GENERATIVE MODELLING AND MACHINE LEARNING METHODS FOR TRAVEL BEHAVIOUR ANALYSIS

by
Melvin Wong
Bachelor of Engineering, Nanyang Technological University, 2015

A dissertation
presented to Ryerson University

in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the program of
Civil Engineering

Toronto, Ontario, Canada, 2019
© Melvin Wong, 2019

Author's Declaration

I hereby declare that I am the sole author of this dissertation. This is a true copy of the dissertation, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this dissertation to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this dissertation by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my dissertation may be made electronically available to the public.

Abstract

Generative Modelling and Machine Learning Methods
for Behaviour Analysis
by
Melvin Wong
PhD in Civil Engineering - Department of Civil
Engineering
Ryerson University
2019

The dissertation outlines novel analytical and experimental methods for discrete choice modelling using generative modelling and information theory. It explores the influence of information heterogeneity on large scale datasets using generative modelling. The behaviourally subjective psychometric indicators are replaced with a learning process in an artificial neural network architecture. Part of the dissertation establishes new tools and techniques to model aspects of travel demand and behavioural analysis for the emerging transport and mobility markets. Specifically, we consider: (i) What are the strengths, weaknesses and role of generative learning algorithms for behaviour analysis in travel demand modelling? (ii) How to monitor and analyze the identifiability and validity of the generative model using Bayesian inference methods? (iii) How to ensure that the methodology is behaviourally consistent? (iv) What is the relationship between the generative learning process and realistic representation of decision making as well as its usefulness in choice modelling? and (v) What are the limitations and assumptions that have needed to develop the generative model systems?

This thesis is based on four articles introduced in Chapters 3 to 6. Chapters 3 and 4 introduces a restricted Boltzmann machine learning algorithm for travel behaviour that includes an analysis of modelling discrete choice with and without psychometric indicators. Chapter 5 provides an analysis of information heterogeneity from the perspective of a generative model and how it can extract population taste variation using a Bayesian inference based learning process. One of the most promising applications for generative modelling is for modelling the multiple discrete-continuous data. In Chapter 6, a generative modelling framework is developed to show the process and methodology of capturing higher-order correlation in the data and deriving a process of sampling that can account for the interdependencies between multiple outputs and inputs. A brief background on machine learning principles for discrete choice modelling and newly developed mathematical models and equations related to generative modelling for travel behaviour analysis are provided in the appendices.

Acknowledgements

I would like to thank everyone who have supported me during my work on this dissertation. Especially I want to thank my supervisor Dr Bilal Farooq and former co-supervisor Dr Guillaume-Alexandre Bilodeau, for their continuous support and advice. Special thanks to my colleagues that I have worked with in LiTrans at Ryerson and Polymtl: Antoine, Zhihui, Alexandra, Dr Anae Sobhani, Arash, Lama and Dr Shadi Djavadian. It was a pleasure working with all of you.

This work has been supported by the following grants:

- Canada Research Chair in Disruptive Transportation Technologies and Services
- Fonds de reserche du Québec - Nature et technologies (FRQ-NT)
- Ryerson University
- The Canada Foundation for Innovation (CFI) research grant

