Latent behaviour modelling using discriminative restricted Boltzmann machines

Melvin Wong\textsuperscript{a}, Bilal Farooq\textsuperscript{a}, Guillaume-Alexandre Bilodeau\textsuperscript{b}

\textsuperscript{a}LITrans, Department of Civil, Geotechnical, and Mining Engineering, Ecole Polytechnique de Montréal, 2500 Chemin de Polytechnique, Montréal, Canada
\textsuperscript{b}LITIV, Department of Computer and software Engineering, Ecole Polytechnique de Montréal, 2500 Chemin de Polytechnique, Montréal, Canada

Abstract

In this study, we explore the use of generative modelling methods to estimate latent variables from choice prior distribution without the conventional use of measurement indicators. A restricted Boltzmann machine design is used to represent latent behaviour factors by analyzing the relationship information between the observed choices and explanatory variables. The algorithm is adapted for a simple discrete choice data and we use a graphical approach to study the semantic meaning from parameter vectors. Our study shows that through exhaustive parameter search, we can extract useful latent information on the behaviour of latent constructs through machine learning methods. Conventional methods of estimating latent behaviour from attitudinal questions are subjective and these survey questions may not always be available. We hypothesize that an alternative approach can be used for latent variable estimation through an undirected graphical model.

Keywords:
Machine learning, generative model, restricted Boltzmann machines, latent variables

1. Introduction

Current applied practice in choice modelling is targeted at drawing conclusion on the mechanism of the stochastic model and not so much about the nature of the data itself, often leading to simple hypothesis assumptions of data relevance and statistical properties of explanatory variables (Burnham and Anderson, 2003). Analyzing data through the statistical properties is generally used for extracting information about the evolution of responses associated with stochastic input variables rather than having good prediction capabilities. On the other hand, algorithmic modelling approaches such as artificial neural networks, decision trees, principal component and factor analysis are based on the ability to predict future responses accurately given future input variables within a ‘black-box’ framework (Breiman et al., 2001). Although improved econometric choice models can be estimated by using hybrid methods which incorporate machine learning into discrete choice analysis to learn the mapping from some latent variable to some posterior distribution (Eric et al., 2008). In this study, we decouple the latent behaviour model underlying the data distribution through exhaustive search without the need for subjective measurement indicators. The proposed method does not predefined a semantic meaning for each latent variable. Instead, we a restricted Boltzmann machine to learn the latent relationships and approximate the posterior probability.

2. Background

A number of different approaches which implements the use of attitudinal variables have been used in existing literature (Ashok et al., 2002; Morey et al., 2006; Hackbarth and Madlener, 2013). The first
approach relies on a top-down modelling framework which makes prior assumptions that individuals are divided into multiple market segments and each segment has its own utility function of underlying attributes. In the most generic form, the assumptions are based on multiple sources of unobserved heterogeneity influencing decisions, e.g. inter- and intra-class variance and “agent effect” (Yazdizadeh et al., 2017). Fig. 1 illustrates the LCM and ICLV model framework which shows the process of deriving latent classes or variables and how it integrates into the structural choice model.

The Latent Class Model (LCM) is one such form which assumes a discrete distribution among market segments (Hess and Daly, 2014). LCM derive clusters using a probabilistic model that describes the distribution of the data. Based on this assumption, similarities within a heterogeneous population are identified through assignment of latent class probabilities. Individuals in the same class share a common joint probability distribution among the observed variables. Under assumption of class independence, the utility is generated with a prior hypothesis from several sub-populations, and each sub-population is modelled separately. The resulting classes are often meaningful and easily interpretable. The unobserved heterogeneity in the population is captured by the latent classes, each of which is associated with different utility vector in the sub-model (Fig. 1a). Another similar class of top-down models are finite mixture models, e.g. Mixed Logit, which allows the parameters to vary with a variance component and that behaviour is dependent on the observable attributes and on the latent heterogeneity which varies with the unobserved factors (Hensher and Greene, 2003).

The use of attitudes and perception latent variables are also particularly interesting and popular in past work (Daly et al., 2012). Choice models with measurement indicator functions treat correlated indicators into multiple latent variables. This factor analysis method is similar to principal component analysis where the latent variables are called principal components. This approach involves the analysis of relationship between indicators and the choice model. Within this domain, there is the sequential and simultaneous estimation process. Sequential approach first estimate a measurement model which derives the relationship between latent variables and indicators. Then, a choice model is estimated, integrating over the distribution of the latent variables. The main disadvantage of this approach is that the parameters may contain measurement errors from the indicator function that were not taken into account during the choice model. To solve this issue, another approach uses simultaneous estimation of structural and measurement model, which includes the latent variable in the choice model framework. This is so called the Integrated Choice and Latent Variable (ICLV) model (Fig. 1b). The ICLV model explicitly uses information from measurement indicators and explanatory variables to derive latent constructs.

