Some Considerations about Mode Choice Model

Xiao-Yun Lu

California PATH
Institute of Transportation Studies
University of California at Berkeley
Richmond Field Station, Bldg. 452
1357 S. 46th Street
Richmond, CA 94804-4648

Tel: (510) 665-3644
e-mail: xylu@Path.Berkeley.edu

January 1st, 2007
Abstract

This paper discusses the representability of discrete logit-type models including multinomial logit and nested logit model from a mathematical approach. It is shown that the logit-type models can be reconstructed from mathematical approximation theory with sigmoidal functions widely used in Neural Network modeling without the basic assumptions such as IIA and iid, and the distribution (or density) function of the unobserved portion of utility. This explains mathematically why logit-type models can approximate the choice probability function to some accuracy. It is hoped that this may suggest the way to improve the accuracy in model specification for logit type models.
Introduction

Discrete choice model (McFadden, 1974; Train, 2002, Bhat, 2003) is popularly used for modeling the probability for human choice behavior among finite number of alternatives. This model has been used in economic system for customer product choice, in transportation system for passenger/goods movement mode choice, or in other similar systems related to decision makings. The discrete models were primarily built based on some economic and human behavior related reasoning and assumptions. Due to those assumptions, the probability distribution function can reach in closed forms with unknown parameters. Those parameters are then determined based on survey data. From a mathematical modeling perspective, the essential function of the discrete choice model is an approximation of the probability distribution function.

In transportation planning, the application of discrete model to traveler behavior for mode choice is important to quantitative evaluation of proposed projects or policies for improving intermodal connectivity of transit system. In order to develop projections which can encourage the use of intermodal facilities and services for improved connectivity and could potentially relieve increased traffic, it is necessary to predict how transit passengers will change their travel patterns in response to any changes in major factors such as mode, connection, service schedule and fare of transportation providers in the system.

A popularly used approach is the logit-type model (Multinomial Logit and Nested Logit) or other deduced types, which belong to the GEV family (Train 2002). Data from transit passenger surveys at selected case study and at fixed time are usually used to calibrate logit-type models (to determine the parameters). This approach has been
applied to intermodal airport ground access planning (Gosling and Lu, 2007; Lu and Gosling, 2007). Main challenges to the discrete model application are (a) how robust the model with respect to the uncertainties in the survey data; and (b) the accuracy of the models for predicting the passenger mode choice behavior subjected to changes of other factors (mode, service, time period and region) in the system. The former is commonly referred to as the “representability”, and the latter the “generality” of the model. Our experience shows that adequate attention need to be paid to those fundamental issues in future research and planning practice. This paper will look at preliminarily a from a different angle the representability of the model.

In systems modeling and approximation, another modeling approach has been developed widely and in parallel, which is called Neural Networks (NNs). However, the model development is in a quite different approach. It is completely based fundamental mathematical approximation theory from Stone-Weierstrass Theorem. Cybenko (1989) showed that sufficient number of layers of finite linear combinations of sigmoidal functions can be used to approximate any continuous functions of N variables to an arbitrary order of accuracy. The model starts from a closed form which is continuously differentiable or even smooth with unknown parameters. The parameters are also to be determined by data from real system in consideration. As stated in Cotter (1990), the powerful results from NN have two desirable properties: (a) larger networks can produce lesser error than smaller networks; and (b) there are no nemesis functions that cannot be modeled by NNs. Particularly, all the continuous functions can be modeled (approximated) by NNs. It is thus expected that probability distribution function in
discrete modeling should also be represented with NNs. This is the motivation to look at the representability from this angle.

NNs are used in two ways for modeling a system: (a) to model a functional relationship between independent variables and dependent variables; (b) to model a dynamical systems which can be represented by a set of ordinary differential equations (ODEs). The former is static modeling in the sense that the time is fixed, which is similar to the modeling of human choice behavior using the discrete model and cross-sectional data. Typical example of application for static modeling is in pattern recognition (Haykin, 1994) and classification. The latter is to represent the right hand side of a set of ODEs using NNs, which is to establish a relationship between the changing rate of the dependent variables and all the dependent variables and, possibly, time. From functional viewpoint, a static modeling (functional relationship) can be imbedded in the dynamic system for fixed initial condition and fixed time. Typical example of application NN to dynamical systems is the modeling of control systems (Cichocki and Unbehauen, 1993; Narendra and Parathasarathy, 1990; Miller et al, 1990; Lu and Spurgeon, 1998). It is noted that a control system can be described as ODEs with control parameter(s), which is essentially a bundle of ODEs.

