Enhancing discrete choice models with representation learning

表現学習を用いた離散選択モデルの強化

Brian Sifringer, Virginie Lurkin, Alexandre Alahi. Enhancing discrete choice models with representation learing. Transportation Research Part B 140 (2020) 236-261.

理論談話会#14 (2024/06/13)

M1 三上侑希

Outline

Abstract

- 離散選択モデルでは，モデルの誤設定により，予測可能性の制限や推定パラメータの偏りにつながる可能性がある.
- In discrete choice modeling, model misspecifications may lead to limited predictability and biased parameter estimates.

Novelty

- 確定効用を**知識駆動項**と**データ駆動項に**分割することで，モデルの解釈可能性を犠牲にせずに予測力を向上させる.
- By dividing the deterministic terms into **knowledge-driven parts** and **data-driven parts**, this formulation increases the predictive power of the models without sacrificing their interpretability.
- MNLとNLをNNから生じる新たな非線形表現で強化した**Learning-MNL(L-MNL)**と**Learning-NL(L-NL)**を提案.
- They suggests Learning-MNL models and Learning-NL models with a new non-linear representation arising from NN.

Usefulness and Reliability

- 合成データと実データを用いた数値実験で，予測性能とパラメータ推定の精度の両方において，L-MNLが従来のモデルを上回ることを示す.
- By conducting experiments using synthetic data or real-world data, L-MNL models outperform the traditional ones both in terms of predictive performance and accuracy in parameter estimation.

- Source Code: https://github.com/BSifringer/EnhancedDCM

Outline

- Abstract
- Introduction
- Related work
- Implementing MNL as a neural network
- Representation learning in discrete choice modeling
- Experiments
- Future directions
- Conclusions

Introduction

解釈可能性と予測性能 / Interpretability and Predictive Power

- MNL
  - 単純でパラメトリックな効用関数を用いる→解釈可能性○：パラメータの比から時間価値などの指標を簡単に求められる
  - Use simple parametric specifications in the utility function. → Good at Interpretability, such as VOT.
  - 解釈可能性を向上させると予測力が犠牲となる→予測性能✕：データを根本的な構造を完全にとらえられない
  - Sacrifice Predictive Power when gaining Interpretability. → cannot adequately capture the underlying structure of the data
- Complex DCM models (MXL, Latent Class Model, …) and Advanced utility specifications (exponential, non-parametric, …)
  - 複雑なモデルと発展的な効用関数の設定で，データ適合量と予測力向上
  - More complex models and More advanced utility specifications allow for a better data fit and a better prediction.
  - モデルの設定が予め分かっている必要があるが，その設定を求めるのが難しい
  - Model specifications need to be known a priori, but determining the specifications remains a difficult task.
- NN
  - 予測性能は優れており，変数間の根底にある真の関係を予め知っておく必要がない
  - Good at Predictive Power and you don’t have to know the nature of true relationships among variables.
  - 出力が得られる過程がブラックボックス→変数を用いた解釈が不可能
  - The process generating outputs is black-box. So, you lose interpretability.

Outline

- Abstract
- Introduction
- Related work
- Implementing MNL as a neural network
- Representation learning in discrete choice modeling
- Experiments
- Future directions
- Conclusions

Related work

交通分野におけるデータ駆動型手法の適用 / Applying data-driven methods in different transportation applications

- 様々なタイプのNNが以下のモデルや手法と比較されてきた
- Different types of NN have been compared to the following models and methods
  - MNL(Agrawal and Schorling, 1996; Lee et al., 2018; Zhao et al., 2020)
  - NL(Mohammadian and Miller, 2002; Hensher and Ton, 2000)
  - Random Utility models(Sayed and Razavi, 2000; Cantarella and de Luca, 2005; Paredes et al., 2017)
  - Statistical methods(West et al., 1997; Karlaftis and Vlahogianni, 2011; Iranitalab and Khattak, 2017; Golshani et al., 2018; Brathwaite et al., 2017)
- 解釈可能性を議論せずに，予測力の観点からモデルを比較している
- Without discussing interpretability, comparing the models in terms of prediction power

機械学習分野 / The machine learning community

- 様々な選択肢を高い精度で予測する研究　Researches aiming at predicting, with high accuracy, a variety of choices
  - Hagenauer and Helbich(2017)：To classify travel mode choice, comparing MNL, NN, Naive Bayes(Rish et al., 2001), Gradient Boosting Machine(Friedman, 2001), Bagging(Breiman, 1996), Random Forests(Breiman, 2001) and Support Vector Machine(Cortes and Vapnik, 1995)
  - And so on…
- 同上　Same as above

Related work

両分野の比較にとどまらない革新的な行動研究 / Innovative behavioral studies beyond the comparison of the both fields

- Wonget et al. (2018)
  - Using a restricted Boltzmann Machine (Ackley et al., 1985) to represent latent behavior attributes.
- van Ctanenburgh and Alwosheel(2019)
  - Develop a novel NN based approach to investigate decision rule heterogeneity amongst travelers.
- その他，ニューラルネットワークを活用した研究がいくつか見られる．Some other studies using NN exist.
- これらの研究には，分かりやすい解釈可能性を維持しながら高い予測性能を可能とする目的はない．
- These studies do not have the objective of finding a utility specification that allows high predictability while maintaining straightforward interpretability.

最も近い研究 ―ブランド選択について― / The closest studies about the brand choice

- NN-MNLモデル(Bentz and Merunka, 2020)
  - Using NN to discover non-linear effects in utility function.
  - Then, the re-specified MNL model is modified to include new variables which is the discovered non-linear effects.
- Using NN only to determine the model specification → Estimating MNL model
- So, interpretability…？→×

Outline

- Abstract
- Introduction
- Related work
- Implementing MNL as a neural network
- Representation learning in discrete choice modeling
- Experiments
- Future directions
- Conclusions

Implementing MNL as a neural network

- Neural Networkの一般的な表現
  - 隠れ層hidden layer $h^{(j)}$を通して，入力空間input space xを関心のある出力interest output Uにマッピングする関数で構成される
$$\boldsymbol{U} = \boldsymbol{h}^{(L)}\left(\boldsymbol{q}^{(\boldsymbol{L-1})}\right) with\ \boldsymbol{q}^{(j)} = \boldsymbol{h}^{(j)}\left(\boldsymbol{q}^{(j-1)}\right), \forall j = 1, \dots, L$$
  - $\boldsymbol{q}^{(0)} = \boldsymbol{x}$で，**L**は最後の表現層
- Convolutional Neural Network
  - 畳み込みを利用，層間を接続するフィルタ状の重みを持つ
$$h_i^{(j+1)} = g\left(\sum_{k=0}^{d} h_{(s\cdot i+k)}^{(j)}\,\beta_k^{(j)} + \alpha_i^{(j)}\right)$$
  - $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_d\}$: the filter of size $(1 \times d)$，s: the stride of the convolution，α_i: a bias term，$g(\cdot)$: an activation function
  - 単層$L = 1$，活性化関数$g(x) = x$，ストライド$s = d$とするとMNLモデルとなる
  - ソフトマックス関数(softmax function)で選択確率が求められる←ロジット型の選択確率式と同じ
$$\left(\boldsymbol{\sigma}(\boldsymbol{V}_n)\right)_i = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{in}}}$$
  - 損失関数には，categorical cross-entropy(CE) (Shannon, 1948)を用いて，最小化←尤度最大化の符号反転
$$\min \sum_{n=1}^{N} H_n(\boldsymbol{\sigma}, \boldsymbol{y}_n) = -\sum_{i \in C_n} y_{in} \log[\sigma_i(\boldsymbol{V}_n)] \Leftrightarrow \max \mathcal{L} = \sum_{n=1}^{N} \sum_{i \in C_n} y_{in} \log[\sigma_i(\boldsymbol{V}_n)]$$

Implementing MNL as a neural network

- An illustrative example

$$U_{in} = \beta_c \cdot x_{1i} + \beta_t \cdot x_{2i} + \varepsilon_{in}, \forall i \in C$$

  - $x_1 = cost$: travel cost, $x_2 = time$: travel time

選択肢集合はすべての個人について同じ
The choice set is the same for all individuals.
SM: Swissmetro

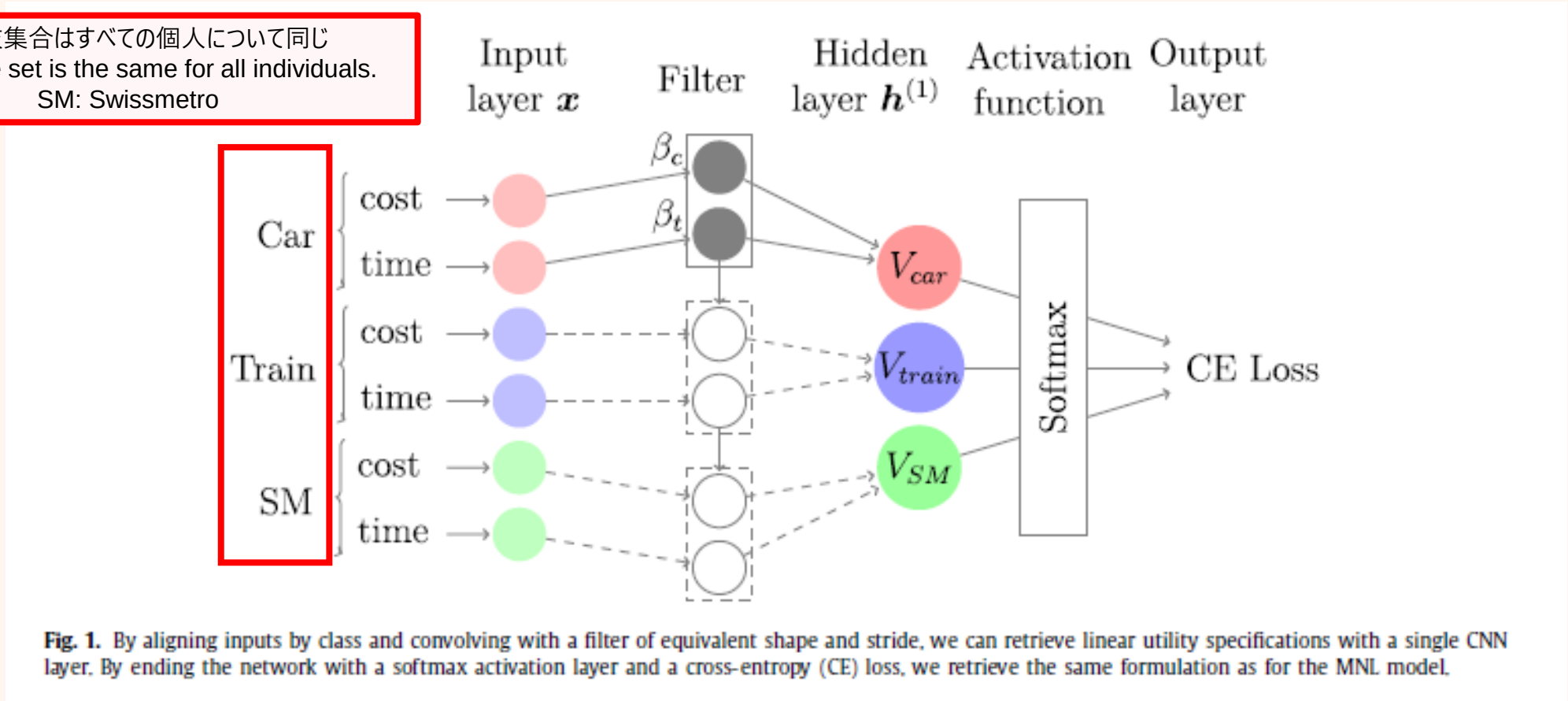

Fig. 1. By aligning inputs by class and convolving with a filter of equivalent shape and stride, we can retrieve linear utility specifications with a single CNN layer. By ending the network with a softmax activation layer and a cross-entropy (CE) loss, we retrieve the same formulation as for the MNL model.

