ResLogit: A residual neural network logit model

Melvin Wong[a,*], Bilal Farooq[a]

[a] *Laboratory of Innovations in Transportation, Civil Engineering Department, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada*

Abstract

We present a Residual Logit (ResLogit) model for seamlessly integrating a data-driven Deep Neural Network (DNN) architecture in the random utility maximization paradigm. DNN models such as the Multi-layer Perceptron (MLP) have shown remarkable success in modelling complex data accurately, but recent studies have consistently demonstrated that their black-box properties are incompatible with discrete choice analysis for the purpose of interpreting decision making behaviour. Our proposed machine learning choice model is a departure from the conventional feed-forward MLP framework by using a dynamic residual neural network learning based approach. Our proposed method can be formulated as a Generalized Extreme Value (GEV) random utility maximization model for greater flexibility in capturing unobserved heterogeneity. It can generate choice model structures where the covariance between random utilities is estimated and incorporated into the random error terms, allowing for a richer set of higher-order substitution patterns than a standard logit might be able to achieve. We describe the process of our model estimation and examine the relative empirical performance and econometric implications on two mode choice experiments. We analyzed the behavioural and theoretical properties of our methodology. We showed how model interpretability is possible, while also capturing the underlying complex and unobserved behavioural heterogeneity effects in the residual covariance matrices.

1. Introduction

Enhancing discrete choice models with DNNs and deep learning optimization algorithms is one of the active areas of research that have shown promising results (Sifringer et al., 2018; Borysov et al., 2019; Wong and Farooq, 2020). The increase in popularity of DNNs is primarily due to the idea that these novel modelling strategies emulate behavioural actions through similar neurological functions measured from the human brain. This is motivated by the argument that multi-layered DNN architecture represents the structure of neuron activity patterns and memory in the human brain, therefore assumed to be an efficient way of generating decision models (Friston and Stephan, 2007). This is often referred to as 'biological plausibility' in deep learning literature (Bengio et al., 2015). These similarities between choice behaviour theory and DNNs have led to many interesting and useful applications in travel behaviour modelling and travel demand forecasting (Cantarella and de Luca, 2005; Lee et al., 2018; Wong et al., 2018; Wang and Zhao, 2019). In general, it is a straightforward intuition that these neural networks are composite functions made up of several

[*] Corresponding Author.

Email addresses: `melvin.wong@ryerson.ca` (Melvin Wong), `bilal.farooq@ryerson.ca` (Bilal Farooq) layers of non-linear operators, which enable the feasibility of estimating complex non-linear models using explanatory variables and discrete output choices.

It is generally assumed that by applying specific non-linear transformations on the input data (e.g. *sigmoid*, *hyperbolic tangent*, or *Linear Rectifier Units (ReLU)*), it improves model *prediction accuracy* over Random Utility Maximization (RUM) based models (Cantarella and de Luca, 2005; Wang and Ross, 2018). However, it has been observed that increasing the number of layers beyond a certain threshold would degrade the model due to overfitting, unreachable optimal solutions, and model identification problems (Glorot et al., 2011). Other major problems with DNNs are the lack of model interpretability, parameter stability and general assumptions of the error distribution (Dougherty, 1995; Hensher and Ton, 2000). Even in cases where DNNs were shown to produce far better and accurate predictions than standard linear utility-based RUM models, the formulation of a DNN can be inconsistent. This issue arises due to the inconsistent hyperparameter selection, validation bias and misspecification errors (Hillel et al., 2019). Moreover, the applicability of machine learning algorithms has not yet been clearly justified in behavioural modelling applications and economic analysis (Karlaftis and Vlahogianni, 2011). Discrete choice experiments are typically only stable at aggregate levels – As a result, training and optimizing a multi-layered model to capture individual level variations have not yet provided the expected benefits beyond few layers (Wang and Zhao, 2019).

This problem can be addressed by introducing the concept of residuals – skipped connections between layers. Recent work has shown that this strategy significantly improves learning in deep neural network architectures with marginal or no loss in performance (Witten et al., 2016). A form of DNN architecture that utilizes this strategy is known as Residual Neural Networks or ResNets, which allows for training of very deep neural network models by facilitating gradient backpropagation throughout the layers (He et al., 2016). We show that this solution can be easily adapted for applications in choice behaviour modelling to identify sources of unobserved heterogeneity. This can provide an opportunity for the discrete choice analysts to leverage on interpretable deep learning algorithms to estimate more robust choice models. The goal of this paper is to present a practical application of machine learning in choice modelling research that leverages on the flexibility of the ResNet architecture.

We propose a tractable method of unifying a neural network model architecture with the Generalized Extreme Value (GEV) choice model, which allows the systematic utility function to modelled using standard econometric specification (McFaddden, 1978). It extends previous work on machine learning techniques for choice modelling tasks in two key aspects. First, we can accommodate the random heterogeneity in the choice selection process in the form of residual matrix parameters to account for the influence of unobserved behaviour variations. Second, it allows for parameter estimation stability, model interpretation, economic analysis, and statistical testing, since the formulation resembles a flexible GEV model structure with additive residual error correction terms in the utility function. We define this model structure as a *ResLogit* model.

This paper is organized as follows: Section 2 provides an overview of the heterogeneity representation in discrete choice models and in deep neural networks. Section 3 presents the specification of our proposed ResLogit model. Section 4 demonstrates the method on a classic Red/Blue bus example. Section 5 outlines the estimation process and the learning algorithm. Section 6 presents two choice experiments that analyze the value of time and choice demand prediction, respectively. Finally, the conclusion in Section 7.

2. Background

2.1. Representation of heterogeneity in discrete choice modelling

The standard framework for analyzing and understanding the consumer behaviour has been the discrete choice models such as Multinomial Logit (MNL), Multinomial Probit, Mixed logit or Nested logit (McFadden and Train, 2000). This framework has proven successful because of its simple yet flexible model formulation for relating observable information and attributes to choice behaviour. Representing the effects of endogeneity and expressing behavioural richness is a fundamental challenge in choice modelling (Louviere et al., 2005). It assumes that the underlying decision processes are hidden from the observer and that decision makers select their preferred alternative by ranking all potential choices and choosing the alternative with the maximum utility through actions, dynamics and contexts (Ben-Akiva et al., 2012). The true utility value is not directly observed, but inferred from observed choice behaviour of the decision maker. The unknown factors are not explicitly captured in the model Thus it needs to be approximated through the error terms, which may result in estimation misspecification and inconsistencies with expected rational behaviour.

The MNL model assumes that the probabilities of each pair of alternatives are uncorrelated with the presence of other alternatives known as independence of irrelevant alternatives (IIA) (McFadden and Train, 2000). When choice alternatives are similar or correlated, assumption of IIA may lead to an incorrect forecast of market share as well as model misspecification. Although this simplifying assumption in MNL models may fit well to simple behavioural models and allows for tractable estimation, it creates unrealistic substitution patterns that are not necessarily accurate in reflecting observed human behaviour. The solution is to allow some correction for heterogeneity in the stochastic component of the utility (McFadden and Train, 2000). For instance, the Nested Logit models partially relax the IIA assumption by segmenting alternatives into subsets such that they are similar within each group (non-zero correlation) but independent between groups (zero correlation). In this way, the uncorrelated unobservable heterogeneity is partially accounted for in the model equation. Similarly, Latent Class and Mixed logit model structures reflect the correlation between alternatives by allowing for variable coefficients to vary between observations, class segments or individuals. These methods of capturing heterogeneity assume a specific, pre-defined and non-dynamical structure that is determined by the analyst and may not reflect the underlying correlation that accounts for the complex behaviour learning process.

2.2. Representation of heterogeneity in DNNs

In the past several years, there have been new innovations in machine learning algorithms for combining DNN and RUM based choice models designed to capture the underlying heterogeneity from large datasets. This stems from the increasing importance of incorporating latent psychological effects in hybrid choice models (Thorhauge et al., 2019). For instance, a way of using machine learning in choice modelling is by learning a non-compensatory decision protocol distribution from the data that generalizes many of the decision rules used in discrete choice, instead of defining fixed assumptions about the error distribution (Vythoulkas and Koutsopoulos, 2003). In DNNs, these latent psychological effects can be represented by the non-linear transformation within the network, deriving a set of stochastic variables from observed explanatory variables.