Both discrete choice model and NN model were caused attention in previous work for modeling and prediction in transportation system planning. Gavidia (2001) and references therein compared both models for modeling and prediction including professional attainment. Denton (1995) compared the two models to see which one performs better under the conditions of (a) sufficient data set available; and (b) insufficient data set available. It concluded that, under less than ideal conditions (e.g.
when the model was not correctly specified, existence of missing data, outliers, heteroscedasticity, or autocorrelation), NN model performance was superior. Nijkamp (1997) used logit model and the feed-forward neural network model to conduct case study on the investigation of the competition effects of two main transport alternatives: rail and road, for predicting the ridership of the high-speed train in Italy. Two dependent parameter variations were considered: (a) ‘time’ attributes only; and (b) both ‘time’ and ‘cost’ attributes. They both showed high probabilities of choosing the improved rail transport mode. The Feed-forward NN model seemed to provide reasonable predictions compared to those obtained by means of a logit model. Mohammadian and Miller (2002) used the nested logit model and the multilayer perceptron artificial NNs for modeling and predicting household vehicle choice. It was shown that both methods generated strong results, although the multilayer perceptron artificial NNs yielded better predictive potential. In the work of Abdelwahab and Abdel-Aty (2001), multilayer perceptron NN model and logit model were used for modeling and analysis of two-vehicle accidents that occurred at signalized intersections. The NN had a better generalization performance of 65.6 and 60.4 percent for the training and testing phases, respectively. The logit model was able to correctly classify only 58.9 and 57.1 percent for the training and testing phases, respectively. Those works inspired the motivation to improve the modeling and prediction capability of the discrete model from mathematical point of view.

This paper will preliminarily discuss some mathematical link between the discrete choice model and the NN model, which has not been considered in previous work. It shows that logit type models can be reconstructed as composite function with sigmoidal functions and other smooth functions as in NN modeling. The representability of NN
model based on mathematical approximation thus explains why the discrete model can represent probability distribution function for choice behavior. It is hoped that this may suggest the way to potentially improve robustness and accuracy of discrete choice model in the model specification based on what we learn form NN approach.

The paper is structured as follows. Section 2 presents the two approaches for modeling: discrete modeling and NN model and the reconstruction of logit-type modeling from NN approach, and some fundamental properties of sigmoidal functions. Section 3 discusses the representability of logit-type models from mathematical and NN approach, coefficient determination methods of the two approaches, and the comparison of the two approaches. Section 4 presents some concluding remarks.

2. Two Viewpoints on Mathematical Modeling

Cybenko (1989) showed that sufficient layers of finite linear combinations of sigmoidal functions can be used to approximate any continuous functions of N variables to arbitrary order of accuracy. In the special case of choice probability distribution function, it can also be approximated by such layered network structure of sigmoidal functions. This explains from a purely mathematical viewpoint why the discrete model could be used to represent probability distribution for human choice behavior.

2.1 Discrete Choice Modeling

The utility function for alternative $i$ is

$$ U^{(i)} = a^{(i)} x^{(i)} + d^{(i)} $$

(2.1)
where $x^{(i)} = \left[ x_1^{(i)}, ..., x_{s^{(i)}}^{(i)} \right]^T$ is the column vector of independent variables represents attributes for alternative $i$ in modeling (critical factors in consideration that affect decision maker’s choice behavior);

$U^{(i)}$ - the utility function corresponding to mode $i$;

$s_i$ - the number of coefficients in utility function $U^{(i)}$;

$M$ - the number of primary alternatives available, which is known;

$a^{(i)} = \left[ a_1^{(i)}, ..., a_{s^{(i)}}^{(i)} \right]$ - the row vector of unknown linear coefficients corresponding to $x^{(i)}$;

$d^{(i)}$ - the unknown constant in the utility function for alternative $i$, the offset.

All the coefficients $a^{(i)}$ and $d^{(i)}$ are to be determined from practical data.

For self consistence, some well-known logit-type models are presented here in a form convenient for next discussion. In particular, this paper will concentrate on the following two logit models:

- Mutinomial Logit
- Nested Logit

Under the following assumptions, one can reach a closed form representation of the probability function with respect to the attributive (independent) variables.