This combined structural model framework has led to many interesting results. E.g. environmental attitudes in rail travel (Hess et al., 2013), image, stress and safety attitudes towards cycling (Maldonado-Hinarejos et al., 2014), and social attitudes towards electric cars (Kim et al., 2014). However, the simultaneous approach still relies on a separate measurement model (latent variable model) which describes the relationship to indicators. Despite the direct benefits of the ICLV model combining factor analysis with traditional discrete choice models, the only advantage to using such an approach is when attitudinal measurement indicators are expected to be available to the modeller and the observed explanatory variables are weak predictors of the choice model (Vij and Walker, 2016). Even when measurement indicators are available, they may not provide any further information that directly influence the choice than through explanatory variables (Chorus and Kroesen, 2014). Consequently, mis-specification and other measurement errors may occur, when the criteria is not associated with the choice model. Without measurement indicators to guide selection of latent variables, we can alternatively derive conditions for latent variables through data mining. This can be implemented through generative modelling methods used in machine learning. Generative modelling in machine learning is a class of models which uses unlabelled data to generate latent features. In logistic regression that learn mappings directly from input space, e.g. \( p(y|x) \) are discriminative models. Generative models learn the underlying choice distribution \( p(y) \) and the latent inference \( p(h|y) \),
then using a Bayesian network to derive the posterior distribution of $y$ given $h$ using $p(y|h) = \frac{p(h|y)p(y)}{p(h)}$, where the denominator is given by $p(h) = \sum_y p(h|y = 1)$. The rapid advancement of machine learning research have led to the development of efficient semi-supervised training algorithms such as conditional restricted Boltzmann machine (C-RBM), a hybrid discriminative-generative model, capable of simultaneously estimating a latent variable model using a priori choice distribution with an latent inference model (see Fig. 2).

To date, machine learning and econometric models are often studied for its contrasting purposes in decision forecasting by behavioural researchers (Breiman et al., 2001). The first, deriving from econometric models, are based on the classical decision theory that individual’s decisions can be modelled rationally based on utility maximization. These models assume that the population will adhere to the strict formulation of the choice model but may not always represent the true decisions. The second, generative model based approach, use clustering and factor analysis developed through algorithmic modelling of data. Associations between decision factors can be classified in this method, obtaining latent information without explicit definition of latent constructs (Poucin et al., 2016). Thus, machine learning algorithms such as artificial neural networks that decouple latent information from ‘true’ distribution generally outperform traditional regression based models in multidimensional problems (Ahmed et al., 2010). Recent works regarding hybrid methods on choice based analysis agree on the potential of improving behaviour models with machine learning. Examples include combining machine learning to improve complex psychological models (Rosenfeld et al., 2012), representing the phenomena of similarity, attraction and compromise in choice models (Osogami and Otsuka, 2014) and inference of priorities and attitudinal characteristics (Aggarwal, 2016).

Despite the many improvements, interpretation of results are still extremely difficult due to the complexity and number of parameters. As a result, machine learning models are not often used for general purpose behaviour understanding but created exclusively for a specific purpose for prediction accuracy. Still, machine learning is a rapidly growing field at the intersection is statistics and data science to find patterns in complex data (Donoho, 2015). Furthermore with the emphasis on real practical examples rather than theoretical hypothesis in today’s massive information driven industry, improving analytical techniques with machine learning is very relevant, although statistics and probability theory will still remain important.
2.1. The basis of latent class and latent variable models

The Latent class model is a simple top-down model approach that imparts generalization properties to the choice model that places a priori discrete number of classes, allowing the parameters to vary with an fixed distribution, while the Mixed Logit model takes a continuous distribution across population. Formally, the LCM choice component can be expressed as:

\[ P(y) = \sum_{n} P(S_n)P(y|x, S_n) \]  

where \( S = [s_1, s_2, ..., s_n] \) are the set of classes and \( P(S_n) \) is the probability that an individual belongs to class \( s \).