The non-linear functions in these neural network based models are assumed to be able to represent taste variations and random heterogeneity in the choice model. For conventional DNNs, such as the Multi-layer Perceptron (MLP), the objective is to find the set of optimal model parameters and hyperparameters that map inputs to outputs through a series of non-linear transformations.

This transformation process allows the underlying information to be inferred through the neural network. Hyperparameters are arbitrary model parameters that specify the learning procedure or controlling the model complexity such as L_1 and L_2 penalties, update step size, decay or initialization conditions. In some situations, hyperparameter tuning can yield state-of-the-art performance. However, from the viewpoint of choice modelling they have no behavioural meaning behind it. A simple MLP model is seen as a 'black-box' model and behaviour analysts have faced difficulty in understanding how the model parameters interpreted. There can be multiple different models defined by the same set of parameters, making it problematic in model identification and identifying the exact beta parameters.

There are also disadvantages to using DNNs in choice modelling. Even though machine learning methods are increasingly being used in travel mode choice prediction, their usefulness has been limited to prediction tasks (Karlaftis and Vlahogianni, 2011). Early research on using machine learning for mode choice modelling work primarily used prediction accuracy as a comparison and suggests that neural networks appear to lack the consistency with economic principles (Hensher and Ton, 2000; Karlaftis and Vlahogianni, 2011).

2.3. General formulation of a neural network architecture

Each neuron in an MLP is a basic processing unit that performs a non-linear transform on its input (Lee et al., 2018). During the training process, the model is estimated by a gradient descent algorithm given an objective function, e.g. maximum log-likelihood. The basic MLP architecture can be represented mathematically as a series of functions:

$$\begin{aligned} h_{m=1} &= f(x; \mathbf{W}_1), \\ h_{m=2} &= f(h_{m=1}; \mathbf{W}_2), \\ &\cdots \\ y &= logit(h_M), \end{aligned} \tag{1}$$

where x is the input, f is an activation function, h_m is the intermediate m^{th} layer in the neural network model and y is the output choice probabilities of the network. The activation function can be as simple as a linear regression or a continuous non-linear function, for example, a sigmoid unit: $f(x) = (1 + e^{-x})^{-1}$. In general, most DNN architecture are non-identifiable due to the non-linear activation function used. For example, it is not required for each DNN parameters to have any particular sign, since subsequent layers can have the opposite sign to reverse the values. Different permutations of weights are also possible. For instance a 1-layer, 1 hidden neuron network with unit weight is equivalent to a 1-layer, 2 hidden neuron network with half weight each, since the product-sum of input and weights are equal.

The naïve intuition is that the MLPs can progressively learn increasingly complex features by adding more layers. However, it has been empirically shown that for a fully connected MLP architecture, there is a maximum threshold to the number of layers before the model overfits and deteriorates in model prediction accuracy and estimated log-likelihood (Srivastava et al., 2015; He et al., 2016). This performance limitation has been observed and discussed recently in choice modelling applications (Alwosheel et al., 2018; Lee et al., 2018). This simple assumption contradicts the belief that DNNs may achieve greater classification accuracy over conventional discrete choice models and has led to a questionable understanding of the practical uses and importance of deep learning in discrete choice modelling. It has been shown that in the MLP models, increasing the number of layers (and increasing non-linearity) may lead to worse performance compared to a simpler model. This observation contradicts the logic that a M-layered network should, in theory, produce a higher model accuracy than a $(M-1)$-layered network by capturing higher level detail in the model (Srivastava et al., 2015).

ResNets, on the other hand, exploit the use of identity shortcuts to enable the flow of information across layers without causing model degradation from repeated non-linear transformations (He et al., 2016). In an MLPs model, the activation function is applied to the entire layer of input data. ResNets apply a non-linear transformation to the residual (variance) of the input, given by the following:

$$\begin{aligned} h_{m=1} &= f(x) + x \\ h_{m=2} &= f(h_{m=1}) + h_{m=1} \\ &\dots \\ y &= logit(f(h_{m=M}) + h_{m=M}) \end{aligned} \tag{2}$$

The hypothesis behind the ResNet architecture from an optimization perspective is that it is easier to optimize a small change to the input rather than improving the entire layer of inputs at once (He et al., 2016). The ResNet framework uses a skip connection mechanism to propagate information through the layers making it easier to train complex behavioural models. From a behavioural perspective, the model framework is able to retain the systematic portion of the utility. Compared to a feedforward MLP neural network architecture, this approach manages to account for the systematic utility function and econometric variables, allowing for statistical analysis of the interactions between the attributes of the utility and the characteristics of the decision maker. For instance, the econometric parameters (e.g. parameter for travel time, or cost of travel) would be able to "explain" the choices with respect to the individual and alternative attributes, rather than an arbitrary assignment to the respective latent variable output from the neural network.

Our proposed ResLogit hybrid choice model approach improves upon the GEV based RUM framework and incorporates a neural network learning model, while preserving consistencies with the RUM paradigm. The general framework of our ResLogit architecture is that it is much more efficient to model the heterogeneity using a neural network rather than applying it to the entire utility function. Whereas, the systematic utility function can still be modelled using the econometric specification. This translates to using the choice endogeneity in the data to generate the interactions and error term variations directly rather than to specify a fixed structural and measurement equation. A comparison of the structure of an MNL, MLP, and the ResLogit model is shown in Figure 1.

3. Specification of the proposed ResLogit model

We propose a tractable method of unifying a neural network model architecture with the Generalized Extreme Value (GEV) choice model (McFaddden, 1978). The GEV model is based on the assumption of extreme value distributed error terms that can allow for correlation across choice alternatives while having a closed form mathematical solution for estimation. The ResLogit model can be classified as a special variant of a GEV model that inherits the theoretical foundations of random utility models. In particular, we propose the following non-negative G function that

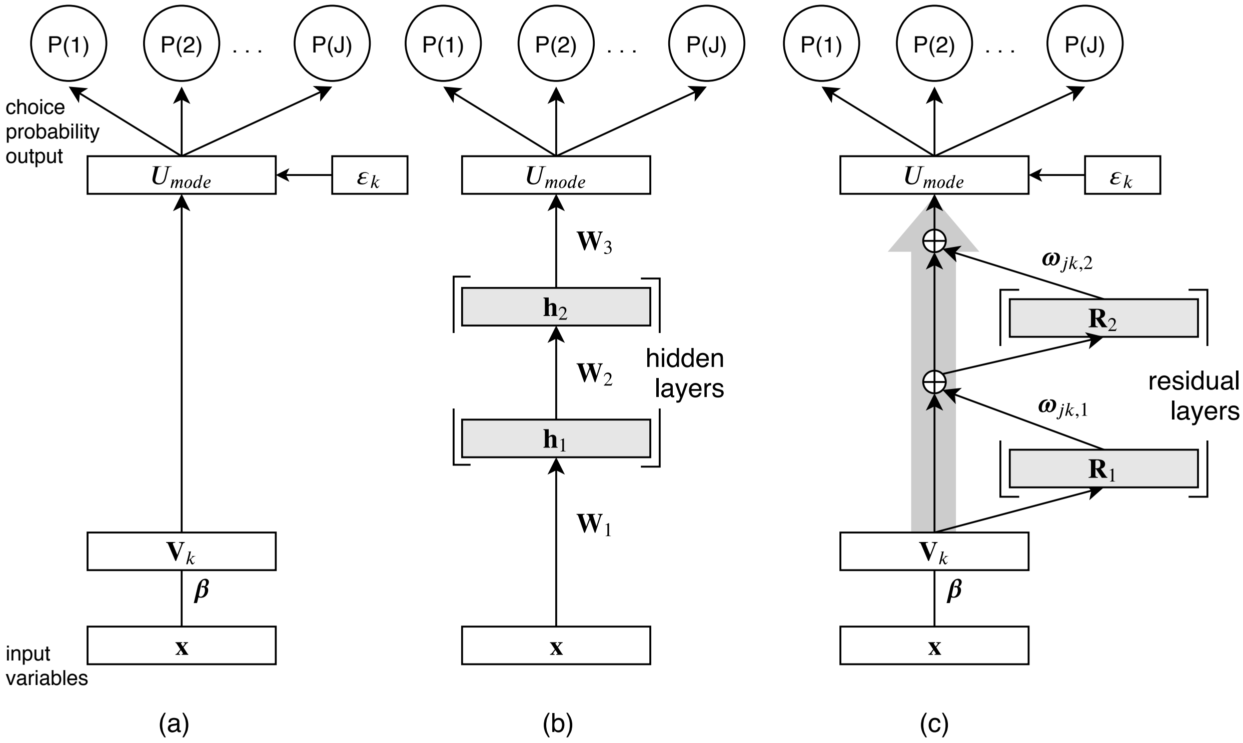


Figure 1: The architecture of: (a) MNL model, (b) MLP-DNN with 2 hidden layers, (c) our proposed ResLogit model.