Implementing MNL as a neural network

- An illustrative example

$$U_{in} = \beta_c \cdot x_{1i} + \beta_t \cdot x_{2i} + \varepsilon_{in}, \forall i \in C$$

  - $x_1 = cost$: travel cost, $x_2 = time$: travel time

選択肢集合はすべての個人について同じ
The choice set is the same for all individuals.
SM: Swissmetro

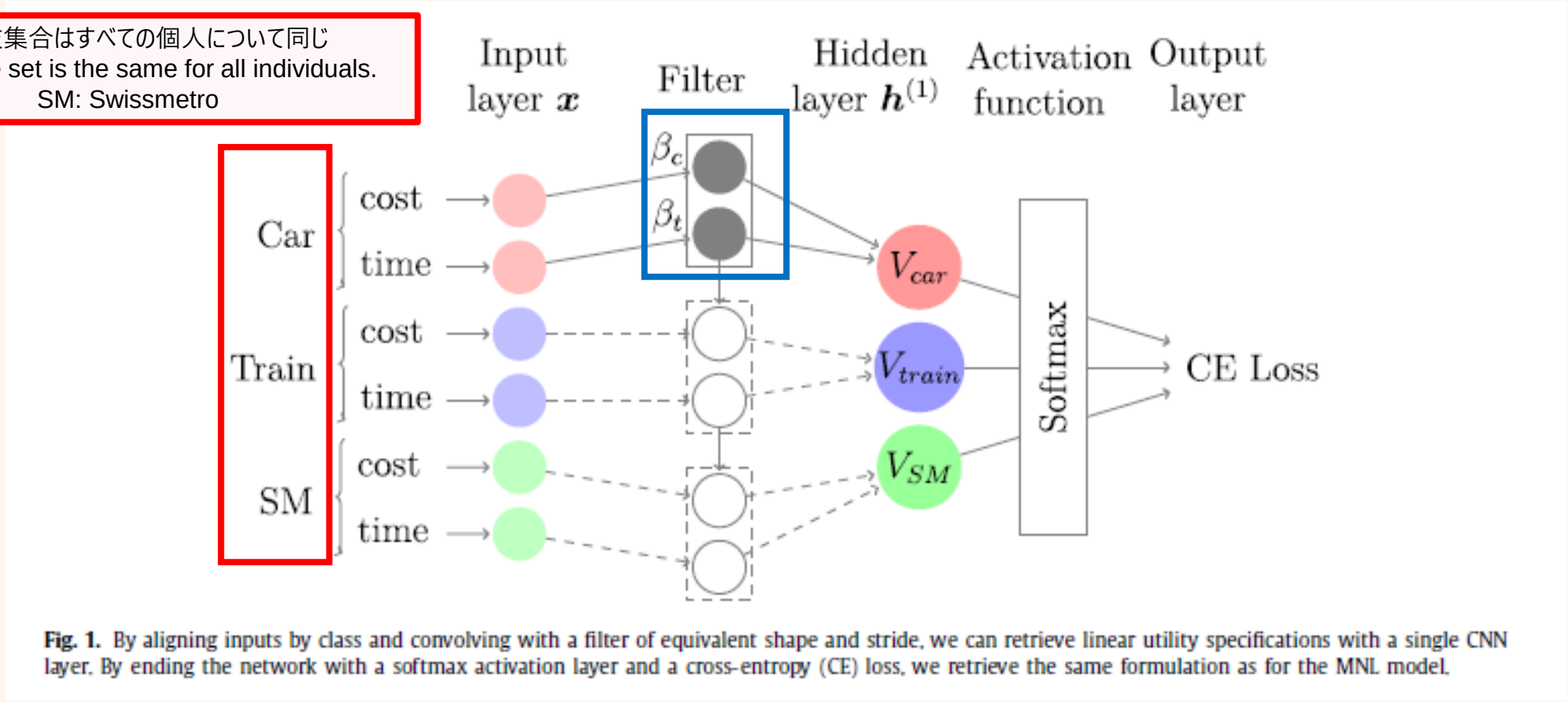

Fig. 1. By aligning inputs by class and convolving with a filter of equivalent shape and stride, we can retrieve linear utility specifications with a single CNN layer. By ending the network with a softmax activation layer and a cross-entropy (CE) loss, we retrieve the same formulation as for the MNL model.

Implementing MNL as a neural network

- An illustrative example

$$U_{in} = \beta_c \cdot x_{1i} + \beta_t \cdot x_{2i} + \varepsilon_{in}, \forall i \in C$$

  - $x_1 = cost$: travel cost, $x_2 = time$: travel time

選択肢集合はすべての個人について同じ
The choice set is the same for all individuals.
SM: Swissmetro

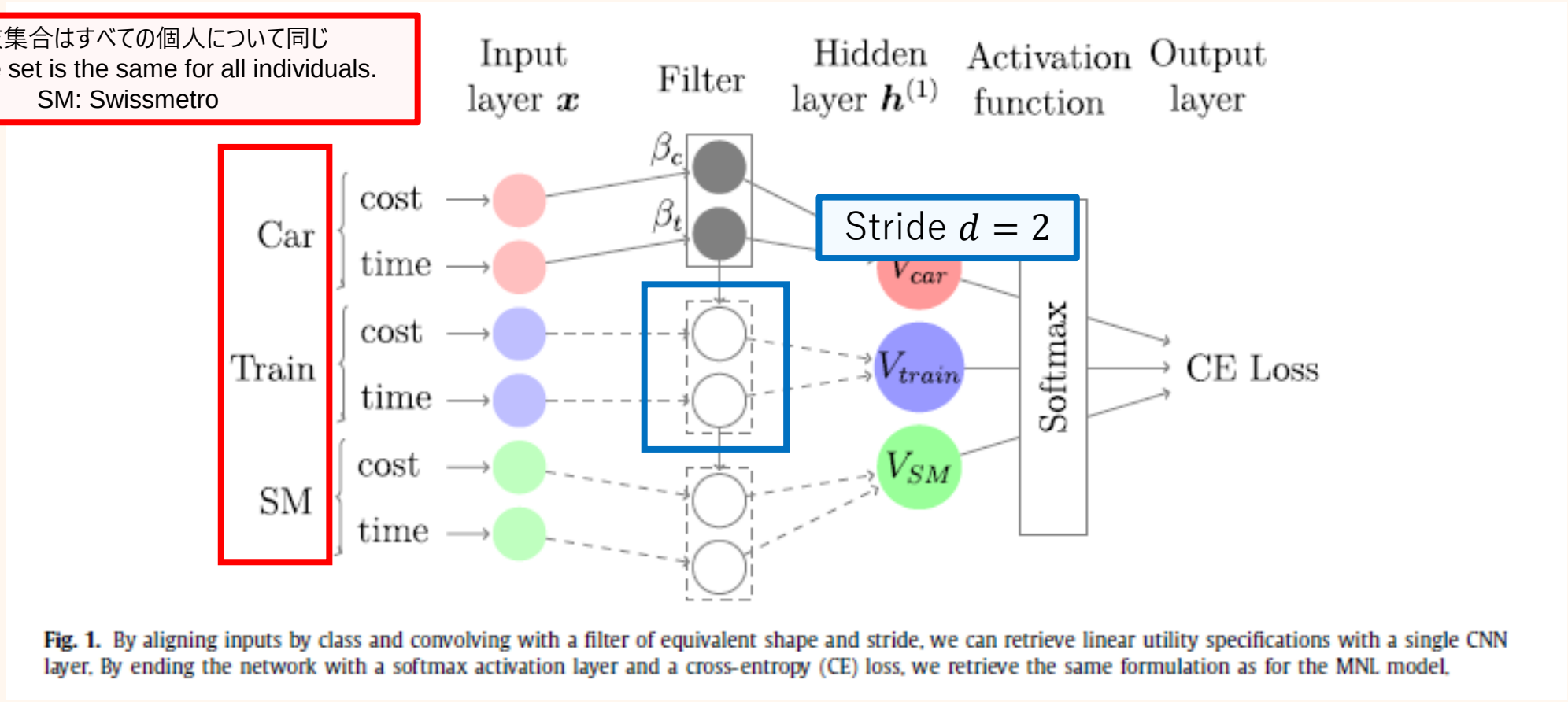

Fig. 1. By aligning inputs by class and convolving with a filter of equivalent shape and stride, we can retrieve linear utility specifications with a single CNN layer. By ending the network with a softmax activation layer and a cross-entropy (CE) loss, we retrieve the same formulation as for the MNL model.

Implementing MNL as a neural network

- An illustrative example

$$U_{in} = \beta_c \cdot x_{1i} + \beta_t \cdot x_{2i} + \varepsilon_{in}, \forall i \in C$$

  - $x_1 = cost$: travel cost, $x_2 = time$: travel time

選択肢集合はすべての個人について同じ
The choice set is the same for all individuals.
SM: Swissmetro

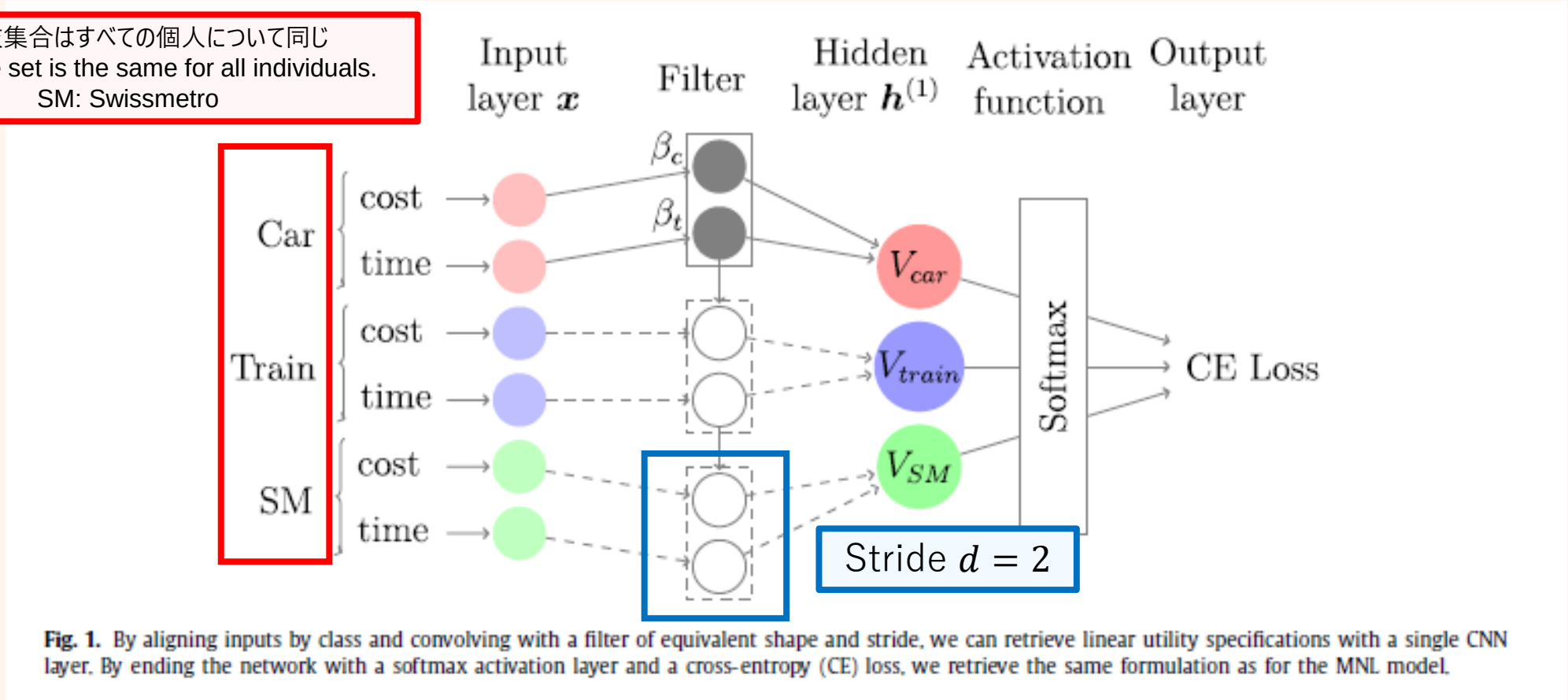

Fig. 1. By aligning inputs by class and convolving with a filter of equivalent shape and stride, we can retrieve linear utility specifications with a single CNN layer. By ending the network with a softmax activation layer and a cross-entropy (CE) loss, we retrieve the same formulation as for the MNL model.

Outline

- Abstract
- Introduction
- Related work
- Implementing MNL as a neural network
- Representation learning in discrete choice modeling
- Experiments
- Future directions
- Conclusions

Representation learning in discrete choice modeling

General formulation

$$V_{in} = f_i(\mathcal{X}_n; \boldsymbol{\beta}) + r_i(\mathcal{Q}_n; \boldsymbol{w})$$

Knowledge-driven part, assumed interpretable
$\mathcal{X}_n$: Input set: explanatory variables, $\boldsymbol{\beta}$: parameters for each $x \in \mathcal{X}_n$

Data-driven part, no a priori relationship is assumed
$\mathcal{Q}_n$: Input set: features, $\boldsymbol{w}$: weights for each $q \in \mathcal{Q}_n$

$$\boldsymbol{U_n} = \boldsymbol{f}(\mathcal{X}_n; \boldsymbol{\beta}) + \boldsymbol{r}(\mathcal{Q}_n; \boldsymbol{w}) + \boldsymbol{\varepsilon_n}$$

- この方程式の第2項は，手作業でモデル化された関数の残差を求めるデータ駆動項として解釈できる
- We may interpret the second term of this equation as a data-driven term finding the residual of a hand-modeled function

- 離散選択モデルの解釈可能性を維持するため，入力集合を2つ$(\mathcal{X}, \mathcal{Q})$に分けているのが特徴
- This formulation differs with the use of two input sets $\mathcal{X}, \mathcal{Q}$ necessary for keeping discrete choice interpretability

Representation learning in discrete choice modeling

Modeling

$$\boldsymbol{U_n = f(\mathcal{X}_n; \beta) + r(\mathcal{Q}_n; w) + \varepsilon_n}$$

have all the benefits of expert modeling in discrete choice. include the availability of the β parameters' Hessian and thus their standard deviation approximation.

produce a new representation of its associated input the Hessian is generally not available due to computational complexity

- ある特徴量tが選択肢iに与える影響(the elasticity of a feature t on alternative i)

$$\frac{\partial U_{in}}{\partial t_{in}} = \frac{\partial f_{in}}{\partial t_{in}} + \frac{\partial r_{in}}{\partial t_{in}}$$

- 解釈可能性を維持するため，弾力性が表現項に依存してはならないとする → $\frac{\partial r_{in}}{\partial t_{in}} = 0$ (The interpretability condition)
- The elasticity must not depend on the representation term.