The ICLV model extends the choice model by describing how perceptions and attitudes affect real choices as well as using separate indicators to estimate latent variables (Ben-Akiva et al., 2002). Latent variables can be classified as either attitudinal (individual characteristics) or perceived (personal beliefs towards responses) (Ben-Akiva and Bierlaire, 1999). The latent variable model (measurement model) forms a sub-part of the structural framework which captures the relationship between the latent variables and indicators and the observed explanatory variables which influence the latent variables. This specification can be used to identify more useful parameters and predict accurate decision outcomes when there is lack of strong significant correlation between explanatory variables and choice outcomes. The functions of the structural and measurement model can be explained in four equations (Vij and Walker, 2016):

\[ x^* = Ax + \nu \]  

\[ I^* = Dx^* + \eta \]  

\[ u = Bx + Gx^* + \epsilon \]  

\[ y_i = \begin{cases} 1 \text{ if } u_i > u'_i \text{ for } i \in \{1, ..., I\} \\ 0 \text{ otherwise} \end{cases} \]
where $A$ represents the relationship between input explanatory variables $x$ and latent variables $x^*$. $D$ represents the relationship between $x$ and the indicator output $I^*$. $B$ and $G$ represents the model parameters with respect to the observed and latent variables. $\nu$, $\eta$ and $\epsilon$ are the stochastic components of the model, assumed to be mutually independent. In a generative model, parameters are shared between $G$ and $D$ that simply defines the joint distribution of $p(y, h)$, i.e. $G = D^\top$ (Fig. 2).

### 2.2. Modelling through generative machine learning methods

In generative machine learning models, hidden units $h$ are the learned features which performs non-redundant generalization of the data. This generalization represents a higher level of complexity from low level, high-bandwidth data (Hinton and Salakhutdinov, 2006). Intuitively, in terms of econometric analysis, hidden units are arbitrary variables that depend on some observed data, for instance, socio-economic attributes such as weather or price information or direct choices such as location and choice of purchase.

We can construct a generative model as a function of these dependent and independent attributes. In the case of factor analysis approach, a common process is to perform feature extraction based on statistical hypothesis testing to determine if the values of the two classes are distinct, for example, using Support Vector Machines (SVMs) or Principal Component Analysis (PCA) to learn low-dimensional classes by capturing only significant statistical variances in the data. The learned classes (or clusters) can then be introduced directly into the model via parameterization. In generative modelling approach, we use the priors directly to learn the distribution of the hidden units. This process we extract latent information directly from the observed choice data instead of using measurement functions which may be prone to errors.

### 2.3. Balancing model inference and accuracy

One common problem that researchers face when constructing latent behaviour models is specifying of the optimal size of latent factors. Since the hypothesis on the number of latent size cannot be tested directly, typical statistical evaluation methods such as AIC and BIC are used to guide class selection. In the case of LVM, through predefinition of measurement function. However, since the number of latent factors determines the capacity of the model to represent the various heterogeneity in the data, it is likely that as we increase $h$, the model become better in capturing complex behaviour effects from individual and choice attributes. On the other hand, if we increase the number of latent segments, the number of parameters will also increase at an exponential rate. Therefore, we gain model accuracy but we would lose model interpretability. The trade-off between inference and accuracy is a challenge when dealing with complex data.

### 3. Generative modelling approach

In this section, we provide a brief overview on restricted Boltzmann machines and how it generates prior over the choice distributions. This is not an exhaustive literature review and we refer readers to (Goodfellow et al., 2016) for more details on generative models and deep learning.

#### 3.1. Restricted Boltzmann machines

A restricted Boltzmann machine (RBM) is an energy-based undirected graphical model that extends from a Markov Random Field distribution by including hidden variables (Salakhutdinov et al., 2007). It is a single layer artificial neural network with no internal layer connections. The model has stochastic visible variables $y \in \{0, 1\}^I$ and stochastic hidden variables $h \in \{0, 1\}^J$. The joint configuration $(y, h)$ of visible and hidden variables is given by the Hopfield energy (Hinton et al., 1984):

$$\text{Energy}(y, h) = -\sum_{i \in \text{vis}} y_i \epsilon_i - \sum_{j \in \text{hid}} h_j \eta_j - \sum_{i,j} h_j D_{ij} y_i.$$  (6)
where $\eta$ and $\epsilon$ are the biases (constants) for the hidden and visible units respectively. $W_{ij}$ is the matrix of parameters representing an undirected connection between the hidden and visible variables. We can express the Boltzmann distribution as an energy model with energy function $F(y)$:

$$p(y, h) = \frac{1}{Z} \exp(-F(y))$$  \hspace{1cm} (7)

where the partition function $Z = \sum_{i,j} \exp(-\text{Energy}(y, h))$ is the normalization function over all possible vector combinations. $F(y)$ is defined as the free energy $F(y) = -\ln \sum_h \exp(-\text{Energy}(y, h))$ and further simplified to

$$F(y) = -y_i \epsilon_i - \sum_{j \in \text{hid}} \ln(1 + \exp(D_{,,j} y + \eta_j))$$  \hspace{1cm} (8)