represents the ResLogit model:

$$G(y_1, \dots, y_j) = \sum_{j \in \mathcal{C}} \Big(\Big[\prod_{m=1}^{M} R_{j,m} \Big] y_j^{\mu} \Big), \tag{3}$$

where $y_j^{\mu} = e^{\mu V_j}$, and V_j is the observed deterministic part of the utility for j alternatives. The product of $R_{j,m}$ terms in the square bracket is the residual correction factor that incorporates the neural network structure into the GEV model. The derivative $\frac{\partial G}{\partial y_j}$ is:

$$\frac{\partial G}{\partial y_j} = G_j(y_1, \dots, y_j) = \mu \Big[\prod_{m=1}^{M} R_{j,m} \Big] y_j^{\mu - 1}. \tag{4}$$

Contrary to the allocation factor used in cross-nested GEV models to reflect the extent at which the alternative j is in a specified nest, the residual correction factor is a stochastic scale value that adjusts the utility function. The justification for moving towards a neural network based structure for the correlation across choice alternatives is such that it becomes a *data-driven* GEV model generating approach using the underlying heterogeneity to estimate the logit structure. We assume that G is a function that exhibits the following properties which can be trivially verified (Ben-Akiva and Lerman, 1985):

1. G is a non-negative function for all y_j;

2. G is homogeneous of degree $\mu > 0$: $\alpha^{\mu} G(y_j) = G(\alpha y_j)$;
3. $\lim\limits_{y_j \to +\infty} G = +\infty$;
4. The l^{th} partial derivative of G with respect to any l distinct y_j is non-negative when l is odd and non-positive when l is even.

Under these conditions, the choice probability of alternative j being selected given a choice set $\mathcal{C}$ derived from the G generator function (3) is defined as follows:

$$\begin{aligned} P(j|\mathcal{C}) &= \frac{y_j G_i}{\mu G}, \\ &= \frac{e^{\mu V_j + \sum\limits_{m=1}^{M} \ln R_{j,m}}}{\sum_{j' \in \mathcal{C}} e^{\mu V_{j'} + \sum\limits_{m=1}^{M} \ln R_{j',m}}} \quad \forall \quad M \geq 1. \end{aligned} \tag{5}$$

for the case of $M = 0$, it is equivalent to a standard MNL model. If the parameters in the residuals R are set to zero, the residual function collapses to a constant ($R = \ln(2)$) and can be factorized out from the logit choice probabilities. Varying the values of the residual does not change the scale of the utility. As we are able to specify the log-likelihood objective function and the gradient on the residual function is continuous, there would be a closed form solution to the choice probability structure.

3.1. Residual function

The residual correction factor $\left[\prod_{m=1}^{M} R_{j,m}\right]$ provides a convenient fully parametric approach for a mixture of the error terms with the underlying unobserved behaviour distribution confounded in the data. We require that $R_{j,m}$ satisfies the following conditions: $0 < R_{j,m} \leq 1$ and $\sum_m R_{j,m} > 0$. Thus, a larger $\prod_{m=1}^{M} R_{j,m}$ value translates to a higher independence of alternative j from the alternatives k, where $k \neq j$. The product of R terms over m represents the joint distribution mixture of higher order unobserved heterogeneity from each m^{th} residual layer. $R_{j,m}$ is formulated as a recursive function:

$$R_{j,m} = \begin{cases} \dfrac{1}{1 + \exp\left(\sum\limits_{k \in \mathcal{C}} \omega_{jk,m} V_k\right)} & \text{if } m = 1, \\ \dfrac{1}{1 + \exp\left(\sum\limits_{k \in \mathcal{C}} \omega_{jk,m} (V_k + \ln R_{j,m-1})\right)} & \text{if } m > 1 \end{cases} \tag{6}$$

where $\omega_{jk,m}$ is the residual relational matrix parameters of the m^{th} residual layer with dimensions $(j \times k)$ to be estimated, which defines the correlation structure of the alternatives. Since the residual correction factor is bounded between 0 and 1 for all $M \geq 1$, $R_{j,m}$ is continuous and asymptotic given any empirical data. The additional optimization term ($\ln R_{j,m}$ in our model) shown in (5) is the measurement quantity for the information content (Louviere et al., 2005). It represents the unobserved interactions and correlated effects of a particular choice action, conditional on the individuals' beliefs or information processing strategy (IPS). The recent use of the rational inattention theory in choice modelling also describes a similar uncertainty term to measure the amount of information used by an individual in decision making (Matějka and McKay, 2015). From a utility-maximization standpoint, the choice outcome may be sub-optimal if this uncertainty is not accounted for.

We can consider the analogy with physical dynamics and information theory. The structure of the ResLogit model corrects for the heterogeneity associated with the information processing cost, moving from a state with high entropy to an optimized final state with low entropy. Although this entropy itself is a latent construct in behavioural modelling, it can be useful to think of it as a form of regularization. It has been shown that imperfect information about the choice distorts the model and leads to choice errors (Matějka and McKay, 2015).

Behavioural researchers have also stressed the importance of accounting for uncertainty in the choice process, such as experiences or habits (Sims, 2003). The uncertainty is modelled as a fixed marginal cost in the utility. As with the rational inattention approach, our ResLogit model does not impose any particular assumptions (e.g. the nesting structure, latent variables etc.) on the underlying prior distribution but derives the structure by learning from the data.

By avoiding the need to create prior assumptions associated with rational expectations in choice behaviour, the error terms become fully flexible and adaptive to the observed data. In theory, any mixture of nesting structure or latent variable error terms could be generated from the ResLogit GEV, provided that the data itself manifests the desired property. Other GEV models and their choice probability distributions can be generated by setting the appropriate matrix parameters ω in the residual function.

3.2. Utility specification

In the ResLogit model, the utility U_{nj} of individual n selecting choice j is obtained as follows:

$$U_{nj} = V_{nj} + \sum_{m=1}^{M} \nu_{nj,m}(\mathbf{V}_{n,m-1}) + \varepsilon_{nj}, \tag{7}$$

where:

V_{nj} = the deterministic linear portion of the utility associated with the characteristics of the individual n and the attributes of the alternatives j;

$\nu_{nj,m}$ = the m^{th} non-linear parametric residual terms of the utility derived from $R_{j,m}$;

ε_{nj} = the unknown random error terms of the utility assumed to be *iid* extreme value distributed with a zero mean.

We denote $\mathbf{V}_n = (V_{n1}, \ldots, V_{nj})$ as the vector of j utility terms for an individual n. For example, the residual components of the utility could look like the following:

$$\begin{aligned}\sum_{m=1}^{M} \nu_{nj,m}(\mathbf{V}_{n,m-1}) &= \nu_{nj,1}(\mathbf{V}_{n,0}) + \nu_{nj,2}(\mathbf{V}_{n,1}) \ldots + \nu_{nj,M}(\mathbf{V}_{n,M-1}) \\ &= -\ln\Big(1 + \exp(\boldsymbol{\omega}_{j,1} \cdot \mathbf{V}_{n,0})\Big) - \ln\Big(1 + \exp(\boldsymbol{\omega}_{j,2} \cdot \mathbf{V}_{n,1})\Big) \ldots \\ &\quad - \ln\Big(1 + \exp(\boldsymbol{\omega}_{j,M} \cdot \mathbf{V}_{n,M-1})\Big),\end{aligned} \tag{8}$$

where $\boldsymbol{\omega}_{j,m} = (\omega_{j1,m}, \dots, \omega_{jk,m})$ is a vector of k parameter terms where $j, k \in \mathcal{C}$, and

$$\begin{aligned} \mathbf{V}_{n,0} &= (V_{n1}, \dots, V_{nj}), \\ \mathbf{V}_{n,m} &= \mathbf{V}_{n,m-1} + \nu_{nj,m}(\mathbf{V}_{n,m-1}) \quad \forall \quad 1 \leq m < M \end{aligned} \tag{9}$$

4. Red/Blue bus example

We show an example of how a simple nesting structure can be obtained using the estimated $\boldsymbol{\omega}$ matrix. Let us consider the classic red/blue bus problem. The choice scenarios are summarized in Table 1. Assuming that we have 2 initial choices: bus (V_{bus}) and car (V_{car}), each have identical observed utility: $V_{car} = 1$, $V_{bus} = 1$, under RUM theory, the probability of choosing either bus or car is $P_{car} = P_{bus} = 0.5$ (Scenario 1).