Contents

Author's Declaration ii

Abstract iii

Acknowledgements v

List of Tables ix

List of Figures x

I Introduction **1**

1 Overview 2

1.1 Contributions 4

1.2 Dissertation Outline 5

2 Background 9

2.1 Discrete Choice Models for Behaviour Analysis 9

2.1.1 Limitations of Random Utility Maximization 12

2.1.2 The motivation for information theory 13

2.1.3 Data-driven Travel Behaviour Analysis 14

2.2 A New Approach to Behavioural Modelling 15

2.2.1 Machine learning based travel behaviour and decision making 15

2.2.2 Relationship between generative modelling and economic behavioural theory 21

2.2.3 How discrete choice analysis can benefit from using a generative modelling framework? 24

2.3 Summary 26

II Dissertation Articles 28

3 Modelling Latent Travel Behaviour Characteristics: A Generative Machine Learning Approach . . . 29

3.1 Introduction . . . 31

3.2 Literature Review . . . 32

3.3 Framework and Estimation of Latent Variable Models . . . 35

3.4 Case Study: Train Hôtel Mode Choice Dataset . . . 40

3.5 Results . . . 42

3.6 Conclusion . . . 48

4 Discriminative Conditional Restricted Boltzmann Machines for Discrete Choice and Latent Variable Models . . . 49

4.1 Introduction . . . 52

4.2 Literature Review . . . 53

4.3 Methodology . . . 60

4.4 Case Study: Consumer Choice Preference Dataset . . . 64

4.5 Results . . . 68

4.6 Conclusion . . . 75

5 Information Processing Constraints in Travel Behaviour Modelling: A Generative Learning Approach . . . 78

5.1 Introduction . . . 81

5.2 Information Theory in Behaviour Models . . . 84

5.3 Methodology . . . 87

5.4 Case Study: Montreal Trajet Dataset . . . 100

5.5 Discussions and Conclusion . . . 111

6 A Bi-partite Generative Model Framework for Analyzing and Simulating Large Scale Multiple Discrete-Continuous Travel Behaviour Data . . . 114

6.1 Introduction . . . 117

6.2 Literature Review . . . 119

6.3 Proposed generative machine learning approach . . . 125

6.4 Case Study: Montreal Trajet Dataset . . . 133

6.5 Results . . . 139

6.6 Conclusion . . . 149

III Concluding Remarks **151**

7 Conclusion 152
7.1 Research Contributions 153
7.2 Limitations 155
7.3 Future Work 156

Appendices **159**

A Advanced machine learning principles for discrete choice modelling 160
A.1 Bayesian statistical approach 160
A.2 Optimization algorithm 162
A.3 Undirected graph model for behavioural representation 163
A.4 Model training algorithm 166
B Mathematical Models and Equations 167
B.1 Multiple discrete and continuous conditional probability generation . 167
B.2 Model elasticity 170
B.3 Example of derivation of the joint probability 171
C Code listing for generative model estimation 173
D Dataset description 185
References 189

List of Tables

3.1 Optimal Parameters for ICLV and C-RBM Choice Model 45
3.2 Optimal ICLV and C-RBM indicators 46
3.3 Alternative Specific Variables and Constants Estimated Within the ICLV and C-RBM Models 47
4.1 List of choice alternatives (**y**) 66
4.2 Explanatory variable descriptive statistics (**x**) 67
4.3 Model training results 69
4.4 Parameter sensitivity rank and standard error difference for estimated parameters **B** for sampling-based sensitivity analysis 74
5.1 Result of maxent estimates on β-parameters. 110
6.1 Notations. 125
6.2 k-th central moment of the original data samples 146
6.3 k-th central moment of the generated data samples (H=5) 147
6.4 k-th central moment of the generated data samples (H=25) 147
6.5 k-th central moment of the generated data samples (H=100) 147
6.6 Variable pair correlation 148
6.7 5% sample size analysis of the generative choice model outputs 148
D.1 Table of descriptive statistics for Train Hôtel dataset 186
D.2 Table of descriptive statistics for Santander dataset 187
D.3 Table of descriptive statistics for Mtltrajet dataset 188

List of Figures

1.1 Dissertation Overview . . . 8

2.1 Common transform functions used in deep learning . . . 17

2.2 Example MLP topology with 1 hidden layer, 3 neurons, 5 inputs and 4 outputs. . . . 18

2.3 Example RBM topology. . . . 20

2.4 RBM framework which models different configurations of input and outputs, (a) generic RBM with 1 hidden and 1 observed layer, (b) the observed layer is divided into input and conditional output, (c) the output is defined as a function of input directly with the hidden layer acting as a regularizer. . . . 21

3.1 Origin Destination nodes modelled for computational experiments. . . . 41

3.2 A joint tri-partite RBM structure for travel analysis with latent behaviour variables. . . . 42

4.1 Classical structural framework for (a) latent class model and (b) integrated choice and latent variables model . . . 54

4.2 Framework for a C-RBM choice model conditional on explanatory variables and choice distribution. . . . 57

4.3 C-RBM (a) positive $\phi+$ and (b) negative $\phi-$ phases during semi-supervised discriminative training. Weights (connections) are learned to reduce reconstruction $\tilde{y}$ error. . . . 65

4.4 During the choice prediction phase, (a) latent variables are sampled using explanatory variables, and (b) the choice model is estimated with variables x and h. . . . 65