The probability of assigning a visible vector $y$ is given by the sum of all possible hidden vector states:

$$p(y) = \frac{1}{Z} \sum_h \exp(-F(y)),$$  \hspace{1cm} (9)

The RBM model is used to learn aspects of an unknown probability distribution based on samples from that distribution. Given some observation, the RBM makes updates to the model weights such that the model best represent the distribution of the observation. To generate data with this method, it is necessary to compute the log likelihood gradient for all visible and hidden units. Hinton introduced a fast greedy algorithm to learn model parameters efficiently using contrastive divergence (CD) method that starts a sampling chain (Gibbs sampling) from real data points instead of random initialization (Hinton, 2010).

3.2. Model Inference

The probability that the RBM network learns a training sample can be raised by adjusting the weights to lower the energy of that training sample and raise the energy of other non-training samples. In order to minimize the negative log likelihood of the probability distribution $p(y)$, we take its gradient derivative of the log probability of a training vector with respect to the model parameters as follows:

$$\frac{\partial \log p(y)}{\partial \theta} = \langle y_i h_j \rangle_{\text{train}} - \langle y_i h_j \rangle_{\text{model}} = \phi^+ - \phi^-$$  \hspace{1cm} (10)

where the components in the angle brackets corresponds to the expectations under the specified distribution. The first and second terms are the positive $\phi^+$ and negative $\phi^-$ phases respectively. This function updates the model parameters using a simple learning rule with a learning rate $\Phi$:

$$\Delta \theta = \Phi(\langle y_i h_j \rangle_{\text{train}} - \langle y_i h_j \rangle_{\text{model}}).$$  \hspace{1cm} (11)

The updates for parameters $\theta = \{D_{ij}, \eta_j, \epsilon_i\}$ can be performed using simple stochastic gradient descent at each iteration of $t$:

$$\theta_t = \theta_{t-1} - \Delta \theta,$$  \hspace{1cm} (12)
To obtain a sample of a hidden unit from $\langle y_i h_j \rangle_{\text{train}}$, we take a random training sample $y$ and sample the state in the hidden layer is given by the following function:

$$p(h_j = 1|y) = \frac{e^{\eta_j + \sum_i D_{ij} y_i}}{1 + e^{\eta_j + \sum_i W_{ij} y_i}} = \sigma(\eta_j + \sum_i D_{ij} y_i),$$

(13)

where $\sigma(x) = e^x/(1 + e^x)$. Similarly, we can obtain a visible state, given a vector of sampled hidden units, via a logistic function:

$$p(y_i|h) = \frac{e^{\epsilon_i + \sum_j D_{ij} h_j}}{\sum_i e^{\epsilon_i + \sum_j D_{ij} h_j}}$$

(14)

Since weights are shared between $D$ and $G$ and they define the distributions of $p(y), p(h), p(y, h), p(y|h)$ and $p(h|y)$, we can express the posterior distribution as $p(y) = \sum_h p(h)p(y|h)$ (Ng and Jordan, 2002). Due to its bidirectional structure, this framework possesses good generalization capabilities. The visible layer represents the data (in the case of choice modelling, data represent selected choices), and the hidden layer represents the capacity of the model as class distributions.

Obtaining a sample from $\langle y_i h_j \rangle_{\text{model}}$ can be done by setting the states of the visible variables to a training sample and then the states of the hidden variables are computed using Eq. 13. Once a “state” is chosen for the hidden variables, a “reconstruction” phase produces a new vector $\tilde{y}$ with a probability given by Eq. 14, and the gradient update rule is given by:

$$\Delta \theta = \Phi(\langle y_i h_j \rangle_{\text{train}} - \langle y_i h_j \rangle_{\text{reconstruction}}).$$

(15)

This learning rule is known as the Contrastive Divergence (CD) algorithm, which is the difference between the Kullback-Liebler (KL) divergences (Hinton, 2002). We approximate the gradient function by using a Gibbs sampler which runs for a total number of $N$ chain steps initialized from a fixed point from the data distribution and then averaged across all examples (Carreira-Perpinan and Hinton, 2005).