(a) IIA: Independence from Irrelevant Alternatives, which leads to the proportional substitution across alternatives. It can be understood as either a restriction imposed by the model or as the natural outcome of a well-specified model that captures all sources of
correlation over alternatives into representative utility such that only white noise remains. IIA needs and can be tested from data if it is satisfied to use the model. Mathematically, it can be interpreted as the ration of the probability for choosing one alternative is independent of other alternatives. For Mutinomial Logit, IIA is required to hold among all the alternatives. For Nested Logit model, IIA is assumed to hold within each nest, but does not hold for two alternatives between nests in general.

(b) Distribution of unobserved portion of utility: For Multinomial Logic Model, it requires to be univariate and independent extreme value error structure (Bhat, 2003). For GEV (Generalized extreme value): unobserved portions of utility for all alternatives have a generalized joint extreme value. As a special case of GEV, the Nested Logit model has this property within a nest but independent between nests.

(1) Multinomial Logit Model

For the utility function in (2.1), the probability for decision maker to choose alternative \( i \) is determined as

\[
P^{(i)}(x) = \frac{e^{U^{(i)}(x^{(i)})}}{\sum_{k \in M} e^{U^{(k)}(x^{(k)})}}
\]  

(2.2)

where \( P^{(i)} \) is the probability for alternative \( i \) to be chosen.

(2) Nested Logit Model
Suppose the alternatives are divided into disjoint nests. $N_m$ is the set of indices of modes in nest $m$. The conditional probability for a decision maker to choose alternative $i$ among nest $m$ is

$$P^{(i|m)} = \frac{\left( \sum_{k \in N_m} \left( \frac{e^{U^{(i)} / \mu_i}}{\mu_k} \right) \right)^{\mu_m}}{\sum_{l \in G_{nest}} \left( \sum_{k \in N_l} \left( \frac{e^{U^{(k)} / \mu_k}}{\mu_l} \right) \right)^{\mu_l}} \quad (2.3)$$

which is basically a multinomial logit model within a nest $m$. The marginal probability of an alternative chosen by the decision maker falling in nest $m$ is

$$P^{(m)} = \frac{\left( \sum_{j \in N_m} \left( \frac{e^{U^{(j)} / \mu_i}}{\mu_m} \right) \right)^{\mu_m}}{\sum_{l \in G_{nest}} \left( \sum_{k \in N_l} \left( \frac{e^{U^{(k)} / \mu_k}}{\mu_l} \right) \right)^{\mu_l}} \quad (2.4)$$

where $P^{(m)}$ is the marginal probability that a passenger will choose a mode in nest $m$ in a Nested Logit model (i.e. the probability that the chosen mode falls in nest $m$); $G_{nest}$ is the set of indices of all nests (suppose only consider one layer of nest); $\mu_m$ is the nest-related constant unknown parameter to be determined. The microeconomic meaning of the parameter $\mu_m$ has been explained in (Train, 2002). From Bayesian Theorem, the probability for alternative $i$ to be chosen is (Train, 2002)
It is obvious that Multinomial Logit model is a special case of Nested Logit model where each nest contains only one alternative.

2.2 Mathematical Reasoning from NN Approach

It is interesting to investigate the representability of logit-type model from a purely mathematical viewpoint. Particularly, it is necessary to answer the following questions:

- Why they can be used to represent a probability distribution function?
- How accurate they can represent the probability distribution function?
- Can one specify the model without the extreme value distribution assumption for the unobserved portion of utility?

To get the answers, it is necessary to show that the logit-type models can be reconstructed from sigmoidal functions as in a NN approach. As a preparation, it is necessary to investigate the fundamental properties of the sigmoidal functions.

2.2.1 Sigmoidal Functions

The following terminologies used for discussion later on are referred to any standard textbook in Mathematical Analysis such as Luthar (2005).
A function can be considered as a mapping between sets. Let \( f(\cdot) \) be a mapping between sets \( D \subset \mathbb{R}^n \) and \( S \subset \mathbb{R} \), denoted as \( D \rightarrow^{f(\cdot)} S \).