Suppose that now we have a red bus (V_{red}) and blue bus (V_{blue}) option in place of V_{bus}, $\mathbf{V} = (V_{car}, V_{red}, V_{blue})$, given a set of simulated utilities where all 3 options have identical utilities: $V_{car} = 1$, $V_{red} = 1$, $V_{blue} = 1$. The outcome of rational behaviour choice probabilities assuming IIA should be $P_{car} = 0.5$, $P_{red} = 0.25$, and $P_{blue} = 0.25$. However, the actual probabilities when estimated by a MNL model would produce $P_{car} = 0.33$, $P_{red} = 0.33$, and $P_{blue} = 0.33$ which violates IIA conditions (Scenario 2).

Under the ResLogit model, the correlation between the red and blue bus is corrected by a residual function ν, with parameter matrix $\boldsymbol{\omega}$. Our model would be able to capture the endogeneity between the red/blue bus choice options (Scenario 3).

Table 1: Illustration of red/blue bus choice scenarios showing the effect of residual correction factors of a 1-layer model.

Choice	V_j	$\nu(V_1, \dots, V_j)$	$e^{V_j + \nu(V_1, \dots, V_j)}$	Prob.
Scenario 1				
car	1	0	2.718	0.5
bus	1	0	2.718	0.5
Scenario 2				
car	1	0	2.718	0.33
red bus	1	0	2.718	0.33
blue bus	1	0	2.718	0.33
Scenario 3				
car	1	-0.127	2.394	0.468
red bus	1	-0.693	1.359	0.265
blue bus	1	-0.693	1.359	0.265

Using a 1-layer ResLogit model and a residual function defined by $\nu = -\ln(1 + \exp(\boldsymbol{\omega} \cdot \mathbf{V}))$ as an example, we create a synthetic choice simulation with the three alternatives $\mathcal{C} = \{car, red, blue\}$. From our simulated example, we assume an estimated model with the following $\boldsymbol{\omega}$ residual matrix from a zero matrix initialization ω_{init}:

$$\boldsymbol{\omega}_{init} = \begin{array}{c} \\ V_{car} \\ V_{red} \\ V_{blue} \end{array} \begin{array}{c} \begin{array}{ccc} \nu_{car} & \nu_{red} & \nu_{blue} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \rightarrow \boldsymbol{\omega}_{final} = \begin{array}{c} \\ V_{car} \\ V_{red} \\ V_{blue} \end{array} \begin{array}{c} \begin{array}{ccc} \nu_{car} & \nu_{red} & \nu_{blue} \end{array} \\ \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \end{array},$$

where $\boldsymbol{\omega}_{init}$ are the initial model parameters before estimation and $\boldsymbol{\omega}_{final}$ are the estimated model parameters. In this example, we assume that the scales of the utilities are all equal, i.e. the diagonal elements of $\boldsymbol{\omega}_{diag} = 0$. A value of $\omega_{jk} = 1$ signifies that increasing the utility of an alternative j would cause a similar increase in alternative k. With a value of $\omega_{jk} = -1$, increasing the utility of an alternative j would cause a change in utility value in the opposite direction for an alternative k. Given a utility vector $\mathbf{V} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$, the residual is:

$$\nu_{1\ldots j} = -\ln\Big(1 + \exp \begin{bmatrix} \omega_{11} & \omega_{12} & \omega_{13} \\ \omega_{21} & \omega_{22} & \omega_{23} \\ \omega_{31} & \omega_{32} & \omega_{33} \end{bmatrix} \cdot \begin{bmatrix} V_{car} \\ V_{red} \\ V_{blue} \end{bmatrix}^\top \Big), \tag{10}$$

$$= -\ln\Big(1 + \exp \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}^\top \Big), \tag{11}$$

$$= \begin{bmatrix} -0.127 \\ -0.693 \\ -0.693 \end{bmatrix} \tag{12}$$

$$P(car, red, blue) = \begin{bmatrix} 0.468 & 0.265 & 0.265 \end{bmatrix} \tag{13}$$

The residual matrix $\boldsymbol{\omega}_{final}$ indicates that the red/blue bus alternatives are correlated with a positive value $\omega_{red,blue} > 0$: Increasing the utility of the red bus will cause a similar increase in utility for the blue bus. If $\omega_{jk} = 0$, then alternative j and k are assumed to be independent (IIA condition holds).

5. Estimation process

The size of the residual matrix varies according to the number of alternatives, such that each element in the matrix corresponds to how the alternative specific utilities are influenced by other alternatives. The matrix diagonal elements represent the variances of each alternative, e.g. covariance with itself. If this residual matrix is an identity matrix, that means that there is zero correlation between alternatives (IIA holds), and the model collapses into a standard MNL model.

5.1. Depth of the neural network

Increasing the depth of the neural network increases the number of recursive addition of residual terms in the utility function. This mathematical formulation allows the model to extract the underlying prior information to reflect individual taste heterogeneities, with the residual layers representing the nature of the complex behavioural distribution of decision makers that are not captured by the observed explanatory variables. The exact number of layers used does not become a constraint in the model, and the residuals allow learning of DNNs with a low potential of overfitting or degradation in model accuracy. This is the primary advantage of having a residual model over a direct application of a feedforward neural network such as the MLP model. The key implication of this on choice models is that we can operationalize the unobserved heterogeneity in the GEV model as a product of the neural network learning process, while retaining the same econometric parameters in the structural equation.

5.2. Objective function

The expression for the probability that an individual n selects a particular choice given I explanatory variables $x_1, ..., x_I$ in a ResLogit is:

$$P_n(j|\mathcal{C};\beta,\boldsymbol{\omega}) = \frac{e^{V_{nj}+\sum_{m=1}^{M}\nu_{nj,m}(\boldsymbol{\omega})}}{\sum_{j'\in\mathcal{C}} e^{V_{nj'}+\sum_{m=1}^{M}\nu_{nj',m}(\boldsymbol{\omega})}}, \tag{14}$$

where

$$V_{nj} = \sum_{i=1}^{I}\beta_{ji}x_{ni} \tag{15}$$

The set of optimal parameters $\hat{\beta}, \hat{\boldsymbol{\omega}}$ are estimated by maximizing the log-likelihood through a batched stochastic gradient descent algorithm:

$$\mathcal{L}_t = \frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{J}\ln P_t(j|\mathcal{C};\beta,\boldsymbol{\omega}) \tag{16}$$

We then perform the updates to the model parameters with a fixed learning rate η. In each batch training step t, we compute the likelihood $\mathcal{L}_t$ and shift the parameter estimate opposite to the gradient direction:

$$\hat{\beta}_{t+1} \leftarrow \beta_t - \eta\left(\nabla_\beta \mathcal{L}_t\right) \tag{17}$$

$$\hat{\boldsymbol{\omega}}_{t+1} \leftarrow \boldsymbol{\omega}_t - \eta\left(\nabla_{\boldsymbol{\omega}} \mathcal{L}_t\right) \tag{18}$$

6. Case studies

We present two different choice experiments on the ResLogit model. The first experiment analyzes the Value of Time (VoT) of a Stated Preference (SP) travel survey. The second experiment investigates the effects of model depth on a large scale Revealed Preference (RP) dataset to investigate the properties of the residuals, model degradation, and mode choice prediction performance.

6.1. Case study 1: Value of travel time analysis on an SP survey

We compared our ResLogit modelling method with a random parameters Mixed Logit model on the value of travel time savings. The dataset used for this analysis is from the 2016 *Train Hotel* SP based travel demand analysis study for a new intercity travel mode, Train Hotel (“TrH”), which provides an overnight sleeper and dining service between business and tourist destinations (Wong and Farooq, 2018). Users of TrH are also expected to save on the hotel cost through this service. This study covers metropolitan regions in the province of Quebec and Northeastern U.S.A.