4.5 MNL model parameters. White: +ve values, Black: -ve values, Blue: >95% significant 70
4.6 (a) C-RBM model with 2 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant 70
4.7 (a) C-RBM model with 4 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant 71
4.8 (a) C-RBM model with 8 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant 71
4.9 (a) C-RBM model with 16 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant 72
5.1 Framework for generative modelling. 83
5.2 Visualization of number of trip trajectory origin points by city district from the dataset. 100
5.3 Learning curve of the sample negative loglikelihood from the choice model.102
5.4 Mode share forecast. 103
5.5 Distribution of data generating parameters. 105
5.6 Comparison of data generating output on activity type data. 106
5.7 Comparison of data generating output on trip distance data. 107
5.8 Comparison of data generating output on trip duration data. 107
5.9 β-parameter estimates using mode choice as the dependent variable, horizontal axis represent number of latent variables. 109
6.1 The generative bi-partite framework. 127
6.2 Graphical illustration of an RBM with connections represented by **W** . 131
6.3 Visualization of the trip trajectories across the Greater Montréal Region from the pilot study. 135
6.4 Training and validation likelihood curve (H=5) 138
6.5 Training and validation likelihood curve (H=25) 138
6.6 Training and validation likelihood curve (H=100) 139
6.7 Histogram of parameter value distribution by mode choice and number of latent variables. Vertical dashed line represent distribution mean. . . 140
6.8 Distance elasticities on mode choice 143
6.9 Discrete and continuous data generated from the model. 144

6.10 Data generation for a joint MDC output 145

B.1 Generating various multiple discrete-continuous outputs using generative models. (a) Example 1, (b) Example 2, (c) Example 3. 168

Part I

Introduction

Chapter 1

Overview

The transport and mobility market is currently undergoing a fundamental transformation, driven by three domains of disruptive technologies: Mobility-as-a-Service (MaaS), Connected and Automated Vehicles (CAV) and Artificial Intelligence (AI) [1, 2, 3]. These three domains offer new challenges for travel behaviour analysts and choice modelling practitioners. These challenges would demand a greater understanding of how the current mobility transformation will have an impact on society. Most notably, new consumer travel options such as ride-hailing services, on-demand car sharing and bike sharing are now ubiquitous in cities worldwide and are rapidly becoming the mainstay of transportation and mobility systems [4, 5, 6]. These new trends in the emerging disruptive mobility market can be attributed to the evolving socio-demographics, consumer behaviour and demand for more dynamic modes of travel [7]. In the European Union, the market for these mobility systems is expected to reach $700 billion by 2025 [8]. By 2030, the worldwide disruptive mobility ecosystem is projected to reach $1 trillion [9]. From a value-added perspective, there would be a significant impact on the global transport market through the development of new behavioural modelling techniques that would be able to exploit data-driven technologies, while also providing econometric interpretability for new policy decisions. This dissertation focuses on the third domain: Artificial Intelligence – through the development of novel behaviour modelling techniques that integrates generative modelling and information theory.

As we move towards a more data-driven and connected society, these new disruptive mobility services and systems will generate more data than ever [10]. Despite the increasing amount of information being collected from passive online services, sensors and other Big Data sources, gaps remain in existing travel behaviour modelling literature as to how these new data-driven technologies can provide significant and beneficial improvements to behaviour modelling [10]. This dissertation seeks to connect aspects of data-driven applications with travel behaviour analysis by investigating the potential of machine learning and generative modelling.

Traditionally, travel behaviour modelling uses hypothesis-driven models derived from discrete choice theory. Hypothesis-driven models allow simple model interpretability, provide insight into behavioural patterns and offer a direct method to identify changes in consumer preferences based on observed characteristics of the choice or individual [11]. These (mostly) linear in parameter yet robust descriptive models are generally suited for small data samples with a relatively small population and decision heterogeneity. However, they do not tend to scale well to large datasets. As the number of observations in the data increase over time, statistical or measurement indicators may give a misleading perspective of model fit, if not corrected for variations in the data as the data grow. Conventional discrete choice analysis methods are beginning to encounter the limits of their usefulness in the data-driven and connected transportation landscape.

Meanwhile, with the increase in data availability, computing resources and new algorithms, we are now able to train complex and deep learning models efficiently. At the same time, the development of deep learning data-driven models have been progressing at a rapid rate and are considered the benchmark for building predictive models[1]. These data-driven modelling techniques are increasingly being used in other fields of research and for real-world applications to great success, for example in computer vision recognition, healthcare and recommendation systems [12, 13]. They differ from discrete choice analysis in their model structure, and ability to capture non-linear correlations using multiple layers. However, data-driven models have recently been questioned on their interpretability, and the "black-box" structure and the primary limitations of deep learning models have been its difficulty in explaining

[1] Industry-wide Data Science and Machine Learning Survey conducted by Kaggle ($n = 16,000$): Kaggle, 2017. The State of ML and Data Science. Retrieved from: https://www.kaggle.com/surveys/2017, June 2019.

the estimation process and how the model connects with behavioural and policy changes [14, 15].