\( f(\cdot) \) is surjective if, for arbitrary \( y \in S \), there exists \( z \in D \) such that \( y = f(z) \);

\( f(\cdot) \) is bijective if, for arbitrary \( z_1, z_2 \in D \), \( f(z_1) = f(z_2) \) if and only if \( z_1 = z_2 \);

If \( g(\cdot) \) is defined in an interval on real line: \( D = [t_0, t_f] \subset \mathbb{R} = [-\infty, \infty], t_0 < t_f \), \( g(\cdot) \) is monotone increasing, if \( t_1, t_2 \in D, t_1 \leq t_2 \), then \( g(t_1) \leq g(t_2) \); and \( g(\cdot) \) is strictly monotone increasing, if \( t_1 < t_2 \) implies \( g(t_1) < g(t_2) \).

The following properties are true:

(i) A strictly monotone increasing function is bijective;

(ii) If \( f(\cdot) \) and \( g(\cdot) \) are bijective, then the composite mapping (function) \( f(g(\cdot)) \) is bijective.

A Sigmoidal function \( \sigma(t) \) is strictly monotone and continuous and satisfying

\[
\sigma(t) = \begin{cases} 
1, t \rightarrow \infty, \\
0, t \rightarrow -\infty.
\end{cases}
\] (2.6)

Cybenko (1989) showed that sufficient layers of finite linear combinations of sigmoidal functions can be used to approximate any continuous functions of \( n \) variables to an arbitrary order of accuracy. Such a type of function is called a neuron in Neural Network Theory and Application. Examples of sigmoidal functions include the following:

Example 1.
\[ \sigma(t) = \frac{1}{1 + e^{-t}}, \quad t \in (-\infty, \infty) \]  

(2.7)

Example 2.

\[ \sigma_1(t) = \frac{2}{1 + e^{-t}}, \quad t \in (-\infty, \infty) \]  

(2.8)

Example 3.

\[ \tanh(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}}, \quad t \in (-\infty, \infty) \]

\[ \sigma_2(t) = \frac{1 + \tanh(t)}{2}, \quad t \in (-\infty, \infty) \]  

(2.9)

In fact, they are also bijective and surjective between \( \mathbb{R} \) and [0,1].

2.2.2 Properties of Sigmoidal Function

Some basic properties of sigmoidal functions are presented and proved here, which will be used for later discussion.

**Property 1.** Let \( \phi_j(\cdot), j = 1, \ldots, n \) be sigmoidal functions, \( \alpha_j \in \mathbb{R}, \alpha_j \geq 0, j = 1, \ldots, n \) and \( \sum_{j=1}^{n} \alpha_j = 1 \), then

\[ \psi(\cdot) = \sum_{j=1}^{n} \alpha_j \phi_j(\cdot) \]

is a sigmoidal function.

**Proof.** This is clearly true that the property (2.6) holds. It is sufficient to prove that \( \psi(\cdot) \) is strictly monotone increasing. This is also obvious since

\[ \psi(t_1) - \psi(t_2) = \sum_{j=1}^{n} \alpha_j \left[ \phi_j(t_1) - \phi_j(t_2) \right] \]

has the same sign as \( (t_1 - t_2) \) noticing that \( \alpha_j \geq 0, j = 1, \ldots, n \) and \( \sum_{j=1}^{n} \alpha_j = 1 \).
This property states that the convex hull of finite number of sigmoidal functions is a sigmoidal function.

**Property 2.** Let $\phi_j(\cdot), j = 1, \ldots, n$ be sigmoidal functions, then

$$
\psi(\cdot) = \prod_{j=1}^{n} \phi_j(\cdot)
$$

is a sigmoidal function.

**Proof.** It is sufficient to prove the case of $n = 2$ by induction. It is clear that $\phi_1(\cdot) \cdot \phi_2(\cdot)$ is strictly monotone and continuous and that

$$
\psi(t) = \begin{cases} 
1, & t \to \infty, \\
0, & t \to -\infty.
\end{cases}
$$

if both $\phi_1(\cdot)$ and $\phi_2(\cdot)$ have the same properties. Thus $\psi(t)$ is a sigmoidal function.

**Property 3.** Let $\phi(\cdot)$ be a sigmoidal function, then $\psi(\cdot) = \phi^\delta(\cdot)$ is a sigmoidal function for any $\delta \in \mathbb{R}, \delta > 0$.

**Proof.** If $\phi(\cdot)$ is strictly monotone increasing, then $\phi^\delta(\cdot)$ is also strictly monotone increasing and the property (2.6) holds.