Representation learning in discrete choice modeling

Modeling

$$\boldsymbol{U}_n = \boldsymbol{f}(\mathcal{X}_n; \boldsymbol{\beta}) + \boldsymbol{r}(\mathcal{Q}_n; \boldsymbol{w}) + \boldsymbol{\varepsilon}_n$$

have all the benefits of expert modeling in discrete choice. include the availability of the β parameters' Hessian and thus their standard deviation approximation.

produce a new representation of its associated input the Hessian is generally not available due to computational complexity

$$\bar{\varepsilon}_{in} = r_i(\mathcal{Q}_n; \boldsymbol{w}) + \varepsilon_{in}$$

- $\boldsymbol{f}(\mathcal{X}_n; \boldsymbol{\beta})$において不偏のパラメータ推定値を得るためには，内生性として知られる効用項とランダム項との相関を回避し，各選択肢間のランダム項の相関を回避し，全体として効用設定の誤りを回避しなければならない.
- To have unbiased parameter estimates in $\boldsymbol{f}(\mathcal{X}_n; \boldsymbol{\beta})$, one must avoid correlation between the specification and the random terms known as endogeneity, avoid correlation of the random terms between each alternative and overall avoid utility misspecification.

- 我々のモデルは，データ駆動型の手法によって，誤った効用の設定と省略変数バイアスによるアンダーフィットを補正することを目的としている.
- Our model aims at correcting for underfit due to misspecification and omitted variable bias thanks to data-driven methods.

Representation learning in discrete choice modeling

L-MNL Model formulation

- Probability of selecting the choice alternative i for individual n

$$P_n(i) = \frac{e^{f_i(\mathcal{X};\boldsymbol{\beta})+r_i(\mathcal{Q}_n;\boldsymbol{w})}}{\sum_{j\in C_n} e^{f_i(\mathcal{X};\boldsymbol{\beta})+r_i(\mathcal{Q}_n;\boldsymbol{w})}}$$

- Using a Dense Neural Network as the learning method. Representation term r_{in} is the resulting function of a DNN

$$r_{in} = \sum_{k=1}^{H} w_{ik}^{(L)} g\left(\boldsymbol{q}_n^{(L-1)}\boldsymbol{w}_k^{(L-1)} + \alpha_k^{(L-1)}\right) + \alpha_i^{(L)}$$

  - $g(\cdot)$ is the rectifier linear units (ReLU) activation function and $\boldsymbol{q}_n^{(j)}$ is recurrently defined as

$$\left[\boldsymbol{q}_n^{(j+1)}\right]_i = \sum_{k=1}^{H} w_{ik}^{(j+1)} g\left(\boldsymbol{q}_n^{(j)}\boldsymbol{w}_k^{(j)} + \alpha_k^{(j)}\right) + \alpha_i^{(j+1)}$$

  - $\boldsymbol{q}_n^{(0)}$ being the vector of input features $\mathcal{Q}_n$.

Representation learning in discrete choice modeling

L-MNL Model formulation

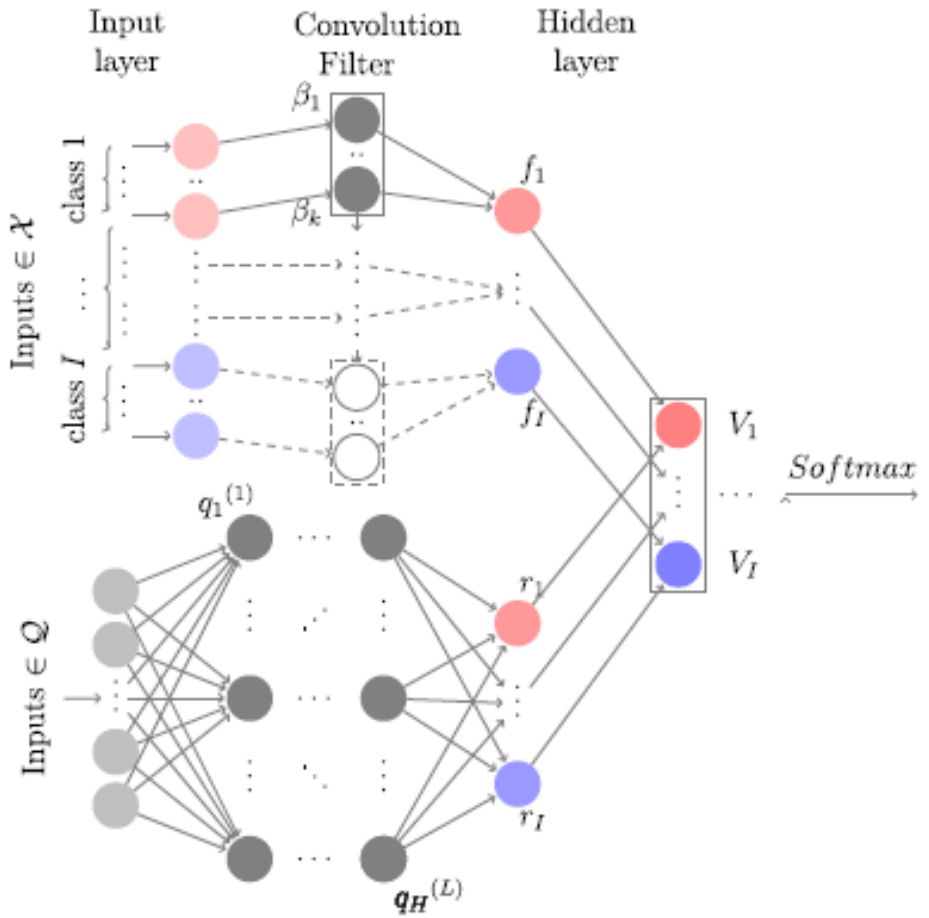

Fig. 2. L-MNL model architecture. On the top, we have the I class generalization of a linear-in-parameter MNL model, as depicted in Fig. 1. At the bottom, we have a deep neural network (*i.e.*, multilayer and fully connected) that enables us to obtain the representation learning term r_i. The terms from each part are added together defining the new systematic function of Eq. (11).

Outline

- Abstract
- Introduction
- Related work
- Implementing MNL as a neural network
- Representation learning in discrete choice modeling
- Experiments
- Future directions
- Conclusions

Experiments

- 実験では，L-MNLモデルが，解釈可能性を保ちながら，予測性能をどの程度向上させるかを実証する
- The experiments demonstrate how our L-MNL models increase the predictability of the MNL and NL models respectively while keeping their interpretability .

- Definition
  - Predictability: directly quantified in terms of likelihood 尤度により直接定量化される
  - Interpretability: the ability of the model to have both a quantifiable uncertainty of its parameters as well as an understandable or meaningful association with its variables
  - 解釈可能性: そのモデルが，①定量化可能なパラメータの不確実性(つまり標準偏差)と②理解可能な or 意味のある変数との関連性の両方を持つ能力

Experiments

Benchmarking models

Models	Description	Knowledge-Driven Part	Data-Driven Part	# of parameters and weights
Logit($\mathcal{X}$)	The standard MNL model with linear-in-parameter specification, and no learning component.	Input: $x \in \mathcal{X}$ Linear utility specification	$\mathcal{Q} = \emptyset$	Parameters: $\lvert\mathcal{X}\rvert$
DNN($n, \mathcal{Q}$)	A Dense Neural Network for every alternative with softmax loss. This is the NN-MNL model proposed by Hruschka et al. (2004).	$\mathcal{X} = \emptyset$	Input: $q \in \mathcal{Q}$ n neurons / hidden layer	Weights: $n(\lvert\mathcal{Q}\rvert + \lvert\mathcal{C}\rvert + 1)$
DNN_L($n, \mathcal{X} = \mathcal{Q}$)	A modified version of the NN-MNL model proposed by Hruschka et al. (2004) . Both of the linear utility specification and the neural network. The latter can be seen as finding the residual from the linear input of $\mathcal{X}$.	Input: $x \in \mathcal{X} (= \mathcal{Q})$	Input: $q \in \mathcal{Q} (= \mathcal{X})$	Weights: $(n+1)(\lvert\mathcal{Q}\rvert + \lvert\mathcal{C}\rvert)$
L-MNL($n, \mathcal{X}, \mathcal{Q}$)	Learning logit model with n neurons in the hidden layer ($H = n$). This model satisfies the interpretability condition for all x . we limit the complexity to a single layer ($L = 1$).	Input: Part of $x \in \mathcal{X}$ Linear utility specification	Input: $q \in \mathcal{Q}$	Parameters: $\lvert\mathcal{X}\rvert$ Weights: $n(\lvert\mathcal{Q}\rvert + \lvert\mathcal{C}\rvert)$

- All models have run for 200 epochs with an Adam optimizer (Kingma and Ba, 2014) running on default parameters from Keras python deep learning library (Chollet et al., 2015).
- Every model with a Neural Network has a 20% dropout regularizer following the DNN layer.

Experiments

Synthetic data experiments

- L-MNLモデルの性能を予測性能と推定精度の両面から分析するのに合成データを使用する.
- We use synthetic data to better analyze the performance of our L-MNL model in terms of both prediction perfor- mance and estimates accuracy.

- Outline
  - 合成データの生成方法　How we generated synthetic data
  - モンテカルロ実験で，すべてのベンチマークモデルをパラメータ推定に関して比較
  - We perform Monte Carlo experiments to compare all benchmarking models on parameters estimation.
  - ニューロン数の増加について，NNアーキテクチャの影響分析
  - A scan on increasing number of neurons is introduced to analyze the impact of the NN architecture better.
  - 入力集合$\mathcal{X}$,$\mathcal{Q}$の間に強い相関がある場合の分析
  - We investigate the case of strongly correlated variables between input sets $\mathcal{X}$ and $\mathcal{Q}$.

Experiments

Data generation

- 単純な二項選択モデルを考える．Guevara (2015) の生成プロセスに従って，以下の効用関数をもとにデータ生成する．
- Consider a simple binary choice model. Following the generation process of Guevara (2015), we generate data based on the following utility function.

$$U_{in} = V_{in} + \varepsilon_{in},$$

with

$$V_{in} = \underbrace{\beta_p \cdot p_{in} + \beta_a \cdot a_{in} + \beta_b \cdot b_{in}}_{\text{known relation}} + \underbrace{\beta_{qc} \cdot q_{in} c_{in}}_{\text{unknown interactions}}$$

$$p_{in} = 5 + z_{in} + 0.03 wz_{in} + \varepsilon_{pin}$$

$$q_{in} = 2h_{in} + k_{in} + \varepsilon_{qin}$$

$$k_{in} = h_{in} + \varepsilon_{kin} \quad \text{for } i = 1, 2$$

Parameters$(\beta_p = -1, \beta_a = 0.5, \beta_b = 0.5, \beta_{qc} = 1)$
Variables$\sim \mathcal{U}([-1, 1])$
Error terms following a uniform distribution

モデル化の段階で未知で未発見の因果関係を表現するための相互作用項
モデル作成者が効用を誤って設定するような状況をシミュレートする
an interaction term to represent unknown and undiscovered causalities during the modeling phase
simulate the situations where the modeler would misspecify the utility due to an undiscovered interaction

- 選択結果はベルヌーイ確率変数　The result of choice is a Bernoulli random variable

$$y_{1n} \sim \text{Bern}\big(P_{1n}(x_n)\big), \qquad y_{2n} = 1 - y_{1n} \qquad \forall n = 1, \ldots, N$$

Experiments

Monte Carlo experiment

- Training set: 1000 synthetic individual observations
- Test set: 200 synthetic individual observations
- *Logit($\mathcal{X}_{true}$) is a “True Model”
- Performing experiments 100 times.