Table 2: Train Hotel survey characteristics

	mode share preference	mean travel cost ($)	mean travel time (h)
mode			
car	0.239	115	7.26
car rental	0.021	179	6.29
bus	0.057	84	11.72
plane	0.056	178	4.04
train	0.060	48	6.12
TrH	0.567	164	13.78

– Montreal, Sherbrooke, New York City, Boston, Maine, etc. The mode share preference and the statistical characteristics of the two explanatory variables used are presented in Table 2.

This survey revealed that more than half of the respondents (56.7%) prefer the proposed TrH alternative despite having the longest average travel time (13.78 hours). This indicates the possibility of obtaining negative VoT from a positive travel time β. In a simple choice model experiment, we consider travel cost and time in the deterministic component of the utilities, as shown in the Logit utility equation:

$$U_{mode} = \beta_{cost} X_{mode\_cost} + \beta_{TT} X_{mode\_TT} + \varepsilon_{mode} \tag{19}$$

For the Mixed Logit model with a Normal distribution $\mathcal{N}$ over the β parameters:

$$U_{mode} = (\bar{\beta}_{cost} + \sigma_{cost}\mathcal{N}(0,1)) X_{mode\_cost} + (\bar{\beta}_{TT} + \sigma_{TT}\mathcal{N}(0,1)) X_{mode\_TT} + \varepsilon^*_{mode} \tag{20}$$

and for the ResLogit model:

$$U_{mode} = \beta_{cost} X_{mode\_cost} + \beta_{TT} X_{mode\_TT} - \ln(1 + \exp(\boldsymbol{\omega}_{j,1} \cdot V_0)) + \varepsilon^*_{mode} \tag{21}$$

where *mode* is one of *car, car rental, bus, plane, train* or *TrH.* We use a 1-layer ResLogit model for our model training and compared the estimated β and alternative specific constants against a Mixed Logit and Logit model.

The three SP choice models were coded and estimated in Python using PythonBIOGEME library (Bierlaire, 2016). Table 3 presents the model estimation and VoT results for the different mode alternatives. It can be seen that for the ResLogit model, the sign of β_{TT} is positive (0.05), indicating that the respondents are willing to choose the mode of travel with *greater* travel time that also provides greater flexibility and comfort. The model performance indicates that the ResLogit model performs the best in terms of log-likelihood. The residual term in (21) accounts for the unobserved perceived value of the travel modes and the individual's attitude towards the new hypothetical TrH option. Another effect of the residual term is that the alternative specific parameters (ASC) converge to zero, as shown by the insignificant values.

Table 4 verifies our hypothesis of unobserved value-added benefits from the TrH option. By removing the proposed alternative, the ResLogit model becomes similar to the Mixed Logit model in terms of β values. Likewise the VOT sign changes to positive (20.59) capturing the consistency

Table 3: Estimated model parameters and VoT calculation. Std. error in parenthesis.

	Logit	Mixed Logit	ResLogit
Parameters			
ASC_{car}	-2.53 (0.25)	-2.58 (0.26)	2.19 (2.84)
$\text{ASC}_{\text{car rental}}$	-1.7 (0.12)	-1.71 (0.12)	2.48 (4.61)
ASC_{bus}	-3.46 (0.23)	-3.6 (0.25)	6.18 (25.1)
$\text{ASC}_{\text{plane}}$	-2.36 (0.25)	-2.42 (0.27)	-0.62 (1.32)
$\text{ASC}_{\text{train}}$	-2.27 (0.21)	-2.38 (0.22)	1.3 (0.77)
ASC_{TrH}	ref.	ref.	ref.
β_{cost}	-0.58 (0.11)	-0.64 (0.12)	-0.48 (0.13)
β_{TT}	-0.072 (0.02)	-0.064 (0.024)	0.05 (0.01)
σ_{cost}		0.71 (0.22)	
σ_{TT}		0.2 (0.06)	
VoT (CAD$/hour)	12.41	9.98	-9.75
log-likelihood	-2060.1	-2058.3	-2036.6
sample size	1788	1788	1788
$\bar{\rho}^2$	0.29	0.295	0.302

Table 4: Estimated model parameters and VoT calculation **without TrH mode**. Std. error in parenthesis.

	Logit	Mixed Logit	ResLogit
Parameters			
ASC_{car}	-0.71 (0.37)	-0.67 (0.41)	0.18 (0.37)
$\text{ASC}_{\text{car rental}}$	-0.92 (0.19)	-0.89 (0.21)	3.81 (10.7)
ASC_{bus}	-1.26 (0.30)	-1.27 (0.31)	-1.26 (0.37)
$\text{ASC}_{\text{plane}}$	-0.91 (0.42)	-0.93 (0.53)	-1.01 (0.65)
$\text{ASC}_{\text{train}}$	ref.	ref.	ref.
β_{cost}	-2.62 (1.56)	-0.40 (0.22)	-0.10 (0.06)
β_{TT}	-0.134 (0.04)	-0.15 (0.06)	-0.02 (0.01)
σ_{cost}		0.64 (0.34)	
σ_{TT}		0.09 (0.10)	
VoT (CAD$/hour)	5.11	38.20	20.59
log-likelihood	-884.1	-843.3	-821.0
sample size	775	775	775
$\bar{\rho}^2$	0.214	0.213	0.216

of the existing modes (car, car rental, bus, train and plane) without the TrH alternative.

Although negative VoT obtained from the ResLogit model (-9.75) clearly contradicts rational economic behaviour (where travel time is a factor of disutility), we attribute the exception in our case study to the fact that it includes value-added amenities onboard the overnight train which are not captured by travel time parameters, allowing for positive β_{TT}. The evidence in the literature suggests that travel time offers a source of positive utility if the utility of travel includes activity at the destination, conducted while travelling or enjoyment of the act of travel itself (Redmond and Mokhtarian, 2001). Our experiment has shown that an individual might choose an alternative with a higher travel time, taking into account other unobserved non-monetary benefits, indicated by the positive time beta estimated in the ResLogit model. By contrast, the Logit and Mixed Logit model estimated a negative time beta, which is a rational economic outcome, but does not reflect the survey data and mode specific features accurately. The design of the *Train Hotel* SP travel study allows for the possibility of positive β_{TT}, travellers are willing to pay for a longer travel time without taking into account other benefits of travel. From Table 3, the ResLogit produced a negative VoT, which is consistent with our choice context and the travel survey data results.

6.2. Case study 2: Mode choice prediction on an RP survey

We developed our second experiment on the 2016 *Mtl Trajet* RP travel survey dataset collected from the user's smartphone data on a mobile application (Yazdizadeh et al., 2017). Table 5 shows a list of explanatory variables and the choice set used for this mode choice prediction analysis. The respondents' travel diary includes mode choice, activity choice, trip attributes and GPS trajectories. The travel survey was conducted over 4 months, from September to December 2016. In total, there are 60,365 unique trips made during the period. To control for overfitting in our ResLogit mode choice prediction model, we divide the dataset into two subsets using a 70:30 training/validation split ($N_{training} = 42,256$, $N_{validation} = 18,109$).

The model estimation algorithm has been coded in Python using the Theano deep learning library and estimated on the large scale RP dataset (Theano Development Team, 2016). The

Table 5: Descriptive statistics of the 2016 *Mtl Trajet* RP travel survey dataset.

variable	description	type	mean	std dev
weekend	trip on weekend	dummy variable	0.205	0.001
hour_8_10	trip between 8am to 10am	dummy variable	0.163	0.0015
hour_11_13	trip between 11am to 1pm	dummy variable	0.147	0.001
hour_14_16	trip between 2pm to 4pm	dummy variable	0.209	0.002
hour_17_19	trip between 5pm to 7pm	dummy variable	0.249	0.002
hour_20_22	trip between 8pm to 10pm	dummy variable	0.095	0.001
hour_23_1	trip between 11pm to 1am	dummy variable	0.03	6e-4
hour_2_4	trip between 2am to 4am	dummy variable	0.006	3e-4
hour_5_7	trip between 5am to 7am	dummy variable	0.101	0.005
num_coord	number of trajectory links	continuous	109.8	131.23
trip_dist	trip distance (km)	continuous	8.366	10.42
trip_duration	trip duration (min)	continuous	24.04	20.97
trip_aspeed	trip average speed (km/h)	continuous	22.503	18.815
activity	trip activity type, 1: education, 2: health, 3: leisure, 4: meal, 5: errands, 6: shopping 7: home, 8: work, 9: meeting	categorical		
choice	1:Auto, 2:Bike, 3:Public Transit, 4:Walk, 5:Auto+Transit, 6:Other mode, 7:Other combination			

advantage of using Theano library is that it is able to implement the calculations on a GPU at an abstract level, reducing the computational times for our ResLogit model. The experiments are iterated by varying the depth of the ResLogit model: $M = \{1, 2, 4, 8, 16\}$. The goal of this process is to understand the effects of increasing the size of the residual layers on the choice model. This experiment considers three specific criteria:

1. The effects of the number of residual layers on the model β parameters.
2. Model prediction accuracy and likelihood test results on the validation set.
3. A performance comparison against a conventional MLP-DNN architecture with an identical number of layers and parameters.