### 2.2.3 Neural Network Model Construction

To reconstruct the logit-type model in an NN approach, it would be necessary to understand how an NN model is constructed from sigmoidal function(s). NNs are
developed as generalizations of mathematical models of human cognition, based on the following assumptions as summarized in the work of Fausett (1994): (1) information processing occurs at many simple elements called neurons or neural processing elements (analogous to the neural systems of human brain); (2) signals are passed between neurons over connection links; (3) each connection link has an associated weight, which, in a typical neural net, multiplies the signal transmitted; and (4) each neuron applies an activation function (usually nonlinear) to its net input (sum of weighted input signals) to determine its output signal. The network structure in Figure 1 better explains those considerations.

A NN model with $M$ layers assumes the following general form:

$$F^{(j)} = \phi_j \left( W^{(j)} F^{(j-1)} + b^{(j)} \right)$$

$$j = 0,1,\ldots,M$$

$N$ - the number of layers;

$L_j$ - the number of neurons (sigmoidal functions) in layer $j$;

$\phi_j(\cdot)$ - sigmoidal function in layer $j$; For vector input $z = [z_1, \ldots, z_p]^T$, it is understood as:

$$\phi_j(z) = \left[ \phi_j(z_1), \ldots, \phi_j(z_p) \right]^T$$

$F^{(j)} = [F^{(j)}_1, \ldots, F^{(j)}_{L_j}]^T$ - the intermediate functions in layer $j$;

$W^{(j)} = [w^{(j)}_1, \ldots, w^{(j)}_{L_{j-1}}]$ - the synaptic weights (coefficients) in layer $j$;

$b^{(j)} = [b^{(j)}_1, \ldots, b^{(j)}_{L_{j-1}}]^T$ - the bias or off-set of neurons in layer $j$, $(j = 1,\ldots,M)$;

$F^{(0)} = x$ - the input layer – the layer of independent variables;
$F^{(N)}$ - the vector of output functions;

Figure 1 depicts a typical neural network model structure. Each block represents a Neuron – a sigmoid function. From lower layer to high layer, a linear combination of the lower layer neurons is taken as the input to the next high layer. Usually, the same sigmoid function is used in the same layer, but this is not a restriction according to the work of Cybenko (1989). It is clear that such a model is smooth and even analytic.

![Figure 1. A typical Neural Network model structure](image)

*Here the arrow lines linking lower layer Neurons with upper layer Neurons means linear combinations of all the Neurons in lower layer is taken as input of the Neurons in upper layer.*

For the modeling of a general function of N variables, it is not necessary to specify and particular structure for the last layer. However, for the modeling of probability distribution function, proper specification of the last layer structure could
possibly simplify the parameter determination procedure and generate more robust results.

3. Mathematical Representability of Logit-type Models

After the preparation above, it is ready show how the Logit-type models are reconstructed from the NN approach.

3.1 Reconstruction of Multinomial Logit Model

Let \( t = -\ln(e^{-y_2} + \ldots + e^{-y_N}) \). Then

\[
\frac{1}{1 + e^{-t}} = \frac{1}{1 + e^{-y_2} + e^{-y_3} + \ldots + e^{-y_N}}
\]

Let

\[
y_i = z_i - z_1, i = 2, \ldots, N
\]

Then

\[
\frac{1}{1 + e^{-t}} = \frac{1}{1 + e^{-y_2} + e^{-y_3} + \ldots + e^{-y_N}} = \frac{1}{1 + e^{-(z_2 - z_1)} + e^{-(z_3 - z_1)} + \ldots + e^{-(z_N - z_1)}} = \frac{1}{e^{-z_1} + e^{-z_2} + e^{-z_3} + \ldots + e^{-z_N}}
\]

Now let \( z_i = U_i, i = 1, \ldots, N \) be the utility function, we obtain the multinomial model (2.3).

It shows how the multinomial logic model can be reached by combination of sigmoidal function with some other mappings using the Chain Rule.

Let’s look at the process in details in terms of a series of mappings (transformations) used and relevant domains. For convenience of discussion, let \( y = [z_2, \ldots, z_N]^T \in \mathbb{R}^{N-1} \) and
\[ z = [z_1, \ldots, z_N]^T \in \mathbb{R}^N. \] Also using \( \mathbb{R} = [-\infty, \infty] \) denote the generalized real field, i.e. to include the point at infinity. Let \( \mathbb{R}_+ = [0, \infty] \). It is easy to see that

1. \( y = I(z) : y_i = z_i; y_i = z_i - z_i, i = 2, \ldots, N : \mathbb{R}^N \rightarrow \mathbb{R}^N \) is bijective and surjective;

2. \( w = \zeta(y) = e^{-y_2} + \ldots + e^{-y_N} : \mathbb{R}^{N-1} \rightarrow \mathbb{R}_+ \) is also surjective;

3. \( t = -\ln(w) : \mathbb{R}_+ \rightarrow [-\infty, +\infty] \) is bijective and surjective;

4. \( P = \sigma(t) = \frac{1}{1 + e^{-t}} : [-\infty, +\infty] \rightarrow [0,1] \) is strictly monotone increasing (and thus bijective) and surjective; It is also a sigmoidal function.