- 予測性能の点で最も優れたモデルはNNベースの選択モデル
- The best models in terms of predictive performance are the neural network-based choice models
  - 訓練セットでの過学習にも関わらず，L-MNLモデルはデータの最も良い汎化表現を与え，テストセットで最高の予測性能を達成
  - Despite overfitting on the train set, our L-MNL gives the best general representation of the data, achieving the best predictive performance in the test set

Models	Knowledge-Driven Part Input	Data-Driven Part Input
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}\}$	$\mathcal{Q} = \{\boldsymbol{q}, \boldsymbol{c}\}$
Logit$(\mathcal{X}_1)$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	—
DNN$(25, \mathcal{Q})$	—	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
*Logit$(\mathcal{X}_{true})$	$\mathcal{X}_{true} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{qc}\}$	—

Table 1
Monte Carlo average log-likelihood ($\bar{LL}$) and standard deviation ($s.d.(LL)$) for the different models. Based on the test set value, we conclude that our L-MNL learns the best general representation. An MNL model with true utility specification is given as a reference. Average of accuracies.

Model	Train set		Test set		Accuracies [%]		ρ^2_{test}
	$\bar{LL}$	$s.d.(LL)$	$\bar{LL}$	$s.d.(LL)$	$\bar{Acc}_{train}$	$\bar{Acc}_{test}$	
Logit$(\mathcal{X}_{true})$	-459	15	-94	8	78	77	0.32
Logit$(\mathcal{X}_1)$	-604	11	-123	6	67	66	0.11
DNN$(25, \mathcal{Q})$	-363	21	-112	13	84	74	0.19
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	-367	18	-108	12	83	75	0.22
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	-429	16	**-97**	**8**	80	**76**	**0.30**

Experiments

Monte Carlo experiment

- パラメータの推定精度の観点からモデルを評価する
- evaluate the models in terms of accuracy in interpretable parameter estimation
  - 以下のように相対誤差$e_\beta, e_{\beta_i/\beta_j}$を定義
  - define the relative errors $e_\beta, e_{\beta_i/\beta_j}$ as:

$$e_\beta = \left|\frac{\beta - \hat{\beta}}{\beta}\right|$$

$$e_{\beta_i/\beta_j} = \left|\frac{e_{\beta_i} - e_{\beta_j}}{1 - e_\beta}\right|$$

- L-MNLは，真のモデルから1%以下しかズレのない相対誤差で，パラメータの真値を復元する能力においてすべてのモデルを大きく上回っている
- L-MNL greatly outperforms every model in the ability to recover the true parameter values with a relative error smaller than 1% away from the true model.
- MNLは，2番目に良い結果　MNL is the second best in accuracy.

Models	Knowledge-Driven Part Input	Data-Driven Part Input
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}\}$	$\mathcal{Q} = \{\boldsymbol{q}, \boldsymbol{c}\}$
Logit$(\mathcal{X}_1)$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	—
DNN$(25, \mathcal{Q})$	—	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
*Logit$(\mathcal{X}_{true})$	$\mathcal{X}_{true} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{qc}\}$	—

Table 2
Monte Carlo relative errors for the different models in [%] with $\bar{e}$ the average relative error, *s.d.* its standard deviation and β_p, β_a are from Eq. (21).

Model	$\bar{e}_{\beta_p}$	$s.d.(e_{\beta_p})$	$\bar{e}_{\beta_a}$	$s.d.(e_{\beta_a})$	$\bar{e}_{\beta_p/\beta_a}$	$s.d.(e_{\beta_p/\beta_a})$
Logit$(\mathcal{X}_{true})$	6.4	± 4.9	14.4	± 10.7	10.8	± 9.7
Logit$(\mathcal{X}_1)$	26.7	± 6.2	26.7	± 14.7	15.5	± 12.4
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	60	± 32.4	74	± 54	460	± 166
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	**7.1**	**± 5.1**	**15.2**	**± 11.7**	**11.3**	**± 10.3**

Experiments

Monte Carlo experiment

- パラメータの推定精度の観点からモデルを評価する
- evaluate the models in terms of accuracy in interpretable parameter estimation
  - 以下のように相対誤差$e_\beta, e_{\beta_i/\beta_j}$を定義
  - define the relative errors $e_\beta, e_{\beta_i/\beta_j}$ as:

$$e_\beta = \left|\frac{\beta - \hat{\beta}}{\beta}\right|$$

$$e_{\beta_i/\beta_j} = \left|\frac{e_{\beta_i} - e_{\beta_j}}{1 - e_\beta}\right|$$

- DNN_Lではパラメータ推定に大きな誤差
- DNN_L generate high errors in parameter estimates.
  - データ駆動項も部分的に$\boldsymbol{p}$や$\boldsymbol{a}$の線形依存性を学習
  - Data-Driven term is also partially learning linear dependencies of $\boldsymbol{p}$ or $\boldsymbol{a}$.

Models	Knowledge-Driven Part Input	Data-Driven Part Input
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}\}$	$\mathcal{Q} = \{\boldsymbol{q}, \boldsymbol{c}\}$
Logit$(\mathcal{X}_1)$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	—
DNN$(25, \mathcal{Q})$	—	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
*Logit$(\mathcal{X}_{true})$	$\mathcal{X}_{true} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{qc}\}$	—

Table 2
Monte Carlo relative errors for the different models in [%] with $\bar{e}$ the average relative error, s.d. its standard deviation and β_p, β_a are from Eq. (21).

Model	$\bar{e}_{\beta_p}$	$s.d.(e_{\beta_p})$	$\bar{e}_{\beta_a}$	$s.d.(e_{\beta_a})$	$\bar{e}_{\beta_p/\beta_a}$	$s.d.(e_{\beta_p/\beta_a})$
Logit$(\mathcal{X}_{true})$	6.4	± 4.9	14.4	± 10.7	10.8	± 9.7
Logit$(\mathcal{X}_1)$	26.7	± 6.2	26.7	± 14.7	15.5	± 12.4
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	60	± 32.4	74	± 54	460	± 166
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	**7.1**	**± 5.1**	**15.2**	**± 11.7**	**11.3**	**± 10.3**

Experiments

Monte Carlo experiment

- 推定パラメータと真のパラメータが統計的に異なるか検定
- We conduct hypothesis testing to determine whether the estimated parameters are statistically different from the true ones.
  - Null hypothesis H_0: $\hat{\beta} = \beta$
  - Alternative hypothsis H_1: $\hat{\beta} \neq \beta$
- L-MNLを除くすべてのモデルで，係数はほぼ常に真の係数と統計的に異なる
- The coefficients are almost always statistically different from the true ones for every model except L-MNL.
- DNN_Lでは，データ駆動項がパラメータ間の比率を損なう
- The data-driven term of DNN_L compromises the ratios between parameters.

Models	Knowledge-Driven Part Input	Data-Driven Part Input
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}\}$	$\mathcal{Q} = \{\boldsymbol{q}, \boldsymbol{c}\}$
Logit$(\mathcal{X}_1)$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	—
DNN$(25, \mathcal{Q})$	—	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$	$\mathcal{Q} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}, \boldsymbol{c}\}$
*Logit$(\mathcal{X}_{true})$	$\mathcal{X}_{true} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{qc}\}$	—

Table 3
Monte Carlo hypothesis testing for the different models, for β_p and β_a taken separately and for their ratio. Parameters β_p, β_a are from Eq. (21).

Model	% of experiments not rejecting H_0	
	β_p and β_a	β_p/β_a
Logit$(\mathcal{X}_{true})$	97.5	94
Logit$(\mathcal{X}_1)$	34	**97**
DNN_L$(25, \mathcal{X} = \mathcal{Q})$	25.5	37
L-MNL$(25, \mathcal{X}, \mathcal{Q})$	**95**	**94**

Experiments

Choice of neural network architecture

- NNのサイズ=層あたりのニューロン数と，尤度・パラメータ推定値の関係
- The relationship between the size of NN, the number of neurons per layer, and the likelihood and the values of parameter β.
- $n = 0{\sim}10$
  - Underfit: the NN has not yet captured all the non-linearities of the original utility specification. This can be seen by the higher values in likelihood.
- $n = 10{\sim}100$
  - Stable: The model performs almost as well as the true model, depicted by the boundary lines, leading us to believe that the NN component has successfully learned the non-linearities of the data.
- $n > 100$
  - Overfit: a drop in train likelihood and an increase in test likelihood.

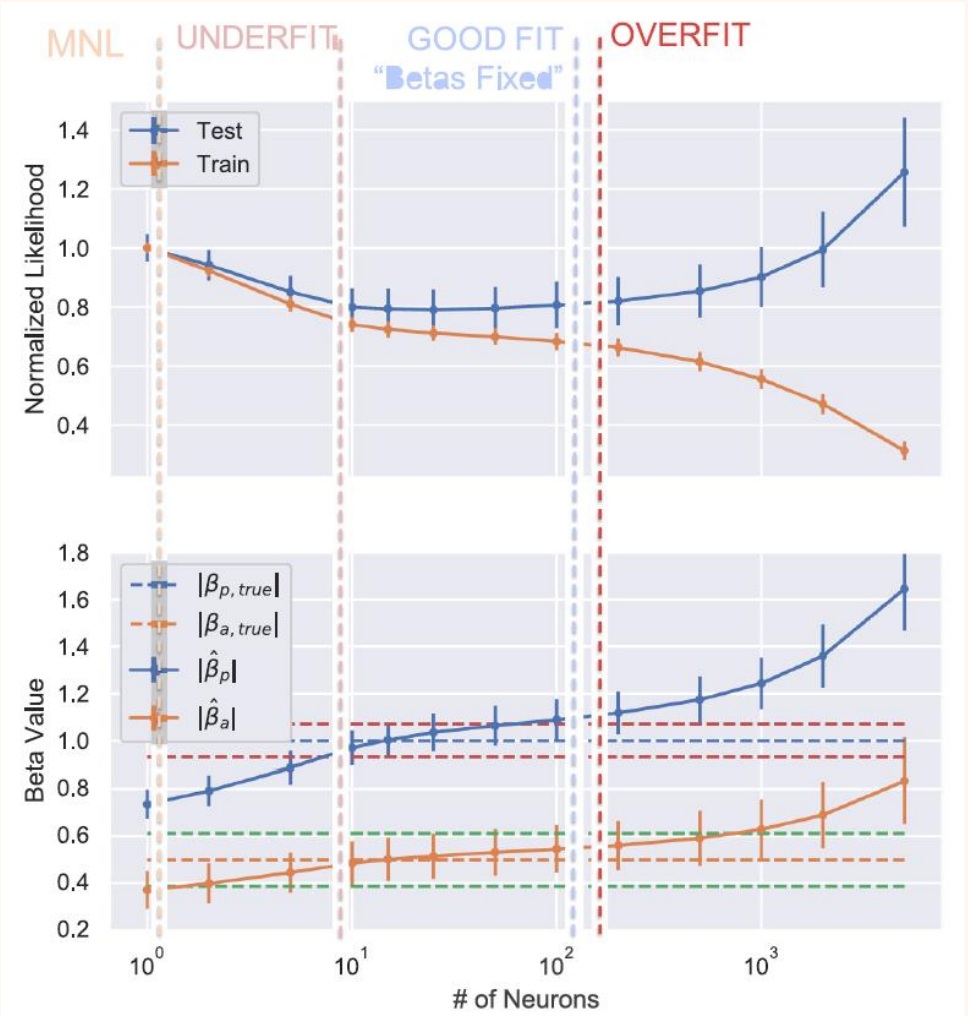

Fig. 3. Likelihood and values of parameter estimates for an increasing number of neurons in the hidden layer. Error bars show the standard deviation for 100 experiments. Red and Green lines show the standard deviation spread of the true model's parameter estimation. Best results are obtained with $n \in [10, 100]$.

Experiments

Choice of neural network architecture

- NNのサイズ=層あたりのニューロン数と，尤度・パラメータ推定値の関係
- The relationship between the size of NN, the number of neurons per layer, and the likelihood and the values of parameter β.
- $n = 0{\sim}10$
  - Underfit: the NN has not yet captured all the non-linearities of the original utility specification. This can be seen by the higher values in likelihood.
- $n = 10{\sim}100$
  - Stable: The model performs almost as well as the true model, depicted by the boundary lines, leading us to believe that the NN component has successfully learned the non-linearities of the data.
- $n > 100$
  - Overfit: a drop in train likelihood and an increase in test likelihood.