This dissertation makes the case that deep learning and neuroscience are significant for behavioural analysis and choice modelling. Machine learning systems operate similarly to human behaviour – by minimizing a cost function that learns an optimal strategy while shaping the internal representation of the machine learning system. This strategy is in contrast to discrete choice models, which is are rooted in the theory of utility maximization. It suggests that choice selection behaviour can encompass an internal choice mechanism hidden from the observer that minimizes energy and risk, while at the same time maximizing utility and reward [16]. This dissertation also explores the concept of information heterogeneity and how we can develop Bayesian-based inference models that capture the underlying unobserved variations in a behaviourally plausible way through machine learning algorithms. Lastly, this dissertation highlights the implementation of welfare analysis in machine learning models, specifically, by calculating the elasticities of model parameters and deriving statistical interpretability.

1.1 Contributions

The work in this dissertation will contribute to the following aspects:

Travel behaviour theory. New experimental and analytical methods for behaviour modelling are investigated using semi-supervised generative models while exploring their underlying connections to behaviour theory. Recent methodologies and frameworks from neuroscience, statistics, optimization and machine learning are specifically into behavioural models for travel behaviour analysis and applications.

Data-driven applications. Experiments were devised to test the proposed framework using Big Data sources as well as traditional stated/revealed preference surveys. Our experiments and results show how forecasting and simulation can be implemented using a generative model framework. In model interpretation and analysis, several techniques for evaluating generative models are shown using statistical and functional properties that are compatible with discrete choice analysis.

Model development and statistical analysis. An end-to-end travel behaviour analysis framework is developed using a generative model that performs estimation, statistical analysis, forecasting and simulation. A new behaviour choice model specification is defined that can incorporate latent constructs from a generative learning process while retaining econometric features that are essential to behavioural analysis. In the development of new estimation methods, this dissertation outlines the use of variational Bayesian inference techniques to train generative models. Finally, this dissertation provides a blueprint for future research and data-driven model development as well as advancing the state-of-the-art in travel behaviour analysis.

1.2 Dissertation Outline

The main contributions of the dissertation are divided into four articles/chapters, each focusing on integrating different aspects of generative modelling and machine learning methods into behaviour analysis. In model development, we bridge the gap between hypothesis-driven, random utility models with data-driven, information theory-based machine learning algorithms. An overview of the dissertation is shown in Fig. 1.1.

Chapter 2 presents a background on the fundamental theory of travel behaviour modelling and generative modelling. This chapter discusses the recent developments in discrete choice modelling and how it can be enhanced through machine learning. The proposed methods in this dissertation based on these data-driven approaches enable the use of deep learning algorithms to model behaviour heterogeneity more accurately and realistically. This chapter also addresses the drawbacks and disadvantages of current modelling methods, and emphasizes the importance of econometric interpretation in machine learning-based models.