Thus the following composite mapping is surjective.

\[
\begin{align*}
z & \xrightarrow{I(.)} y \xrightarrow{\zeta(.)} w \xrightarrow{-\ln(.)} x \xrightarrow{\phi(.)} P
\end{align*}
\]

which means that the multinomial logit model considered as a mapping from the dependent variables to \([0,1]\) can be decomposed into several surjective mapping and a sigmoidal mapping. The composite mapping (function) can be written as

\[
P = \phi\left[-\ln\left(\zeta\left(I(x)\right)\right)\right] : \mathbb{R}^N \rightarrow [0,1]
\]

### 3.2 Nested Logit Model Reconstruction from NNs

A nested logit model (2.5) can be decomposed into the production of two parts with each part similar to a logit model structure (Train, 2002). Let

\[
e^{t_w} = \left(\sum_{j=N_u}^N \left( e^{U_j} \right)^{1/\mu_u} \right)^{\mu_u}
\]

18
Then the second part in (2.5) becomes

\[ p^{(m)} = \frac{e^{J_m}}{\sum_{l \in G_{out}} e^{J_l}} \]  

which is a logit-type with the following transformation:

\[ I_m = \mu_m \cdot \ln \left( \sum_{j \in N_m} e^{U(j)/\mu_m} \right) \]  

and the probability is

\[ p^{(i)} = p^{(m)} p^{(m)} = \left( \frac{e^{J_m}}{\sum_{l \in G_{out}} e^{J_l}} \right) \left( \frac{e^{U(i)/\mu_i}}{\sum_{k \in N_m} e^{U(k)/\mu_k}} \right) \]  

Now both the conditional probability and the marginal probability function have the structure similar to the multinomila logit model. With similar argument as above, it can be concluded that they can be represented as sigmoidal function in composite with some surjective and bijective function.

Besides, according to Property 2, the multiplication of two sigmoidal functions is also sigmoidal. Thus (3.5) indicate that a nested logit model can be decomposed into the composite of a finite number sigmoidal functions and 1-1 surjective functions.

3.3 Representability of Logit-type Models

As is shown above, all the logit-type functions can be obtained from sigmoidal functions in conjunction with some other surjective and bijective mappings. It is well-known mathematically that 1-1 surjective mapping will not change the representability of a function. Based on the results of Cybenko (1989), we can reach the following conclusion:
(i) Logit-type models can be used to approximate probability distribution functions to some order of accuracy from mathematical point of view;

(ii) This also suggests that in order to improve the modeling accuracy, one can add more layers of sigmoidal functions in model specification as in an NN approach.

It is noted that, although the sigmoidal function has been used twice in the nested logit model, they are not used in a layered (composite structure). Thus, one could not conclude from it that the nested logit model can represent the probability function with higher order of accuracy. It is also noted that the spatial extension of the Multinomial Logit model proposed by Mohammadian and Kanaroglou (2003) could also be interpreted this way.

With this approach, the only assumption necessary is that the human choice behavior has some probability distribution. All the other assumptions used in the discrete model approach deduction (Train, 2002) can be dropped. Obvious advantages for this viewpoint are the naturally closed forms and the flexibility in specifying the models.

3.4 Comparison of the Two Approaches

The following table summarizes the advantages and disadvantages of the two approaches.

<table>
<thead>
<tr>
<th>Modeling Approach</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete Modeling</td>
<td>• Econometric meaning of the function structure is clear;</td>
<td>• to obtain closed form for probability density function for the choice,</td>
</tr>
<tr>
<td></td>
<td>• Econometric meaning of the coefficients are clear;</td>
<td>either of the following conditions are necessary:</td>
</tr>
<tr>
<td></td>
<td>• Alternative choice corresponds</td>
<td>a) IIA or IID assumptions</td>
</tr>
</tbody>
</table>
Approach to utility maximization;

- Linear parameters in the utility leads to global concave of log-likelihood function which is good for fitting methods such as the Least Squares.