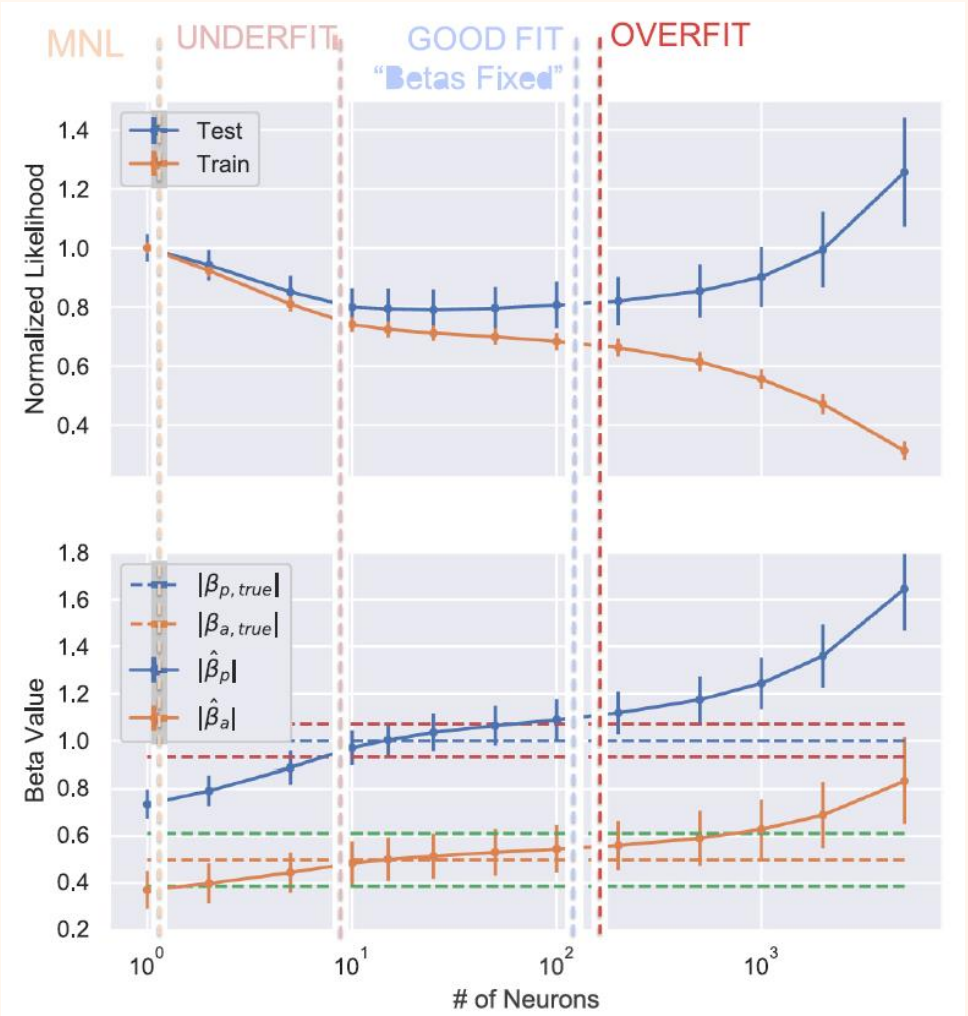

Fig. 3. Likelihood and values of parameter estimates for an increasing number of neurons in the hidden layer. Error bars show the standard deviation for 100 experiments. Red and Green lines show the standard deviation spread of the true model's parameter estimation. Best results are obtained with $n \in [10, 100]$.

Experiments

Models	Knowledge-Driven Part Input	Data-Driven Part Input
L-MNL$(100, \mathcal{X}, \mathcal{Q})$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}\}$	$\mathcal{Q} = \{\boldsymbol{q}', \boldsymbol{c}\}$
Logit$(\mathcal{X}_1)$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{q}', \boldsymbol{c}\}$	—
Logit$(\mathcal{X}_2)$	$\mathcal{X} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}\}$	—
*Logit$(\mathcal{X}_{true})$	$\mathcal{X}_{true} = \{\boldsymbol{p}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{qc}\}$	—

Impact of strongly correlated variables

- 相関のある変数の影響
  - Replace q_{in} with a new variable q'_{in} that is defined to be correlated to p_{in}.

$$q'_{in} = s \cdot p_{in} + \sqrt{1 - s^2} \cdot q_{in}$$

- bias with high variance due to the correlated variables (Logit$(\mathcal{X}_1)$, L-MNL)
- bias due to misspecification (Logit$(\mathcal{X}_1)$, Logit$(\mathcal{X}_2)$).
- s≤0.8以下では，L-MNLはロジットモデルよりも良い推定値を持つ
- For a correlation coefficient below $s \leq 0.8$ we can see the L-MNL has much better estimates with respect to the true model than the Logit models.

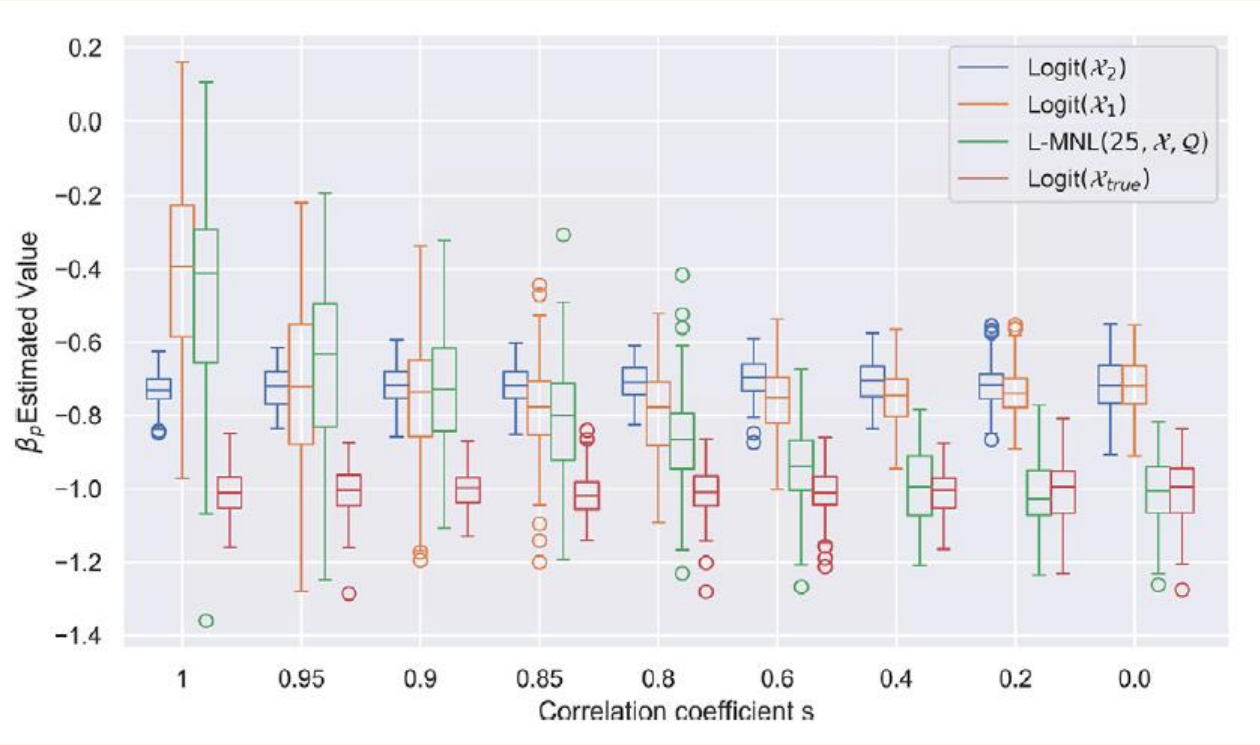

Fig. 4. Impact of correlated variables on parameter estimates, where $\boldsymbol{p} \in \mathcal{X}$ is correlated to $\boldsymbol{q}' \in \mathcal{Q}$ (see Eq. (30)). We see in this case that the L-MNL correlation bias is smaller than the bias due to underfit for all $s \leq 0.8$. For each coefficient, we performed 100 experiments.

- ニューラルネットワークの入力特徴量が知識駆動項の変数と同じでないか，強い相関がないか注意深くチェックする必要がある
- Modelers has to carefully check that the variables that enter as input features in the neural network are not the same or not too strongly correlated with the variables in the linear utility specification.

Experiments

A real case study: The swissmetro dataset (Bierlaire et al., 2001)

- Swissmetroデータセットは，1998年3月にスイスで収集された調査データで構成されている．回答者は，画期的な磁気浮上式高速鉄道システムであるSwissmetro(地下鉄ではなくスイス版リニア新幹線らしい，実現せず)に代表される，新しい革新的交通手段の影響を分析するための情報を提供した．(SP調査データということになる)
- The Swissmetro dataset consists of survey data collected in Switzerland during March 1998. The respondents provided information to analyze the impact of a new innovative transportation mode, represented by the Swissmetro, a revolutionary maglev underground system. This is SP survey data.
- Training set: 7234 observations, Test set: 1802 observations

Table 4
Swissmetro benchmark utility function, from Bierlaire et al. (2001).

Variables		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

Table 5
Variables in the Swissmetro dataset used for modeled component of the utility specification.

Variable	Description
TT	Door-to-door travel time in [minutes], scaled by 1/100.
Cost	Travel cost in [CHF], scaled by 1/100.
Freq	Transportation headway in [minutes]
GA	Binary variable indicating annual pass holders (=1).
Age	IntegeEs variable scaled with the traveler's age.
Luggage	Integer variable scaled with amount of luggage during travel.
Seats	Binary variable for special seats configuration in Swissmetro (=1).

Experiments

Model comparison

- 真のモデルは未知のため，　Bierlaire et al. (2001) のMNLの結果と比較
- Since the true values of parameter estimates are unknown, we compare the values obtained with our L-MNL to the ones obtained using the MNL model described in Bierlaire et al. (2001).
- 効用設定に表現学習の要素を加えると，対数尤度が有意に増加する．旅行者の選択を説明するのに役に立つ情報を含むことを示唆．
- Adding the representation learning component in the utility specification significantly increases the log-likelihood, suggesting that these variables contain information that helps to explain travelers' choice.

Table 6
Unused variables in the Swissmetro dataset.

Variable	Description
Purpose:	Integer variable indicating the trip purpose (business, leisure, etc.)
First :	Binary variable indicating if first class (=1) or not (=0)
Ticket:	Integer variable indicating the ticket type (one-way, half-day, etc.)
Who:	Integer variable indicating who is paying the ticket (self, employer, etc.)
Male:	Binary variable indicating the traveler's gender (0 = female, 1 = male)
Income:	Integer variable indicating the traveler's income per year.
Origin:	Integer variable indicating the canton in which the travel begins.
Dest:	Integer variable indicating the canton in which the travel ends.

Table 7
Comparison of log-likelihood and parameters estimates for different models with utility specification of Bierlaire et al. (2001). Number of observations = 7234.

Model	Parameters	Estimates	Std errors	t-stat	p-value
MNL $\rho^2_{test} = 0.28$ $\mathcal{L}(\hat{\beta}) = -5764$ $\mathcal{L}_{test}(\hat{\beta}) = -1433$	ASC_{Car}	1.08	0.162	6.67	0.00
	ASC_{SM}	1.05	0.153	6.84	0.00
	β_{age}	0.146	0.436	3.35	0.00
	β_{cost}	-0.695	0.0423	-16.42	0.00
	β_{freq}	-0.733	0.1132	-6.47	0.00
	β_{GA}	1.54	0.167	9.24	0.00
	$\beta_{luggage}$	-0.114	0.0488	-2.338	0.02
	β_{seats}	0.432	0.115	3.76	0.00
	β_{time}	-1.34	0.051	-26.18	0.00
L-MNL(100, $\mathcal{X}_1$, $\mathcal{Q}_1$) $\rho^2_{test} = 0.41$ $\mathcal{L}(\hat{\beta}) = \mathbf{-4511}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1181}$	ASC_{Car}	0.106	0.174	0.61	0.54
	ASC_{SM}	0.454	0.163	2.80	0.01
	β_{age}	0.390	0.045	8.63	0.00
	β_{cost}	-1.378	0.048	-28.45	0.00
	β_{freq}	-0.860	0.127	-6.77	0.00
	β_{GA}	0.214	0.194	1.10	0.27
	$\beta_{luggage}$	0.116	0.0529	2.19	0.03
	β_{seats}	0.104	0.109	0.95	0.34
	β_{time}	-1.563	0.056	-27.97	0.00
DNN_L(100,$\mathcal{X}_1 = \mathcal{Q}$) $\rho^2_{test} = 0.37$ $\mathcal{L}(\hat{\beta}) = -4964$ $\mathcal{L}_{test}(\hat{\beta}) = -1257$	ASC_{Car}	0.365	0.165	3.61	0.00
	ASC_{SM}	0.549	0.162	2.22	0.03
	β_{age}	0.087	0.0423	2.07	0.04
	β_{cost}	-0.897	0.046	-19.46	0.00
	β_{freq}	-0.639	0.123	-5.20	0.00
	β_{GA}	1.40	0.172	8.15	0.10
	$\beta_{luggage}$	0.186	0.0523	3.52	0.00
	β_{seats}	0.233	0.102	2.29	0.02
	β_{time}	-1.146	0.049	-23.32	0.00
Logit($\mathcal{X}_{dum}$) (all 41 inputs) $\rho^2_{test} = 0.33$ $\mathcal{L}(\hat{\beta}) = -5451$ $\mathcal{L}_{test}(\hat{\beta}) = -1322$	β_{cost}	-1.062	0.059	18	0.00
	β_{freq}	-0.79	0.118	6.69	0.00
	β_{time}	-1.326	0.053	25.02	0.00
	...	...	...	...	...
L-MNL(100, $\mathcal{X}_2$, $\mathcal{Q}_2$) $\rho^2_{test} = 0.44$ $\mathcal{L}(\hat{\beta}) = \mathbf{-3895}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1108}$	β_{cost}	-1.671	0.0523	-31.94	0.00
	β_{freq}	-0.865	0.0765	-11.30	0.00
	β_{time}	-1.769	0.0389	-45.4	0.00

the result obtained using the MNL model described in Bierlaire et al. (2001)

L-MNL model including all variables

Experiments

Model comparison

- DNN_L($100, \mathcal{X}_1 = \mathcal{Q}$)は、両方の入力集合で変数間に1対1の相関があるが、パラメータの有意性を失っていない.
- DNN_L($100, \mathcal{X}_1 = \mathcal{Q}$) has one-to-one correlation among variables in both sets and does not loose significance in its parameters.
- L-MNLは，どの変数と相互作用が良くデータに相関しているかNNが学習することで，推定パラメータの有意性を失った
- The coefficients in L-MNL have lost their significance due to the neural network's ability to learn better which variables and interactions are most correlated to the data.