We established our benchmark using a standard MNL model with all the parameters shown in Table 5. For the ResLogit model, we initialized the residual parameters using an identity matrix. We used a SGD learning algorithm with a batch size of 32 (i.e. gradient is computed over a sample of 32 observations from the training dataset) to train our models.

Figure 2 shows the validation log-likelihood estimates and a comparison between the MLP-DNN and ResLogit models of different depth sizes. In the Reslogit model (Figure 2, right), we observed that as we increase the depth of the model, the log-likelihood remains consistent and higher than the benchmark MNL. By contrast, the MLP-DNN validation log-likelihood estimates reveal that it does not perform well as a ResLogit model of the same number of layers and hidden units, and it overfits the training data as we increase the number of layers in the model. An indication of overfitting in the MLP-DNN model is shown by the "spikes" in the training curve, which suggests that the MLP-DNN architecture is unable to generalize to new samples. The ResLogit model does not show such behaviour in the validation and log-likelihood results. We note that for our experiment consistency, we did not implement any other forms of regularization, e.g. L_1, L_2 regularizer or Dropout techniques. In theory, as we increase the complexity and non-linearity in the model, it would fit the data better. However, the MLP model suffers from the problem of overfitting as we increase the number of layers, an overfit effect is observed with depth $M \geq 4$, where the validation log-likelihood no longer improves and performed *worse* than the MNL benchmark. In the ResLogit model, we do not see this detrimental effect, even at a depth of 16 layers. This surprising fact highlights how a simple comparison between a MNL and machine learning algorithm may sometimes be misleading without first understanding the structure of the model.

In terms of behavioural interpretation, the ResLogit model is able to represent the underlying choice heterogeneity far better due to the inherent structure of the residual layers. The residual layers add another level of complexity, without interfering with the explanatory variables and the model β parameters. In other words, the neural network component in the ResLogit model extracts the variations of the β parameters through a data-driven learning process, rather than pre-defining a fixed distribution over the β parameters. The model β parameters estimated by the ResLogit model tend to be closer to the 'true' mean values while the taste variance is explained by the $\boldsymbol{\omega}$ parameters.

Table 6 shows the model performances across the training, validation and predictive accuracy on the validation set. The ResLogit model averaged a predictive accuracy of 76.09% ± 0.609% while the MLP-DNN model fared worse with an average of 70.95% ± 1.62% compared to the benchmark MNL model (72.01%).

Table 7 shows selected β parameter estimates from the benchmark MNL and the 16-layer ResLogit model. The results revealed that the parameters estimated by the ResLogit model are relatively different from the MNL model, indicating that the taste variation between mode choice

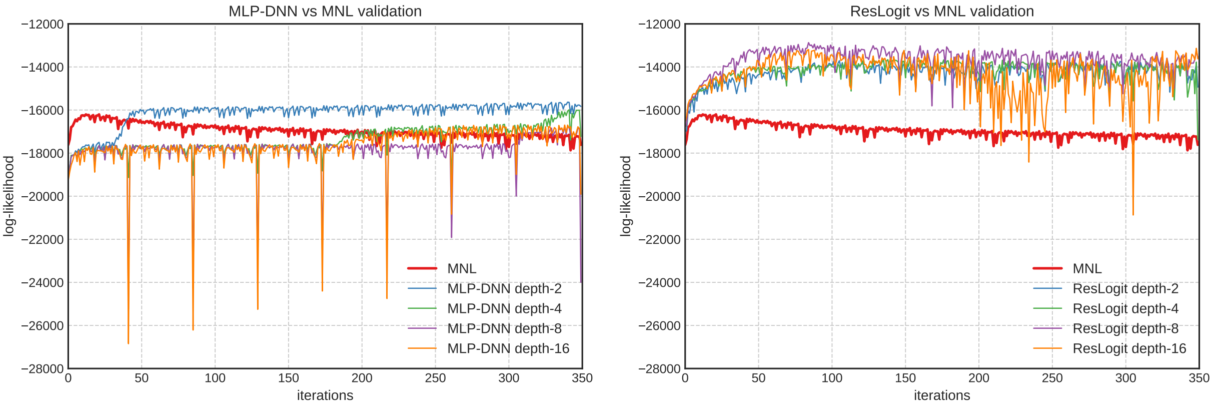


Figure 2: Training and validation results.

Table 6: Prediction accuracy, training and validation log-likelihood results.

	MNL			MLP-DNN			ResLogit		
M	train LL	valid LL	acc.	train LL	valid LL	acc.	train LL	valid LL	acc.
0	-38790	-16145	0.720	-	-	-	-	-	-
2	-	-	-	-37208	-15583	0.726	-40217	-13675	0.752
4	-	-	-	-37820	-15894	0.719	-32342	-13583	0.762
8	-	-	-	-38496	-16736	0.698	-30592	-12870	0.773
16	-	-	-	-42240	-16667	0.695	-31887	-13121	0.767

validation set sample size: 18,109

and the underlying unobserved heterogeneity may have influenced the choice behaviour outcome. The residual-corrected β values in the ResLogit model are shown to have a much smaller standard error. This demonstrates that the heterogeneity confounded in the β parameters is explained away by the residual layers. In terms of parameter significance, the ResLogit parameters have a nominal p-value < 0.05. We note that our results are based on the fixed assumptions of independent and identically distributed (IID) observations and we used the training set to calculate the standard errors at a fixed number of iterations (250). Figure 3 shows the first 4 layers of residual weight matrices from the ResLogit model. The weight parameters do not have a significant meaning to the other β parameters. However, these parameters influence the variance of the error terms. The higher level layers do not have a significant meaning or interpretation, but are useful in capturing the higher-order variances (variance of variance components).

7. Conclusion

We present a novel hybrid choice model that integrates a residual neural network architecture into the RUM model structure. A ResLogit model is designed based on the understanding of residual connections in DNNs that significantly improves learning, out-of-sample-prediction and obtain unbiased model parameter estimates. Two key objectives are accomplished that resolve the shortcomings in machine learning for discrete choice modelling – overfitting due to systematic error of biased model estimates in DNN and lack of flexible economic interpretability. We present a new approach using neural networks for discrete choice analysis in the form of a residual learning model, namely a ResLogit model. Our proposed solution is able to maintain economic interpretability without the loss in prediction performance.

The direct implications of our approach in research as well as practice are demonstrated by a classic red/blue bus example in Section 4, analysis of VoT in Section 6.1, and choice prediction performance in Section 6.2. Unlike earlier studies that only examined the performance of machine learning algorithms and their comparisons with discrete choice models in out-of-sample predictions, this paper studies the impact of the *model structure*, which utilizes deep neural networks consistent with GEV models.

One of the main criticisms of DNNs such as MLPs is that it cannot be easily adapted for econometric analysis due to its 'black-box' nature On the other hand, machine learning and DNNs have emerged as a powerful modelling strategy due to its theoretical ability to provide an accurate out-of-sample prediction. Understanding how we can leverage these machine learning strategies effectively is one of the primary challenges in choice modelling. The key contributing feature of our proposed ResLogit model is the non-degenerate depth-invariant structure – this model structure does not exhibit the same performance degradation seen in MLPs with increasing depth. Therefore, it allows learning of very complex behavioural models from data without significant loss in performance over simpler non-linear discrete choice models. Also, our method leverages on large datasets to learn a richer set of higher-order substitution patterns rather than the standard way of pre-determined error distributions over the model parameters. Finally, our method can be formulated as a GEV random utility maximization model.