3.3. Conditional RBM

Conditional RBM (C-RBM) expands the model to include “context variables” and allows for discriminative learning (Mnih et al., 2012), that can be adapted for prediction problems, i.e. $p(y|x, h)$. $k$ input explanatory variables are introduced as context variables so that they can be used to influence the latent variables, even though Eq.14 does not reconstruct these explanatory variables. This influence is represented by a weight matrix $B_{ik}$. The intuition is that for each latent variable, it acts as a function of the observed choice $y$, conditional on $x$ (see Fig. 2). In the choice prediction stage, a vector of new input samples $x$ generate latent variables $h$. Conditional on the explanatory and latent variables, a probability function describing the choice behaviour is given as:

$$p(y_i|h, x) = \frac{e^{\sum_k B_{ik} x_k + \sum_j D_{ij} h_j + \epsilon_i}}{\sum_{i'} e^{\sum_k B_{i'k} x_k + \sum_j D_{ij} h_j + \epsilon_{i'}}}$$

(16)

Likewise, sampling of the hidden state is extended to incorporate $x$:

$$p(h_j = 1|y) = \sigma(\eta_j + \sum_i D_{ij} y_i + \sum_k A_{jk} x_k),$$

(17)
where the update parameters are $\theta = \{D_{ij}, B_{ik}, A_{jk}, \eta_j, \epsilon_i\}$. During the reconstruction phase, the condition probability (Eq. 16) is equivalent to a MNL model with latent variables (where $h$ and $x$ represents the latent and observed variables respectively). Good latent variables $h$ best capture information along the orthogonal direction where choices $y$ and observed inputs $x$ vary the most. The training and choice estimation phase is illustrated in Fig. 3 and 4.

4. Data

In this section, we develop a choice model in terms of explanatory variables and latent variables where the latter indicates a factorial representation of the observed decisions which is simultaneously modelled in conjunction with the interaction with choice model. We developed a structured choice subset from a financial dataset, obtained from the Kaggle database\(^1\). The data shows a monthly basis record of each financial product purchase by customers of Santander. The time span of the data is from January 2015 to June 2016. The choice set is extracted from the original dataset but constrained only to single products were chosen out of a set of 13 product choices, 20 explanatory were used in the experiment. Table 1 lists the 13 choices and data distribution. Input data were scaled and normalized, examples with more than 1 product chosen in a month are removed from the dataset. Given the above conditions, a total of 253,803 valid

\(^1\)Dataset: https://www.kaggle.com/c/santander-product-recommendation/data
responses were recorded with 13 available choices. A descriptive list of mean and standard deviation values of the explanatory variables are shown in Table 2. The experimental question is straightforward: “Given a set of examples with explanatory variables, what product is the individual most likely to purchase in the given month?” In a typical situation, the decision maker chooses an alternative that yields the maximum utility, making an inference about the behaviour of the decision maker using the predictive model.

Table 1: Choices (y)

<table>
<thead>
<tr>
<th>Choice index</th>
<th>Name</th>
<th>Total sample distrib.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Guarantees</td>
<td>0.002%</td>
</tr>
<tr>
<td>2</td>
<td>Short-term deposits</td>
<td>0.83%</td>
</tr>
<tr>
<td>3</td>
<td>Medium-term deposits</td>
<td>0.07%</td>
</tr>
<tr>
<td>4</td>
<td>Long-term deposits</td>
<td>3.79%</td>
</tr>
<tr>
<td>5</td>
<td>Funds</td>
<td>0.98%</td>
</tr>
<tr>
<td>6</td>
<td>Mortgage</td>
<td>0.02%</td>
</tr>
<tr>
<td>7</td>
<td>Pensions</td>
<td>0.15%</td>
</tr>
<tr>
<td>8</td>
<td>Loans</td>
<td>0.035%</td>
</tr>
<tr>
<td>9</td>
<td>Taxes</td>
<td>2.68%</td>
</tr>
<tr>
<td>10</td>
<td>Cards</td>
<td>21.93%</td>
</tr>
<tr>
<td>11</td>
<td>Securities</td>
<td>1.42%</td>
</tr>
<tr>
<td>12</td>
<td>Payroll</td>
<td>22.04%</td>
</tr>
<tr>
<td>13</td>
<td>Direct debit</td>
<td>46.05%</td>
</tr>
</tbody>
</table>