Table 7
Comparison of log-likelihood and parameters estimates for different models with utility specification of Bierlaire et al. (2001). Number of observations = 7234.

Model	Parameters	Estimates	Std errors	t-stat	p-value
MNL $\rho^2_{test} = 0.28$ $\mathcal{L}(\hat{\beta}) = -5764$ $\mathcal{L}_{test}(\hat{\beta}) = -1433$	ASC_{Car}	1.08	0.162	6.67	0.00
	ASC_{SM}	1.05	0.153	6.84	0.00
	β_{age}	0.146	0.436	3.35	0.00
	β_{cost}	-0.695	0.0423	-16.42	0.00
	β_{freq}	-0.733	0.1132	-6.47	0.00
	β_{GA}	1.54	0.167	9.24	0.00
	$\beta_{luggage}$	-0.114	0.0488	-2.338	0.02
	β_{seats}	0.432	0.115	3.76	0.00
	β_{time}	-1.34	0.051	-26.18	0.00
L-MNL$(100, \mathcal{X}_1, \mathcal{Q}_1)$ $\rho^2_{test} = 0.41$ $\mathcal{L}(\hat{\beta}) = \mathbf{-4511}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1181}$	ASC_{Car}	0.106	0.174	0.61	0.54
	ASC_{SM}	0.454	0.163	2.80	0.01
	β_{age}	0.390	0.045	8.63	0.00
	β_{cost}	-1.378	0.048	-28.45	0.00
	β_{freq}	-0.860	0.127	-6.77	0.00
	β_{GA}	0.214	0.194	1.10	0.27
	$\beta_{luggage}$	0.116	0.0529	2.19	0.03
	β_{seats}	0.104	0.109	0.95	0.34
	β_{time}	-1.563	0.056	-27.97	0.00
DNN_L$(100, \mathcal{X}_1 = \mathcal{Q})$ $\rho^2_{test} = 0.37$ $\mathcal{L}(\hat{\beta}) = -4964$ $\mathcal{L}_{test}(\hat{\beta}) = -1257$	ASC_{Car}	0.365	0.165	3.61	0.00
	ASC_{SM}	0.549	0.162	2.22	0.03
	β_{age}	0.087	0.0423	2.07	0.04
	β_{cost}	-0.897	0.046	-19.46	0.00
	β_{freq}	-0.639	0.123	-5.20	0.00
	β_{GA}	1.40	0.172	8.15	0.10
	$\beta_{luggage}$	0.186	0.0523	3.52	0.00
	β_{seats}	0.233	0.102	2.29	0.02
	β_{time}	-1.146	0.049	-23.37	0.00
Logit$(\mathcal{X}_{dum})$ (all 41 inputs) $\rho^2_{test} = 0.33$ $\mathcal{L}(\hat{\beta}) = -5451$ $\mathcal{L}_{test}(\hat{\beta}) = -1322$	β_{cost}	-1.062	0.059	18	0.00
	β_{freq}	-0.79	0.118	6.69	0.00
	β_{time}	-1.326	0.053	25.02	0.00
	...	...	...	...	...
L-MNL$(100, \mathcal{X}_2, \mathcal{Q}_2)$ $\rho^2_{test} = 0.44$ $\mathcal{L}(\hat{\beta}) = \mathbf{-3895}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1108}$	β_{cost}	-1.671	0.0523	-31.94	0.00
	β_{freq}	-0.865	0.0765	-11.30	0.00
	β_{time}	-1.769	0.0389	-45.4	0.00

L-MNL model including all variables

DNN_L

Experiments

Model comparison

- 元のMNLのパラメータの有意性はアンダーフィットによる可能性
- The significance of the same parameters in the initial MNL model can originate from a bias due to the model's underfit.
- 多くの説明変数が省略されているため、モデルが内生性、すなわち従属変数と誤差項の間の相関を受けやすくなっている
- With many explanatory variables being omitted in the initial MNL model, it is worth noting that the model is more likely to be subject to endogeneity, i.e., correlation among the dependent variables and the error term.

Table 7
Comparison of log-likelihood and parameters estimates for different models with utility specification of Bierlaire et al. (2001). Number of observations = 7234.

Model	Parameters	Estimates	Std errors	t-stat	p-value
MNL $\rho^2_{test} = 0.28$ $\mathcal{L}(\hat{\beta}) = -5764$ $\mathcal{L}_{test}(\hat{\beta}) = -1433$	ASC_{Car}	1.08	0.162	6.67	0.00
	ASC_{SM}	1.05	0.153	6.84	0.00
	β_{age}	0.146	0.436	3.35	0.00
	β_{cost}	-0.695	0.0423	-16.42	0.00
	β_{freq}	-0.733	0.1132	-6.47	0.00
	β_{GA}	1.54	0.167	9.24	0.00
	$\beta_{luggage}$	-0.114	0.0488	-2.338	0.02
	β_{seats}	0.432	0.115	3.76	0.00
	β_{time}	-1.34	0.051	-26.18	0.00
L-MNL$(100, \mathcal{X}_1, \mathcal{Q}_1)$ $\rho^2_{test} = 0.41$ $\mathcal{L}(\hat{\beta}) = \mathbf{-4511}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1181}$	ASC_{Car}	0.106	0.174	0.61	0.54
	ASC_{SM}	0.454	0.163	2.80	0.01
	β_{age}	0.390	0.045	8.63	0.00
	β_{cost}	-1.378	0.048	-28.45	0.00
	β_{freq}	-0.860	0.127	-6.77	0.00
	β_{GA}	0.214	0.194	1.10	0.27
	$\beta_{luggage}$	0.116	0.0529	2.19	0.03
	β_{seats}	0.104	0.109	0.95	0.34
	β_{time}	-1.563	0.056	-27.97	0.00
DNN_L$(100, \mathcal{X}_1 = \mathcal{Q})$ $\rho^2_{test} = 0.37$ $\mathcal{L}(\hat{\beta}) = -4964$ $\mathcal{L}_{test}(\hat{\beta}) = -1257$	ASC_{Car}	0.365	0.165	3.61	0.00
	ASC_{SM}	0.549	0.162	2.22	0.03
	β_{age}	0.087	0.0423	2.07	0.04
	β_{cost}	-0.897	0.046	-19.46	0.00
	β_{freq}	-0.639	0.123	-5.20	0.00
	β_{GA}	1.40	0.172	8.15	0.10
	$\beta_{luggage}$	0.186	0.0523	3.52	0.00
	β_{seats}	0.233	0.102	2.29	0.02
	β_{time}	-1.146	0.049	-23.32	0.00
Logit$(\mathcal{X}_{dum})$ (all 41 inputs) $\rho^2_{test} = 0.33$ $\mathcal{L}(\hat{\beta}) = -5451$ $\mathcal{L}_{test}(\hat{\beta}) = -1322$	β_{cost}	-1.062	0.059	18	0.00
	β_{freq}	-0.79	0.118	6.69	0.00
	β_{time}	-1.326	0.053	25.02	0.00
	...	...	...	...	...
L-MNL$(100, \mathcal{X}_2, \mathcal{Q}_2)$ $\rho^2_{test} = 0.44$ $\mathcal{L}(\hat{\beta}) = \mathbf{-3895}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1108}$	β_{cost}	-1.671	0.0523	-31.94	0.00
	β_{freq}	-0.865	0.0765	-11.30	0.00
	β_{time}	-1.769	0.0389	-45.4	0.00

the initial MNL model

Experiments

Model comparison

- ダミー符号化多項ロジットモデル：$\mathcal{X}_{dum} = \mathcal{X}_2 \cup \mathcal{Q}_2$
- L-MNLモデルとの最終的な尤度の大きな差は，このデータセットが，モデル作成者の捉えることが困難な複雑な関数や変数間の相互作用を実際に含んでいることを示唆
- The big difference in final likelihoods with our L-MNL models further demonstrate that this dataset indeed contains complex functions and interactions among its variables which may be difficult to capture by the modeler.

Table 7
Comparison of log-likelihood and parameters estimates for different models with utility specification of Bierlaire et al. (2001). Number of observations = 7234.

Model	Parameters	Estimates	Std errors	t-stat	p-value
MNL $\rho^2_{test} = 0.28$ $\mathcal{L}(\hat{\beta}) = -5764$ $\mathcal{L}_{test}(\hat{\beta}) = -1433$	ASC_{Car}	1.08	0.162	6.67	0.00
	ASC_{SM}	1.05	0.153	6.84	0.00
	β_{age}	0.146	0.436	3.35	0.00
	β_{cost}	-0.695	0.0423	-16.42	0.00
	β_{freq}	-0.733	0.1132	-6.47	0.00
	β_{GA}	1.54	0.167	9.24	0.00
	$\beta_{luggage}$	-0.114	0.0488	-2.338	0.02
	β_{seats}	0.432	0.115	3.76	0.00
	β_{time}	-1.34	0.051	-26.18	0.00
L-MNL$(100, \mathcal{X}_1, \mathcal{Q}_1)$ $\rho^2_{test} = 0.41$ $\mathcal{L}(\hat{\beta}) = \mathbf{-4511}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1181}$	ASC_{Car}	0.106	0.174	0.61	0.54
	ASC_{SM}	0.454	0.163	2.80	0.01
	β_{age}	0.390	0.045	8.63	0.00
	β_{cost}	-1.378	0.048	-28.45	0.00
	β_{freq}	-0.860	0.127	-6.77	0.00
	β_{GA}	0.214	0.194	1.10	0.27
	$\beta_{luggage}$	0.116	0.0529	2.19	0.03
	β_{seats}	0.104	0.109	0.95	0.34
	β_{time}	-1.563	0.056	-27.97	0.00
DNN_L$(100, \mathcal{X}_1 = \mathcal{Q})$ $\rho^2_{test} = 0.37$ $\mathcal{L}(\hat{\beta}) = -4964$ $\mathcal{L}_{test}(\hat{\beta}) = -1257$	ASC_{Car}	0.365	0.165	3.61	0.00
	ASC_{SM}	0.549	0.162	2.22	0.03
	β_{age}	0.087	0.0423	2.07	0.04
	β_{cost}	-0.897	0.046	-19.46	0.00
	β_{freq}	-0.639	0.123	-5.20	0.00
	β_{GA}	1.40	0.172	8.15	0.10
	$\beta_{luggage}$	0.186	0.0523	3.52	0.00
	β_{seats}	0.233	0.102	2.29	0.02
	β_{time}	-1.146	0.049	-23.32	0.00
Logit$(\mathcal{X}_{dum})$ (all 41 inputs) $\rho^2_{test} = 0.33$ $\mathcal{L}(\hat{\beta}) = -5451$ $\mathcal{L}_{test}(\hat{\beta}) = -1322$	β_{cost}	-1.062	0.059	18	0.00
	β_{freq}	-0.79	0.118	6.69	0.00
	β_{time}	-1.326	0.053	25.02	0.00
	...	...	...	...	...
L-MNL$(100, \mathcal{X}_2, \mathcal{Q}_2)$ $\rho^2_{test} = 0.44$ $\mathcal{L}(\hat{\beta}) = \mathbf{-3895}$ $\mathcal{L}_{test}(\hat{\beta}) = \mathbf{-1108}$	β_{cost}	-1.671	0.0523	-31.94	0.00
	β_{freq}	-0.865	0.0765	-11.30	0.00
	β_{time}	-1.769	0.0389	-45.4	0.00

L-MNL model including all variables

a dummy coded Multinomial Logit

L-MNL model including only time, cost and frequency in $\mathcal{X}_2$

Experiments

Model comparison

- 各モデルのVOTとVOFの比較　Comparison of VOT and VOF for the different models
- L-MNLとLogit($\mathcal{X}_1$)の対数尤度の差は，MNLモデルがアンダーフィットに直面し，有意な比率の不一致を示唆
- The difference in log-likelihood values between the L-MNL and the Logit($\mathcal{X}_1$) suggests that the MNL model suffers from underfitting that leads to significant ratio discrepancy.
- ➢ データセットに強い非線形性がある　Only under strong non-linearities in the dataset

- ➢ どちらの追加したDCM手法も，尤度が高くなるにつれて，L-MNLの比率に近づいている．この傾向は，効用の仕様が改善されれば，我々の新しい選択モデルによって，比率と尤度がより似た値になることを示唆
- ➢ Both added DCM methods, that their ratio values move towards those of L-MNL as their likelihood increases. This trend implies a better specification in their utility would allow them to reach even more similar values in ratios and likelihood with our new choice model.

