNL	Hybrid	Simple Hybrid
β_{cost}	100.0%	204.3%	239.1%
β_{freq}	100.0%	114.6%	116.5%
β_{time}	100.0%	121.7%	132.5%
Value of Time	0.54	0.89	0.96
Value of Frequency	0.98	1.75	2.01
Final Log-Likelihood	-5766.71	-5009.00	-4894.54
Number or parameters	9	10	6

Take-home message:

- There is a high possibility of utilizing machine learning techniques to improve behavioral models, while satisfying basic requirements such as having solid microeconomic foundations