4.1. Method for assessing model performance

We can estimate the weights for the latent inference model $B_{ik}$ and $D_{ij}$ by optimizing the lower bound of the KL-divergence using gradient backpropagation. Intuitively $D_{ij}$ represents the parameters for the explanatory variables and $B_{ik}$ represents the parameters for the latent variables. We selected models with 2, 4, 16 and 32 latent variables to observe the effects of increasing model complexity. One disadvantage of this step is that it results in a large number of estimated parameters: $N_{\text{params}} \in \mathbb{R}^{(I \times J) + (K \times I) + (K \times J) + K + I}$. With $J = 4$, we ended up with 409 parameters. To counteract overfitting due to this problem, we trained on 70% of our data and validate the model on the other 30%. When the validation error stops decreasing, the optimal estimation is reached (Goodfellow et al., 2016). A baseline comparison is set up using a standard multinomial logistic regression model with all explanatory variable and compared to the discriminative C-RBM modelling approach, followed by comparing the log-likelihood, $\rho^2$ model fit and predictive accuracy across all data models. The criteria for measuring performance of a categorical based model include: rho-square model fit and prediction error. The $\rho^2$ fit denotes the predictive ability between the trained model and a model without covariates. In the prediction error evaluation, the elements in the diagonal cells of a confusion matrix over the total number of examples denotes the accuracy of the model in predicting the correct choice and the error is

$$\text{Error}_{\text{valid}} = 1 - \sum_i P(y_{\text{pred}} = 1 | x, h, y_i = 1)$$  \hfill (18)

$y_i$ is the actual choice and $\text{Error}_{\text{valid}}$ is the sum of all the error probabilities for correct assessment for each choice. We fit the model on the training set and evaluate on the validation set.
Table 2: Explanatory variable descriptive statistics (x)

<table>
<thead>
<tr>
<th>Explanatory variable</th>
<th>Description</th>
<th>mean</th>
<th>std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Customer age</td>
<td>42.9</td>
<td>13.0</td>
</tr>
<tr>
<td>loyalty</td>
<td>Customer seniority (in years)</td>
<td>8.03</td>
<td>6.0</td>
</tr>
<tr>
<td>income</td>
<td>Customer income (€)</td>
<td>141,838</td>
<td>262,748</td>
</tr>
<tr>
<td>sex</td>
<td>Customer sex (1=male)</td>
<td>0.387</td>
<td>0.487</td>
</tr>
<tr>
<td>employee</td>
<td>Employee index, 1 if employee</td>
<td>0.0006</td>
<td>0.024</td>
</tr>
<tr>
<td>active</td>
<td>Active customer index</td>
<td>0.95</td>
<td>0.199</td>
</tr>
<tr>
<td>new_cust</td>
<td>1 if customer loyalty &lt; 6 mo.</td>
<td>0.045</td>
<td>0.207</td>
</tr>
<tr>
<td>resident</td>
<td>Resident index (Spain)</td>
<td>0.999</td>
<td>0.007</td>
</tr>
<tr>
<td>foreigner</td>
<td>Foreign citizen index</td>
<td>0.045</td>
<td>0.21</td>
</tr>
<tr>
<td>european</td>
<td>EU citizen index</td>
<td>0.995</td>
<td>0.006</td>
</tr>
<tr>
<td>vip</td>
<td>VIP customer index</td>
<td>0.116</td>
<td>0.32</td>
</tr>
<tr>
<td>savings</td>
<td>Savings Account type</td>
<td>0.0002</td>
<td>0.012</td>
</tr>
<tr>
<td>current</td>
<td>Current Account type</td>
<td>0.572</td>
<td>0.495</td>
</tr>
<tr>
<td>derivada</td>
<td>Derivada Account type</td>
<td>0.0009</td>
<td>0.03</td>
</tr>
<tr>
<td>payroll_acc</td>
<td>Payroll Account type</td>
<td>0.416</td>
<td>0.493</td>
</tr>
<tr>
<td>junior</td>
<td>Junior Account type</td>
<td>0.0001</td>
<td>0.0098</td>
</tr>
<tr>
<td>masparti</td>
<td>Mas Particular Account type</td>
<td>0.017</td>
<td>0.128</td>
</tr>
<tr>
<td>particular</td>
<td>Particular Account type</td>
<td>0.168</td>
<td>0.373</td>
</tr>
<tr>
<td>partiplus</td>
<td>Particular Plus Account type</td>
<td>0.113</td>
<td>0.316</td>
</tr>
<tr>
<td>e_acc</td>
<td>e-Account type</td>
<td>0.255</td>
<td>0.436</td>
</tr>
</tbody>
</table>