Table 8
Parameter ratio comparison.

Model	Value of Time	Value of Frequency	Train Log-Likelihood	Test Log-Likelihood
Logit($\mathcal{X}_1$)	0.52	0.95	-5764	-1433
Logit($\mathcal{X}_{dum}$)	0.80	1.34	-5451	-1322
DNN_L($\mathcal{X}_1$)	0.78	1.40	-4964	-1257
L-MNL($\mathcal{X}_1$)	0.88	1.60	-4511	-1181
L-MNL($\mathcal{X}_2$)	0.94	1.93	**−3895**	**−1108**
CNL($\mathcal{X}_1$)	0.59	1.52	-5711	-1415
TPM($\mathcal{X}_1$)	0.72	1.59	-4752	-1350

Experiments

Model comparison

- VOTとVOFの安定した比率が，10～200ニューロンを持つニューラルネットワークで得られる
- Stable ratios for VOT and VOF are obtained for a neural network having from 10 to 200 neurons.
- MNLモデルのアンダーフィット
- We see the MNL ratios for $n = 0$ neuron highlights the underfit of the MNL model.
- オーバーフィットは，500ニューロンから観測され，テスト尤度はもはや改善されない
- An overfit effect is observed starting from 500 neurons, where the test likelihood no longer improves.

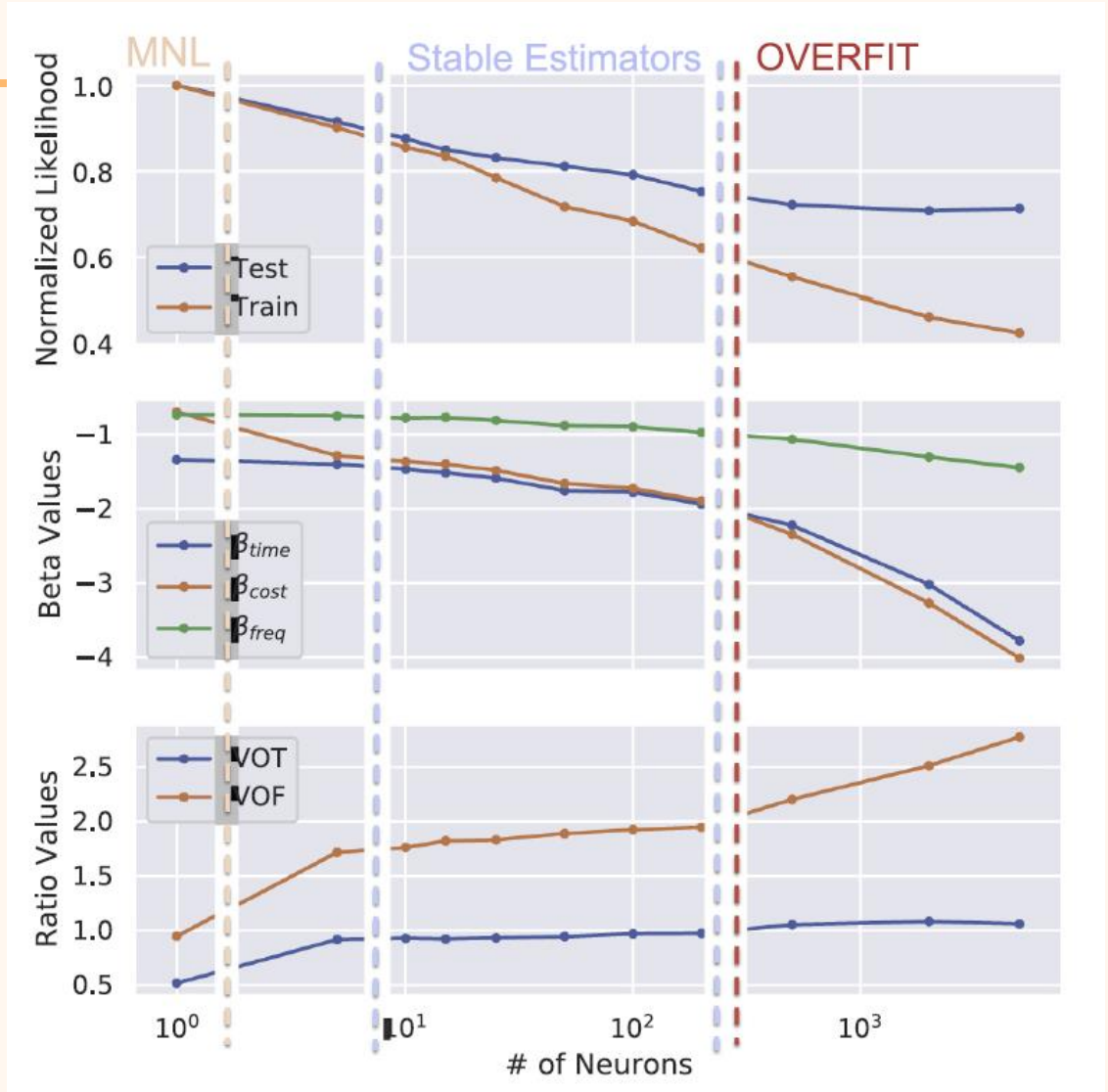

Fig. 5. Scan of Likelihood and Beta values over number of neurons n in the densely connected layer for L-MNL(100,$\mathcal{X}_2$, $\mathcal{Q}_2$). For $n \in [10, 200]$ we have VOT ≈ 1 and VOF ≈ 1.9. Over $n = 200$, we have signs of overfit.

Experiments

Revealed preference study: Optima dataset

- 収集したデータのサイズの制約により，表現学習がより困難なケースを分析
- We investigate a case much more difficult for representa- tion learning due to the size limitation of gathered data.
- Optima datasetは，2009年から2010年にかけてスイスで収集された小規模な調査であり，回答者は，移動手段の選択というテーマについて，移動にかかる時間や費用，社会経済属性など，広範な情報を回答
- The project named Optima is a small survey collected in Switzerland between 2009 and 2010 where respondents filled up extensive information in the topic of mode choice, including time and cost of performed trips, socio-economic characteristics.

- Benchmark models
  - L-MNL model
  - A well-specified base multinomial logit model (Fernández-Antolín et al., 2016)
  - The Integrated Choice and Latent Variable (ICLV) model (Fernández-Antolín et al., 2016)

Experiments

Models description

- The MNL specification from Fernández-Antolín et al. (2016) can be seen in Table 10.
- Compare the following models:
  - L-MNL: $\mathcal{X}_1$ is the same as the MNL. $\mathcal{Q}_1$ includes variables in Table 11.
  - L-MNL2: $\mathcal{X}_2 = \{time, cost, distance\}$. $\mathcal{Q}_2$ includes other variables.
  - NN: $\mathcal{X} = \mathcal{Q}$. 100 neurons per layer.
  - NN2: $\mathcal{X} = \mathcal{Q}$. 30 neurons per layer.
- Training set: 1089 answers, Test set: 287 answers
- ➢ Too small dataset for NN. Difficult to avoid overfitting
- ➢ Added a 30% dropout layer and an L2 regularizer of weight $\lambda = 0.5$.

Table 10
Base model specification from Fernández-Antolín et al. (2016).

Variables		Alternative		
		Public Transportation	Car	Slow modes
ASC	Constant	PT-Const	CAR-Const	
TT	Travel Time [min]	B-Time-PT	B-Time-CAR	
MCost	$\frac{\text{Marginal Cost}}{\text{Income}}$	B-MCost-PT	B-MCost-CAR	
Distance	Trip distance [km]			B-Dist
Work	Work related Trip		B-Work	
French	French Speaking area		B-French	
Student	Occupation is student	B-Student		
Urban	Urban area	B-Urban		
NbChild	Number of Children		B-NbChild	
NbCar	Number of Cars		B-NbCar	
NbBicy	Number of Bicycles			B-NbBicy

Table 11
Added variables for representation learning term in Optima dataset. All variables gave -1 for missing values.

Variable	Description
Age:	Age of the respondent (in years)
HouseType :	1 is individual house (or terraced house), 2 is apartment, 3 is independent room
Gender:	1 is man, 2 is woman, -1 for missing value.
Education:	Highest education achieved[a]. Categories from 1 to 8.
FamilSitu:	Family situation[a]. Categories from 1 to 7.
ScaledIncome:	Integer variable indicating the traveler's income per year.
OwnHouse:	Do you own the place where you are living? 1 is yes, 2 is no
MotherTongue:	1 for german or swiss german, 2 for french, 3 for other.
SocioProfCat:	Socio-professional[a] categories from 1 to 8.

[a] More details at https://biogeme.epfl.ch/data.html.

Experiments

Expert modeling as a regularizer

- 前述の5つのモデルの性能
- The performances of the five models described above
- Run 100 times for every model
- 強力な正則化と学習サイクル 80 epochsより，ばらつきは初期値の変化による
- The observed variation comes from the change in starting values, which ultimately brings to a slightly different optimum given the strong regularizers and fixed number of training cycles 80 epochs.

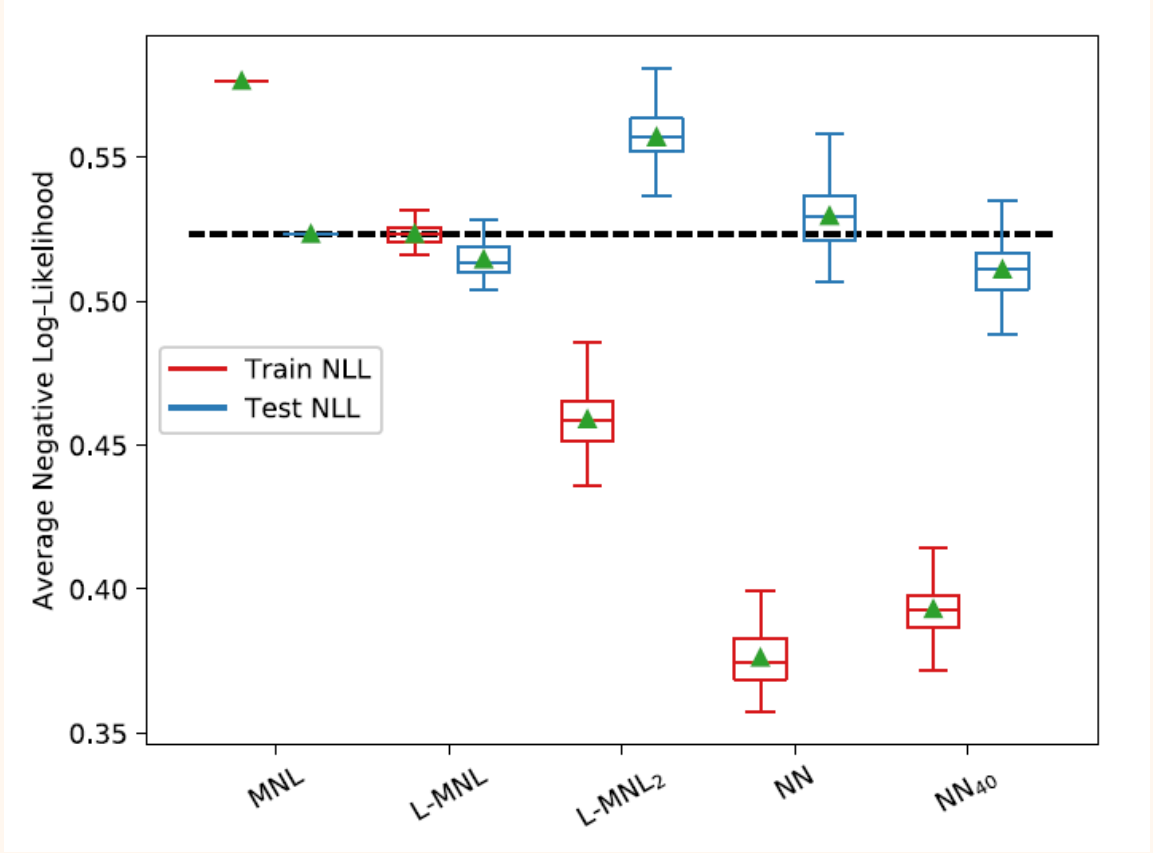

Fig. 6. Performance of multiple models on a small revealed preference dataset (Optima) for 100 minimization iterations. The first two models have expert modeling of the utility specification and generalize well. The last three have small or no modeling of the utility and show clear signs of overfitting as well as high variance in performance results.