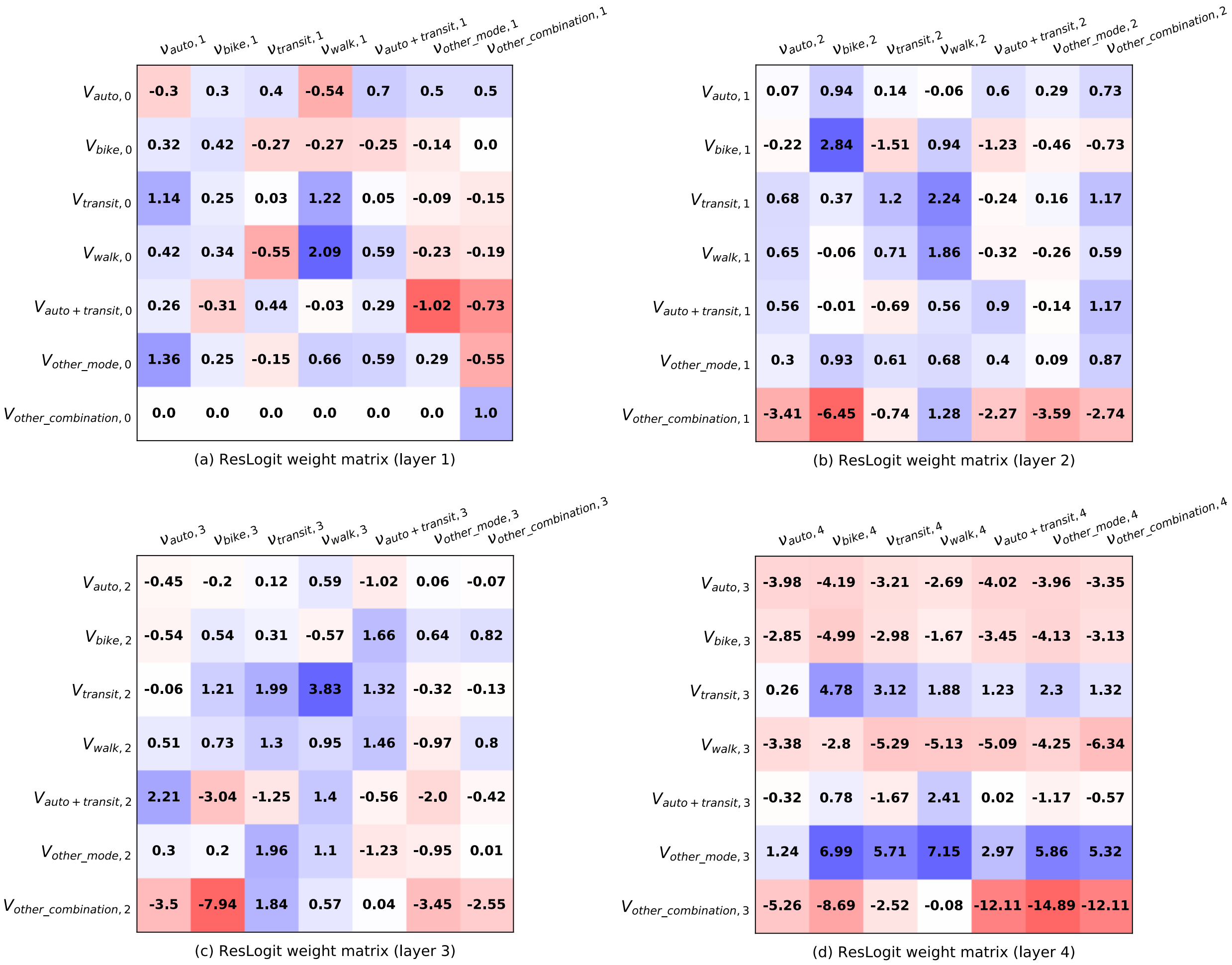
Two case studies are conducted. The first study establishes the use of our method to understand how a neural network structure like our ResLogit model can have significant VoT estimation differences and policy implications. The VoT estimation can be biased if the model specification is poor or not well formulated. Our results showed that even with a Mixed Logit model, the VoT for our case study example could not recover a suitable estimate based on an SP choice survey.

The second case study examines the performance comparison with a MNL-DNN model with an exact number of layers and parameters. The MNL-DNN model fails to optimize parameters on out-of-sample data and our experiment has shown it performs worse than a simple MNL model. For the ResLogit model, it performs better than both the benchmark MNL and the MNL-DNN model. This result shows that a residual learning approach serves a meaningful extension to discrete choice models and offers richer behavioural insights.

The emerging research using Big Data and machine learning has made great strides in choice modelling, but current methods of analysis and modelling have very limited capabilities and flexibility in estimating complex and noisy data. Our proposed ResLogit model offers a practical and logical solution for how choice modelling researchers can systematically implement machine learning algorithms into existing discrete choice models without sacrificing model interpretability. Future research will establish additional models and extensions to our proposed method.

Table 7: Comparison of a subset of parameter estimates between MNL and ResLogit model.

	MNL		ResLogit (16-layer)		
Parameters (β_{mj})	Value	std. err.	Value	std. err.	Value diff.
β weekend_auto	0.02	0.007	0.225	0.006	0.205
β weekend_bike	-1.055	2.776	-0.069	0.044	0.986
β weekend_auto+transit	-0.685	0.309	0.189	0.017	0.874
β hour_8_10_walk	-0.957	0.039	-3.477	0.038	-2.52
β hour_8_10_auto+transit	-2.493	0.328	1.835	0.028	4.328
β hour_11_13_bike	-1.061	1.234	0.132	0.049	1.193
β hour_11_13_auto+transit	-3.031	0.918	1.562	0.031	4.593
β hour_17_19_auto	0.029	0.002	-0.836	0.004	-0.865
β hour_20_22_bike	-1.477	1.462	0.201	0.059	1.678
β trip_dist_auto	0.409	0.022	-0.275	0.001	-0.684
β trip_dist_transit	0.258	0.039	0.113	0.004	-0.145
β trip_dist_walk	-2.008	0.204	-0.536	0.01	1.472
β trip_time_auto	-0.653	0.027	0.24	0.001	0.893
β trip_time_transit	0.84	0.008	1.121	0.002	0.281
β trip_time_walk	0.88	0.272	0.057	0.005	-0.823
β trip_speed_auto	0.919	0.085	-0.109	0.001	-1.028
β trip_speed_walk	-1.386	0.042	-0.878	0.006	0.508
β activity_edu_auto	-1.303	2.458	-0.089	0.011	1.214
β activity_edu_walk	-0.121	0.004	-5.645	0.063	-5.524
β activity_home_auto	-0.069	0.246	-0.015	0.004	0.054
β activity_home_bike	-1.108	0.075	1.357	0.03	2.465
β activity_home_transit	0.085	0.003	1.455	0.012	1.37
β activity_work_auto	-0.016	0.012	-0.077	0.004	-0.061
β activity_work_transit	-0.039	0.002	1.386	0.012	1.425
β activity_work_auto+transit	-1.877	0.745	-0.353	0.023	1.524
β activity_meeting_bike	-2.852	2.417	0.684	0.115	3.536
log-likelihood	-16145		-13121		
sample size	42,255		42,255		
# of estimated parameters	138		922		
validation accuracy	72.01%		76.73%		



	$\nu_{auto,1}$	$\nu_{bike,1}$	$\nu_{transit,1}$	$\nu_{walk,1}$	$\nu_{auto+transit,1}$	$\nu_{other\_mode,1}$	$\nu_{other\_combination,1}$
$V_{auto,0}$	-0.3	0.3	0.4	-0.54	0.7	0.5	0.5
$V_{bike,0}$	0.32	0.42	-0.27	-0.27	-0.25	-0.14	0.0
$V_{transit,0}$	1.14	0.25	0.03	1.22	0.05	-0.09	-0.15
$V_{walk,0}$	0.42	0.34	-0.55	2.09	0.59	-0.23	-0.19
$V_{auto+transit,0}$	0.26	-0.31	0.44	-0.03	0.29	-1.02	-0.73
$V_{other\_mode,0}$	1.36	0.25	-0.15	0.66	0.59	0.29	-0.55
$V_{other\_combination,0}$	0.0	0.0	0.0	0.0	0.0	0.0	1.0

(a) ResLogit weight matrix (layer 1)