5. Results

We compare the different models based on their performance on the validation dataset. For the purpose of this study, we tried both normalized and non-normalized data but found that both produce similar results. All measurements are based on running algorithms using Theano Python libraries\(^2\). The model was trained with stochastic gradient descent on mini-batches of 64 samples for 400 epochs with input normalization (Hinton et al., 2006). The training time was approximately 30 minutes for each model. At the given time, computational demand is not significant enough using only a few (<1000 hidden) units, however, it could be an important consideration when one decides to experiment with many models or very large parameter vectors. The results of the model comparison statistics over the validation set is listed in Table 3. We found that adding latent information about the relationship between explanatory variables and observed decisions was useful in increasing the modelling accuracy. BIC indicates that 8 hidden units is the optimal point before overfitting. To evaluate the efficiency of the models, a Hinton diagram is used to analyze the strengths of the connected weights by plotting the values and significance with choices on y-axis and variables on x-axis (Bremner et al., 1994). Figs. 5 through 9 shows the parameter estimates of the different model after completing the training procedure. The matrix shows the influence of each independent variable on each alternative or latent variable. Significant (> 95%) parameters are highlighted in blue. These values are normalized along the x-axis. Each of these models ran on a total of 76,141 valid examples where the individuals decide which of the 13 products to purchase. The learned parameters of the C-RBM prediction models B, D and \(\epsilon\) are projected onto patch matrix (figs. 6a, 7a, 8a and 9a) while parameters A and \(\eta\) are

\(^2\)Theano Python library: http://github.com/Theano/Theano
Table 3: Model training results

<table>
<thead>
<tr>
<th>Model</th>
<th>Valid error</th>
<th>log-likelihood</th>
<th>$\rho^2$</th>
<th>no. of params</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNL</td>
<td>0.4454</td>
<td>-206808</td>
<td>0.546</td>
<td>273</td>
<td>416915</td>
</tr>
<tr>
<td>CRBM (2 Hidden units)</td>
<td>0.4360</td>
<td>-203558</td>
<td>0.553</td>
<td>341</td>
<td>411237</td>
</tr>
<tr>
<td>CRBM (4 Hidden units)</td>
<td>0.4338</td>
<td>-202066</td>
<td>0.556</td>
<td>409</td>
<td>409075</td>
</tr>
<tr>
<td>CRBM (8 Hidden units)</td>
<td>0.4323</td>
<td>-200846</td>
<td>0.559</td>
<td>545</td>
<td>408279</td>
</tr>
<tr>
<td>CRBM (16 Hidden units)</td>
<td>0.4318</td>
<td>-200223</td>
<td>0.560</td>
<td>817</td>
<td>410321</td>
</tr>
</tbody>
</table>

Projected onto the patch matrix (figs. 6b, 7b, 8b and 9b). $\epsilon$ and $\eta$ are referred to as constants. The sizes of the square indicates the value of the parameter and the colour corresponds to the signs of the parameters (white = +ve, black = −ve). The statistical significance (t-test) of each parameter is calculated using $\frac{\theta}{\sqrt{\sigma}}$, where $\sigma$ is the inverse of the Hessian of the log likelihood with respect to the weights, with sample size adjustment.

6. Analysis

6.1. Comments on the results

We can characterize each hidden unit with the explained significance and strength from the incoming explanatory variables and the significance of each hidden unit with respect to the choices is represented by the weights $D^\top$. For instance, in Fig. 6b, hidden1 is characterized by individuals that are working-age, foreign, non-EU citizens, non-VIP and does not own any special accounts. This latent variable can be inferred with a label ‘high spending and wealth accumulation attitude’. From the model results, they are an indication of purchasing a payroll (+) or cards (+) product from the positive and significant parameter in fig. 6a. Likewise in hidden2, it is represented by older, loyal customers who are VIP and have held various account types. This latent variable can be denoted as ‘self-reliance attitude’ and are indication of less likely to purchase long term deposits (-), funds (-), securities (-) and cards (-) products.

From the presented results, it is clear that the RBM models differ significantly from the MNL model in terms of which parameters are strong and significant. This result seems to be broad-based in the sense that it is not dictated by the number of hidden units and signifies that the observed distribution has some latent factors that can be explored. However, we should mention that the training parameter initialization can have a small random effect on the model. Normalization of incoming data may reduce this internal variation that will be considered in future work. Note that in the parameter plots, the signs and strength contribution to the choice model differ from model to model. It would be interesting to find out if the obtained conclusions is due to the training procedure or from the hidden units. We see that in the parameter plots, the values and signs correspond to the strength of each variable. For instance, the parameters for “Guarantees” choice are not significant, since the distribution is very low (0.002%). The latent models show similar results. For CRBM with 2 and 4 hidden units, almost all of the parameters are significant, except for income, employee, savings derivada and junior variables. This can be attributed to the small mean values (and high deviation).