Experiments

Expert modeling as a regularizer

- 前述の5つのモデルの性能
- The performances of the five models described above
- Swissmetro datasetとは対照的な結果
- As opposed to the case seen in Swissmetro:
  - L-MNL: ○, L-MNL2: ×
  - NN: △, NN2: △

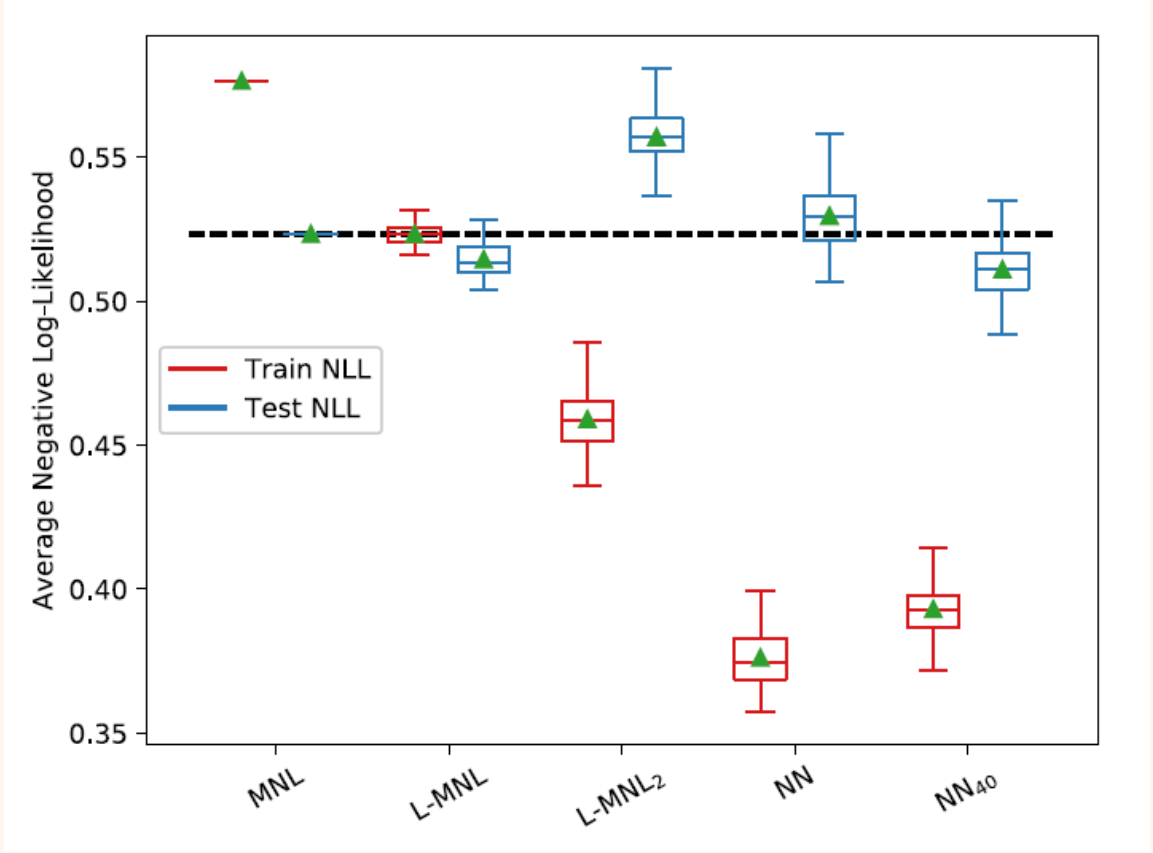

Fig. 6. Performance of multiple models on a small revealed preference dataset (Optima) for 100 minimization iterations. The first two models have expert modeling of the utility specification and generalize well. The last three have small or no modeling of the utility and show clear signs of overfitting as well as high variance in performance results.

Experiments

Benchmarking with ICLV

- 潜在変数$CarLoving$を追加したICLVモデルとの比較　Compare an ICLV model, where the latent variable CarLoving is added
- Two Likert indicators (Likert, 1932) and the following specifications
  - I_1：It is difficult to take the public transport when I travel with my children.
  - I_2：With my car I can go wherever and whenever.

$$t_{car} \cdot I_i = \alpha_i + \lambda_i \cdot Carloving \cdot t_{car} + \omega_i$$

  - $\omega_i \sim \mathcal{N}(0, \sigma_i)$ is the random term, $\alpha_i, \lambda_i, \sigma_i$ are parameters to be estimated

$$CarLoving = \eta_{car} + \omega$$

  - η_{car} and σ are the estimated parameters and $\omega \sim \mathcal{N}(0, \sigma)$ is the random term

Table 12
Models accuracy on training and testing sets of Optima.

Model	Logit($\mathcal{X}_1$)	L-MNL($\mathcal{X}_1$,$\mathcal{Q}_1$)	$NN_{40}(\mathcal{X} \cup \mathcal{Q})$	ICLV($\mathcal{X}_1$)
Accuracy Train [%]	76.8	80.4	86.1	80.0
Accuracy Test [%]	76.7	79.2	81.3	77.7

Experiments

Benchmarking with ICLV

- 潜在変数$CarLoving$を追加したICLVモデルとの比較　Compare an ICLV model, where the latent variable CarLoving is added
- Two Likert indicators (Likert, 1932) and the following specifications
  - I_1：It is difficult to take the public transport when I travel with my children.
  - I_2：With my car I can go wherever and whenever.

$$t_{car} \cdot I_i = \alpha_i + \lambda_i \cdot Carloving \cdot t_{car} + \omega_i$$

  - $\omega_i \sim \mathcal{N}(0, \sigma_i)$ is the random term, $\alpha_i, \lambda_i, \sigma_i$ are parameters to be estimated

$$CarLoving = \eta_{car} + \omega$$

  - η_{car} and σ are the estimated parameters and $\omega \sim \mathcal{N}(0, \sigma)$ is the random term

Table 12
Models accuracy on training and testing sets of Optima.

Model	Logit($\mathcal{X}_1$)	L-MNL($\mathcal{X}_1$,$\mathcal{Q}_1$)	$NN_{40}(\mathcal{X} \cup \mathcal{Q})$	ICLV($\mathcal{X}_1$)
Accuracy Train [%]	76.8	80.4	86.1	80.0
Accuracy Test [%]	76.7	79.2	81.3	77.7

Experiments

Benchmarking with ICLV

- 潜在変数$CarLoving$を追加したICLVモデルとの比較　Compare an ICLV model, where the latent variable CarLoving is added
  - ICLVモデルは，同じ特徴空間を持ちながらもMNLより良い結果
  - The ICLV model performs better than MNL while having the same initial feature space.
  - 潜在変数や，より複雑な構造方程式は追加されると，より高い精度が期待できる
  - A stronger increase in accuracy could be expected with added latent variables or more complex structural equations.

Table 12
Models accuracy on training and testing sets of Optima.

Model	Logit($\mathcal{X}_1$)	L-MNL($\mathcal{X}_1$,$\mathcal{Q}_1$)	$NN_{40}(\mathcal{X} \cup \mathcal{Q})$	ICLV($\mathcal{X}_1$)
Accuracy Train [%]	76.8	80.4	86.1	80.0
Accuracy Test [%]	76.7	79.2	81.3	77.7

Experiments

Benchmarking with ICLV

- 潜在変数$CarLoving$を追加したICLVモデルとの比較　Compare an ICLV model, where the latent variable CarLoving is added
  - NNはL-MNLよりも良い性能を示すが，training setにオーバーフィット
  - The NN performs better than the L-MNL on average while slightly overfitting the training set.
  - 解釈可能なパラメータを含まず，性能のばらつきが大きいので，fine-tuningが大変
  - The obtained model does not contain straightforward interpretable parameters, has higher variance in performance, and required more efforts in fine-tuning for optimal performance.
  - ➢ 離散選択モデルでは，うまくモデル化されたL-MNLとICLVモデルが最も有用
  - ➢ In the DCM sense, the most useful models would be L-MNL with the full expert specification and the ICLV model.

Table 12
Models accuracy on training and testing sets of Optima.

Model	Logit($\mathcal{X}_1$)	L-MNL($\mathcal{X}_1$,$\mathcal{Q}_1$)	$NN_{40}(\mathcal{X} \cup \mathcal{Q})$	ICLV($\mathcal{X}_1$)
Accuracy Train [%]	76.8	80.4	86.1	80.0
Accuracy Test [%]	76.7	79.2	81.3	77.7

Outline

- Abstract
- Introduction
- Related work
- Implementing MNL as a neural network
- Representation learning in discrete choice modeling
- Experiments
- Future directions
- Conclusions

Future directions

The new possibility to investigate many more types of datasets for discrete choice modeling

- 新しい入力を標準的な標準的な離散選択モデルの変数と共存させることを提案
- Propose the coexistence of these new inputs with standard discrete choice modeling variables
  - 表現学習は，連続信号，画像，時系列データなどあらゆる種類の入力に対応できる
  - Representation learning methods exist for all types of inputs, including continuous signals, images, time series, and more.
  - 離散選択モデルの入力変数の幅を広げることができるのではないか

Other discrete choice models such as more advanced GEV models, Mixed Logit or Latent Class models can also benefit from an added data-driven term while keeping high degree of interpretability

- 統一されたフレームワーク/ライブラリに統合
- Integrating them in a unified framework/library
  - MNL: NNを利用してデータ駆動項を実装　Using NN, implement data-driven term
  - NL: 深層学習ライブラリに複数のカスタム損失層の実装による　Thanks to the implementation of multiple custom loss layers in a deep learning library, we have implemented its first nested generalization

Future directions

The proposed architecture of our model may also help in the task of modeling the knowledge-driven term of the utility specification via feature selection

- データ駆動項は何を学習したかを理解して，知識駆動項の設定に役立てる
- Understanding what a data-driven term has learned, make use of modeling knowledge-driven term

The structure and role of the representation term in the utility function

- 効用ごとの複数の項を持つことで，選択された各入力集合$\mathcal{Q}_i$がそれぞれのネットワークに属し，効用に意味のある埋め込みを作成できる
- One could have multiple terms per utility, such that each chosen input set $\mathcal{Q}_i$, belonging to their own respective network, would create a meaningful embedding in the utility.

Tackle small datasets

- 本研究のアーキテクチャを小規模なデータセットに適用した場合，一般的な深層学習手法の正則化ツールとして機能する可能性がある
- Our architecture may perform as a regularization tool for common deep learning methods when applied to small datasets

Outline

- Abstract
- Introduction
- Related work
- Implementing MNL as a neural network
- Representation learning in discrete choice modeling
- Experiments
- Future directions
- Conclusions

Conclusion

- 離散選択モデルの効用仕様に表現技術技術を統合し，利用可能なデータから自動的に良い効用仕様を発見する一般的で柔軟な新しい理論的枠組みを導入
- We introduced a novel general and flexible theoretical framework that integrates a representation learning technique into the utility specification of a discrete choice model to automatically discover good utility specification from available data.
- 解釈可能性を維持しながら，データ駆動項によりモデルの全体的な予測性能を大幅に改善
- While keeping interpretability, data-driven term may account for many forms of misspecifications and greatly improves the overall predictability of the model.
- L-MNLの定式化，および有効性を実証
- Formulating L-MNL and demonstrating the effectiveness of our framework
  - 予測性能とパラメータ推定の精度の両方において，従来の選択モデルや既存のハイブリッドモデルを凌駕する
  - Our models outperformed the traditional choice models and existing hybrid models, both in terms of predictive performance and accuracy in parameter estimation.

所感

- 離散選択モデルと機械学習の関連性・繋がりがすこし見えてきた
  - ロジット型選択確率式とsoftmax関数の形が同じ
  - MNLをCNNで実装する部分で，離散選択モデルの最尤推定と機械学習の損失最小化が同値であることが見えて良かった
- 人の手ではうまく設定できない部分を機械学習に頼ることで，精度を改善している点が良い
  - 通常，変数で説明できない効用の効果は定数項などで表現される
  - L-MNLでは定数項がないので，データ駆動項に取り込んで精度を改善しているのだなと
- 数値実験の節がかなり分量が多く大変だった
  - スライド作成しながら理解は深められたと思う
- 今後の展望で，離散選択モデルの入力の多様化ができそうなのが面白そう
  - 目の前に見えているもの(混雑の程度とか天気とか)を変数として入力することで，行動をもっとよく表現できないか
  - 自動車運転時にどんな行動をとるかという話にも応用できそう