	$\nu_{auto,2}$	$\nu_{bike,2}$	$\nu_{transit,2}$	$\nu_{walk,2}$	$\nu_{auto+transit,2}$	$\nu_{other\_mode,2}$	$\nu_{other\_combination,2}$
$V_{auto,1}$	0.07	0.94	0.14	-0.06	0.6	0.29	0.73
$V_{bike,1}$	-0.22	2.84	-1.51	0.94	-1.23	-0.46	-0.73
$V_{transit,1}$	0.68	0.37	1.2	2.24	-0.24	0.16	1.17
$V_{walk,1}$	0.65	-0.06	0.71	1.86	-0.32	-0.26	0.59
$V_{auto+transit,1}$	0.56	-0.01	-0.69	0.56	0.9	-0.14	1.17
$V_{other\_mode,1}$	0.3	0.93	0.61	0.68	0.4	0.09	0.87
$V_{other\_combination,1}$	-3.41	-6.45	-0.74	1.28	-2.27	-3.59	-2.74

(b) ResLogit weight matrix (layer 2)

	$\nu_{auto,3}$	$\nu_{bike,3}$	$\nu_{transit,3}$	$\nu_{walk,3}$	$\nu_{auto+transit,3}$	$\nu_{other\_mode,3}$	$\nu_{other\_combination,3}$
$V_{auto,2}$	-0.45	-0.2	0.12	0.59	-1.02	0.06	-0.07
$V_{bike,2}$	-0.54	0.54	0.31	-0.57	1.66	0.64	0.82
$V_{transit,2}$	-0.06	1.21	1.99	3.83	1.32	-0.32	-0.13
$V_{walk,2}$	0.51	0.73	1.3	0.95	1.46	-0.97	0.8
$V_{auto+transit,2}$	2.21	-3.04	-1.25	1.4	-0.56	-2.0	-0.42
$V_{other\_mode,2}$	0.3	0.2	1.96	1.1	-1.23	-0.95	0.01
$V_{other\_combination,2}$	-3.5	-7.94	1.84	0.57	0.04	-3.45	-2.55

(c) ResLogit weight matrix (layer 3)

	$\nu_{auto,4}$	$\nu_{bike,4}$	$\nu_{transit,4}$	$\nu_{walk,4}$	$\nu_{auto+transit,4}$	$\nu_{other\_mode,4}$	$\nu_{other\_combination,4}$
$V_{auto,3}$	-3.98	-4.19	-3.21	-2.69	-4.02	-3.96	-3.35
$V_{bike,3}$	-2.85	-4.99	-2.98	-1.67	-3.45	-4.13	-3.13
$V_{transit,3}$	0.26	4.78	3.12	1.88	1.23	2.3	1.32
$V_{walk,3}$	-3.38	-2.8	-5.29	-5.13	-5.09	-4.25	-6.34
$V_{auto+transit,3}$	-0.32	0.78	-1.67	2.41	0.02	-1.17	-0.57
$V_{other\_mode,3}$	1.24	6.99	5.71	7.15	2.97	5.86	5.32
$V_{other\_combination,3}$	-5.26	-8.69	-2.52	-0.08	-12.11	-14.89	-12.11

(d) ResLogit weight matrix (layer 4)

Figure 3: First 4 layers of weight matrices from the ResLogit model.

References

Alwosheel, A., van Cranenburgh, S., Chorus, C.G., 2018. Is your dataset big enough? sample size requirements when using artificial neural networks for discrete choice analysis. Journal of choice modelling 28, 167–182.

Ben-Akiva, M., de Palma, A., McFadden, D., Abou-Zeid, M., Chiappori, P.A., de Lapparent, M., Durlauf, S.N., Fosgerau, M., Fukuda, D., Hess, S., Manski, C., Pakes, A., Picard, N., Walker, J., 2012. Process and context in choice models. Marketing Letters 23, 439–456.

Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete choice analysis: theory and application to travel demand. MIT press, Cambridge MA.

Bengio, Y., Lee, D.H., Bornschein, J., Mesnard, T., Lin, Z., 2015. Towards biologically plausible deep learning. arXiv preprint arXiv:1502.04156 .

Bierlaire, M., 2016. PythonBiogeme: a short introduction. Technical Report TRANSP-OR 160706. EPFL.

Borysov, S.S., Rich, J., Pereira, F.C., 2019. How to generate micro-agents? a deep generative modeling approach to population synthesis. Transportation Research Part C: Emerging Technologies 106, 73–97.

Cantarella, G.E., de Luca, S., 2005. Multilayer feedforward networks for transportation mode choice analysis: An analysis and a comparison with random utility models. Transportation Research Part C: Emerging Technologies 13, 121–155.

Dougherty, M., 1995. A review of neural networks applied to transport. Transportation Research Part C: Emerging Technologies 3, 247–260.

Friston, K.J., Stephan, K.E., 2007. Free-energy and the brain. Synthese 159, 417–458.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, in: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, pp. 315–323.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the 29th IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hensher, D.A., Ton, T.T., 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. Transportation Research Part E: Logistics and Transportation Review 36, 155–172.

Hillel, T., Bierlaire, M., Jin, Y., 2019. A systematic review of machine learning methodologies for modelling passenger mode choice. Technical Report TRANSP-OR 191025. EPFL.

Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. Transportation Research Part C: Emerging Technologies 19, 387–399.

Lee, D., Derrible, S., Pereira, F.C., 2018. Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling. Transportation Research Record 2672, 101–112.

Louviere, J., Train, K., Ben-Akiva, M., Bhat, C., Brownstone, D., Cameron, T.A., Carson, R.T., Deshazo, J., Fiebig, D., Greene, W., et al., 2005. Recent progress on endogeneity in choice modeling. Marketing Letters 16, 255–265.

Matějka, F., McKay, A., 2015. Rational inattention to discrete choices: A new foundation for the multinomial logit model. American Economic Review 105, 272–298.

McFaddden, D., 1978. Modeling the choice of residential location. Spatial Interaction Theory and Planning Models , 75–96.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of applied Econometrics 15, 447–470.

Redmond, L.S., Mokhtarian, P.L., 2001. The positive utility of the commute: modeling ideal commute time and relative desired commute amount. Transportation 28, 179–205.

Sifringer, B., Lurkin, V., Alahi, A., 2018. Enhancing discrete choice models with neural networks. 7th Symposium of the European Association for Research in Transportation conference, .

Sims, C.A., 2003. Implications of rational inattention. Journal of monetary Economics 50, 665–690.

Srivastava, R.K., Greff, K., Schmidhuber, J., 2015. Training very deep networks, in: Advances in neural information processing systems. volume 28, pp. 2377–2385.

Theano Development Team, 2016. Theano: A Python framework for fast computation of mathematical expressions. arXiv e-prints abs/1605.02688.

Thorhauge, M., Cherchi, E., Walker, J.L., Rich, J., 2019. The role of intention as mediator between latent effects and behavior: application of a hybrid choice model to study departure time choices. Transportation 46, 1421–1445.

Vythoulkas, P.C., Koutsopoulos, H.N., 2003. Modeling discrete choice behavior using concepts from fuzzy set theory, approximate reasoning and neural networks. Transportation Research Part C: Emerging Technologies 11, 51–73.

Wang, F., Ross, C.L., 2018. Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model. Transportation Research Record 2672, 35–45.

Wang, S., Zhao, J., 2019. Multitask learning deep neural network to combine revealed and stated preference data. arXiv preprint arXiv:1901.00227 .

Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. 4 ed., Morgan Kaufmann, Cambridge, MA.

Wong, M., Farooq, B., 2018. Modelling latent travel behaviour characteristics with generative machine learning, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 749–754.

Wong, M., Farooq, B., 2020. A bi-partite generative model framework for analyzing and simulating large scale multiple discrete-continuous travel behaviour data. Transportation Research Part C: Emerging Technologies 110, 247–268.

Wong, M., Farooq, B., Bilodeau, G.A., 2018. Discriminative conditional restricted boltzmann machine for discrete choice and latent variable modelling. Journal of Choice Modelling 29, 152–168.

Yazdizadeh, A., Farooq, B., Patterson, Z., Rezaei, A., 2017. A generic form for capturing unobserved heterogeneity in discrete choice modeling: Application to neighborhood location choice, in: Transportation Research Board 96th Annual Meeting, pp. 17–05144.