The CRBM with latent variables outperforms the MNL model, however, the performance increase from increasing the number of latent variables past 4 LV, is small. This would suggest that the upper bound of latent representative capacity is reached with just a small number of latent variables. Using 2 or 4 latent variables is enough to give the model a significant improvement over a MNL structure.

6.2. Discussions

The purpose of this study is to analyze how we can use other means of latent behaviour modelling in the absence of attitudinal indicators. In ICLV modelling, specialized surveys have to be constructed
with attitudinal questions to model latent effects on the decisions. While this has been one of the more popular method in discrete choice analysis, there are several disadvantages to it. First, attitudinal questions are subjective and the behaviour are subjected to changes over time. Next, existing datasets that have no attitudinal questions cannot leverage on the ICLV model, thus latent effects cannot be utilized. We explore generative modelling of the choice distribution to uncover latent variables using machine learning methods, without measurement indicators. We hypothesize that latent effects can come not only from attitudinal questions, but also from the posterior choices. In effect, we are modelling latent components that fits the real choice distribution rather than achieving good statistics on subjective models. For example, there could possibly be some mean behaviour that dictates a more probable influence on purchases given some latent variables.

For this method to be effective, certain conditions have to be present: First, difficulty to get a good discriminative prediction result using only the provided explanatory variables. In this scenario, the C-RBM models were able to learn good latent variable representation and improve the model fit and prediction accuracy while providing latent variable inferrability. Next, when the data lacks attitudinal survey data, this method can find latent effects without the use of subjective measurement indicators.

The limitations of this study are the absence of choice dynamics or explanatory variable dynamics, i.e. changes over time or multiple choices for the same individual was not considered. We hypothesize that this may improve the model significantly but we are still looking for ways to incorporate dynamics into our C-RBM model. In recent studies, we have seen dynamic frameworks such as recurrent neural networks used in modelling temporal data (Taylor et al., 2007; Mnih et al., 2012). Finally, it is worth noting that as the number of latent variable increases, the number of estimated parameters increases exponentially. This will pose problems in large datasets and the ability to reduce dimensionality will give a significant benefit to efficient use of model parameters. In our observation, using cross-validation method and choosing the model which gives the lowest optimal validation error is a justifiable method to prevent overfitting using all the parameters. In the future, we would also look at the possibility of introducing deep learning architecture to choice modelling by stacking RBMs (Otsuka and Osogami, 2016).

7. Conclusion

While ICLV model are optimized to predict the effects of latent constructs on the choice model using measurement indicators to guide latent parameters selection, this method uses observed decisions as an influence source for optimizing latent variables through machine learning. This is not to say that we do not agree with using measurement indicators which may often be subjective and may raise mis-specification problems. On the contrary, when explanatory variables are poor predictors, ICLV models can improve latent effects on choice models directly (Vij and Walker, 2016). However, latent effects may not only be present in attitudes and perceptions but also in the direct observation of choices. Generative modelling in DCA is inspired by state-of-the-art machine learning algorithms that performs unsupervised feature extraction from unlabelled data used in classification problems (Hinton et al., 1984). However we have not seen any traction in DCA using this approach, therefore, we have a strong reason to believe that understanding which source of latent effects are important and desirable and what should be considered when building a discrete choice model. Our current work explores the use of posterior choice distribution for latent behaviour modelling. A future study that would be of interest is to extend this method to datasets with attitudinal questions and survey, for example, rail travel survey (Hess and Stathopoulos, 2013), and perform an analysis on both methods. We believe contrasting studies can guide not only in selecting the appropriate models, but also direct research effect to more promising directions.
Acknowledgements

We would like to acknowledge Fonds de recherche du Québec - Nature et technologies (FRQ-NT) for supporting this work via the team grant No. 2016-PR-189250.

Figure 5: MNL model parameters. White: +ve values, Black: -ve values, Blue: >95% significant
Figure 6: (a) C-RBM model with 2 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant

Figure 7: (a) C-RBM model with 4 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant
Figure 8: (a) C-RBM model with 8 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant

Figure 9: (a) C-RBM model with 16 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant
References


A. Rosenfeld, I. Zuckerman, A. Azaria, S. Kraus, Combining psychological models with machine learning to better predict people’s decisions, Synthese 189 (2012) 81–93.


D. Donoho, 50 years of data science, Tukey Centennial Workshop (2015).