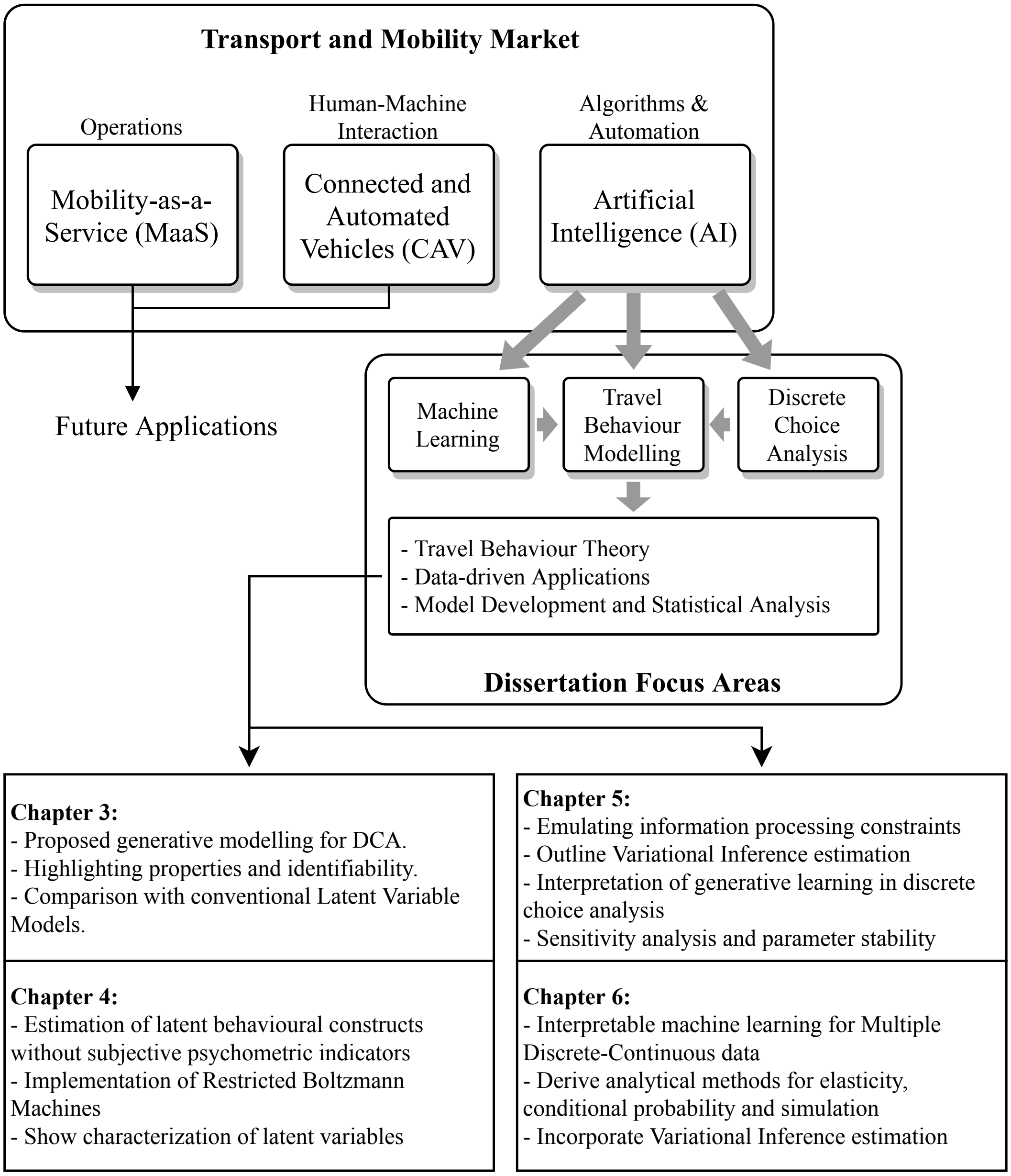
Chapter 3 introduces a novel generative modelling framework seeking to analyze latent travel behaviour characteristics. A variant of the conditional restricted Boltzmann machine framework (a type of generative model) is proposed to incorporate the information relationship between observed and latent variables. The properties of the model framework are highlighted, and the identifiability of these latent behaviour characteristics are explored. Finally, an experiment provides an empirical comparison against the conventional structured equation modelling method.

Chapter 4 extends the flexibility of the generative modelling framework to advanced discrete choice modelling strategies, allowing for the estimation of behavioural heterogeneities without the disadvantages of conventional hypothesis based model specification. The proposed method combines the efficiency of machine learning algorithms and exploits the parallelization gains from the generative modelling approach to identify useful representation in the data. At the same time, the proposed method does not pre-define any semantic meanings to each latent variable, eliminating the reliance on subjective measurement indicators used in integrated choice and latent variable (ICLV) models. In the case study, this work devises experiment and statistical analysis is presented using a Hinton diagram [17] to show how generative modelling can characterize latent variables with semantic meanings without additional psychometric data.

Chapter 5 proposes a data-driven generative machine learning version of rational inattention model [18]. Rational inattention frames the choice problem as a communication channel with finite Shannon capacity, which stems from the similar principles of neuroscience where information theory explains behaviour learning and inference. The methodology of the generative model and the associated learning process is outlined. The principles demonstrated in this chapter can be formulated as a generalized entropy and utility-based multinomial logit model. The effects of information heterogeneity on a travel choice are demonstrated, and the econometric interpretation of the properties of the generative model is analyzed. The findings suggest that individuals may ignore certain exogenous variables and rely on prior information for evaluating decisions under uncertainty and information heterogeneity.

Chapter 6 presents a practical application of generative modelling in the analysis of travel behaviour data. The theoretical background that supports generative machine learning methods provides a simple intuition and a plausible explanation as to how the model emulates a learning behaviour and how individuals account for their information processing cost. A bi-partite, multiple discrete-continuous (MDC) framework extension is proposed to estimate MDC behaviour data and investigate how the generative model can produce accurate data reconstruction. Analytical methods specifically for generative models are developed to show the econometric behaviour compatibility, elasticities and correlation analysis.