Necessary

b) or the distribution (or density) function of the unobserved part in the utility function is known;

- Those assumptions are not necessarily hold or difficult to check if they hold from data;
- Need to assume the extreme value distribution of the unobserved part in utility function;
- Lack of flexibility in model specification;
- The model cannot approximate the probability function to arbitrary order of accuracy.

Neural Network Modeling Approach

- Completely based on function approximation theory in mathematics;
- No restrictive assumptions are necessary;
- Function form structure is very flexible based on the chain rule of composite function.
- The model can approximate the probability function to arbitrary order of accuracy if more layers are specified in the model.

- Physical meaning of the function structure is not clear;
- Physical meaning of the coefficients are not clear;
- Alternative choice probability does not necessarily correspond to utility maximization;

3.5 New Models Specification

Based on the above discussion, one may improve the accuracy of the logit-type models by increasing the number of layers. Here the layer means the layers in the composite function similar to the structure of NN model in (2.10). It is completely arbitrary to specify the additional layers. As a special case, the multinomial logic model can be generalized as follows. For the utility function in (2.1), let

\[ V^{(j)}(x) = \sigma\left( \sum_{k=1}^{M} c_{k,j} U^{(k)}(x^{(k)}) + d_j \right) \] (3.8)
where $\sigma(\cdot)$ is a sigmoidal function. $C = [c_{i,j}]_{M \times M}, d = [d_1, \ldots, d_M]$ are to be determined by data. The probability that passengers choose mode $i$ can be determined by the following generalized multinomial logit model.

$$P^{(i)} = \frac{\left( e^{V^{(i)}} \right)}{\sum_{k \in M} \left( e^{V^{(k)}} \right)} \quad (3.9)$$

For improving modeling accuracy, one can also add more layers similarly.

For nested logit model, one can add a neural network layer to the conditional probability and/or the marginal probability.

(1) One way is to modify the conditional probability expression by adding a layer as follows leave the marginal probability unchanged. Let

$$V^{(j)}(x) = \sigma\left( \sum_{k \in N_m} c^{(m)}_{k,j} U^{(k)}(x^{(k)}) + d^{(m)}_j \right)$$

The conditional probability becomes

$$Q^{(j,m)} = \frac{\left( e^{V^{(j)}/\mu} \right)}{\sum_{k \in N_m} \left( e^{V^{(k)}/\mu} \right)}$$

$$Q^{(m)} = \frac{e^{V^{(j)}}}{\sum_{k \in G_{nest}} e^{V^{(k)}}}$$

$$L^{(k)} = \mu_m \cdot \ln \left( \sum_{j \in N_m} e^{V^{(j)}/\mu_m} \right)$$

$$P^{(i)} = Q^{(i,m)}$$

where $A = [c^{(m)}_{k,j}]$ an $N_m \times N_m$ matrix of unknown parameter, and $d_j^{(m)}, (j = 1, \ldots, N_m)$ are related to nest $m$, which are to be determined by the data. Then $Q^{(i)}$ is the marginal probability for mode $i$ to be selected. Mathematically, this is equivalent to improving the modeling accuracy of the conditional probability.
(2) A similar way is to add a layer to modify the marginal probability expression as follows and keep the conditional probability part unchanged:

\[ L^{(j)} = \sum_{k \in G_{nest}} \sigma \left( r_{k,j} f^{(k)} + h_j \right), \quad j \in N_{nest} \]

where \( R = \left[ r_{k,j} \right]_{G_{nest} \times G_{nest}} \) and \( H = [h_1, \ldots, h_{N_{nest}}] \) are unknown parameters to be determined by data. Now the marginal probability and final probability to choose alternative \( i \) becomes

\[ Q^{(m)} = \frac{e^{L^{(i)}}}{\sum_{k \in G_{nest}} e^{L^{(k)}}} \]
\[ P^{(i)} = Q^{(m)} P^{(jm)} \]

Mathematically, this is equivalent to improving modeling accuracy of the marginal probability.

(3) One may also combine those two methods to add a layer or more layers to both marginal and conditional probability distribution to improve their modeling accuracy.

**Remark 3.1** To add a layer or multiple layers to either marginal probability or conditional probability, or both using the sigmoidal functions in model specification will cause some coupling between the utility function of different alternatives. In general, it mixes up the attributes of different alternatives and different nests. The physical meaning of the original model structure disappears. It makes the model more complicated and may cause it to lose concaveness of the corresponding log-likelihood function even if the utility function in the first layer is linear in parameter. However, one gains the flexibility in model specification and improves accuracy in approximation. Besides, the decoupling (independence between alternatives) is left to coefficient determination process from data. This makes sense because the practical data for mode choice model are usual cross-
sectional instead of completely reflect the mode choice individual’s preference (Bhat, 1998). Besides, coefficient determination in NN approach does not require the Log Likelihood function to be concave for global minimum.

### 3.6. About Coefficient Determination

#### 3.6.1 Different Approaches for Parameter Estimation

Numerical methods for determining the unknown coefficients of the models are usually some curve fitting techniques which involve optimization procedures. Parameter estimation of logit-type model is to maximize the log-likelihood function which is concave in case the utility is linear in parameters and thus has a unique minimum:

In neural networks, the parameter determination (or train) process is a typical unconstrained optimization of the sum squares errors of the model output and the system (plant) output from data

\[
J^{(i)} = \frac{1}{2} \sum_{k=1}^{K} \left( P^{(i)}(k) - P_{d}^{(i)}(k) \right)^2
\]  

(3.10)

where \( K \) is the total number of data point and \( P_{d}^{(i)}(k) \) is obtained from data directly. It is noted that this is a multiple objective function optimization problem. This can be converted into single objective function by setting

\[
J = J^{(1)} + \ldots + J^{(M)}
\]

The objective function or its log-likelihood function is usually not concave. The optimization method used are usually genetic algorithm, steepest decent, conjugate gradient method, and modified Newton’s Methods, etc. (Cybenco, 1989; Gupta and Rao, 1994; Haykin, 1994). Particularly, the Genetic Algorithm does not require unique global minimum.
3.6.2 About Concaveness of the Generalized Logit Models

To achieve global minimum, it would be ideal for the objective function to be concave. It is known that, the original multinomial logit model results in globally concave log-likelihood function if the utility is linear in parameters. How about the modified model with added neural network layers? Simply adding some layers in the model specification as above may not leads to globally concave log-likelihood function even if the utility function in the first layer is linear in unknown parameters.

**Research problem 1:** It would be interesting to investigate how to add some layers to the multinomial logit models to improve its accuracy while keeping the log-likelihood function to be globally concave. If this can be achieved, parameter determination would be easier to achieve a more robust and accurate model.

**Research Problem 2:** It would be interesting to investigate if the GEV family models can be reconstructed this way.

4. Concluding Remarks and Future Work

The discrete models are built based on statistical and economical reasoning and some assumptions for reaching closed forms. The representability of NNs is based on mathematical reasoning for function approximation, eventually on Stone-Weisstrass Theorem. Mathematically, adequate number of layers of sigmoidal functions in a composite structure NNs can approximate a continuous function of $N$ variables to an arbitrary order of accuracy. The logit-type model can be considered as special type of NN model with one layer of sigmoidal function structure. As such, it can represent human choice probability function to some order of accuracy, which is the representability of the
discrete model from mathematical approximation theory. This viewpoint does not rely on any assumption such as IIA or iid or even known distribution of the unobserved portion in the utility function. Those results explains mathematically some cited previous work using two approaches for modeling human choice behaviors for comparison in microeconomic system and transportation system planning. It is hoped that this viewpoint may suggest new ways to improve modeling robustness and accuracy in model specification. However, the specified models need to be validated through practical data, which will be part of the future work.

Acknowledgement

The author is grateful to the anonymous reviewer for the constructive comments.

Reference

Abdelwahab H. T. and Abdel-Aty M. A. (2001), Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections, 1746, Transportation Research Board, Washington, DC 20418


dimensions of travel, 10th International Conference on Travel Behaviour Research
Lucerne, 10-15. August 2003


Luthar, I. S. (2005), *Sets, Functions, and Numbers*, Alpha Science Intl Ltd


Mohammadian, A. and Kanaroglou, P. S., (2003), Applications of Spatial Multinomial Logit Model to Transportation Planning Moving through nets: The physical and
social dimensions of travel, 10th International Conference on Travel Behaviour

Narendra, K. S. and Parathasarathy, K. (1990), Identification and control of dynamical systems using neural networks, *IEEE Trans. on Neural Networks*, 1, no. 1, 4-27