Chapter 7 summarizes the findings from the previous four chapters and the limitations of the developed methodologies. Finally, it provides new research questions for future work and possible extensions to incorporate generative modelling in travel behaviour analysis.


Transport and Mobility Market
Operations
Human-Machine Interaction
Algorithms & Automation
Mobility-as-a-Service (MaaS)
Connected and Automated Vehicles (CAV)
Artificial Intelligence (AI)
Future Applications
Machine Learning
Travel Behaviour Modelling
Discrete Choice Analysis
- Travel Behaviour Theory
- Data-driven Applications
- Model Development and Statistical Analysis
Dissertation Focus Areas

Chapter 3:
- Proposed generative modelling for DCA.
- Highlighting properties and identifiability.
- Comparison with conventional Latent Variable Models.

Chapter 4:
- Estimation of latent behavioural constructs without subjective psychometric indicators
- Implementation of Restricted Boltzmann Machines
- Show characterization of latent variables

Chapter 5:
- Emulating information processing constraints
- Outline Variational Inference estimation
- Interpretation of generative learning in discrete choice analysis
- Sensitivity analysis and parameter stability

Chapter 6:
- Interpretable machine learning for Multiple Discrete-Continuous data
- Derive analytical methods for elasticity, conditional probability and simulation
- Incorporate Variational Inference estimation

Figure 1.1: Dissertation Overview

Chapter 2

Background

2.1 Discrete Choice Models for Behaviour Analysis

The fundamental theory of analyzing discrete choices that linked economic utility theory and choice behaviour was developed in the early 1960s [19, 20]. Since then, discrete choice models have been the primary tool for travel behaviour modelling and analysis.

The standard multinomial logit (MNL) model used in travel behaviour modelling relies on deterministic decision rules – utilities are operationalized by random variables and it is assumed that unobserved heterogeneities are independent and identically distributed (i.i.d.) [21]. This assumption on the MNL model implies that there are no common unobserved factors that affect the utilities. However, not all random utility models follow this strict rule. Flexible error component structural models have been developed to relax the independence of irrelevant alternatives (i.i.a.) assumption by parameterizing an appropriate error variance as a function of individual attributes.

Recent developments in discrete choice modelling have devised new ways for integrating sensitivity effects, latent behaviour constructs and behaviour theory (e.g. Prospect Theory, reference dependence and loss aversion) [22, 23, 24]. Notably, these approaches extend non-deterministic methods (i.e. MNL models) to include effects from preference deviations, the bias in perspectives and evaluation of gains and losses from a reference point.

The presence of endogenous behaviour has a direct influence on the estimation of discrete choice models. In particular, studies have shown that economic values such as willingness-to-pay (WTA) or travel time variability that incorporate choice under risk and uncertainty result in more realistic behaviour models [24, 25].

MNL models assume that the decision-makers have perfect information availability, but from the perspective of the analyst, they have incomplete information about the individual's behaviour [11]. Therefore, behavioural uncertainty has to be taken into account from various sources unknown to the observer, for instance: decision protocols, choice sets, unobserved state variation and unobserved attributes [26]. The utility U_i, is modelled based on a function that associates some linear combination of explanatory variables as the deterministic utility V_i and an unobserved utility term ε_i in an MNL model:

$$U_i = V_i + \varepsilon_i \tag{2.1}$$

The decision maker's selected alternative is one with the highest utility from a choice set C:

$$P(i|C) = p(U_i \geq U_j \ \forall \ j \in C, j \neq i) \tag{2.2}$$

Choice preferences are measured through a weighted sum of the individuals' utilities: $V_i = \beta_{i0} + \sum_m \beta_{im} x_{im}$, where m denotes the explanatory variables and β_{i0} is the alternative specific constant. The evaluation of the utility is based on a specific set of perceived values of choice attributes and socio-economic factors. These early behaviour models have traditionally assumed a rational decision-maker, in which the decision-maker maximizes their utility given some constraints, e.g. time, budget, location. The choice selection behaviour has always been defined as a strict utility specification in which optimization solutions are developed to maximize a log-likelihood objective function [27].

Despite its simplicity, hypothesis-driven discrete choice models are often rigid and may not capture higher-order interactions between individuals, habits and choices. The underlying assumptions are often violated in decision making experiments, and the complexity of human behaviour cannot be adequately represented by utility alone since the analyst does not have access to the underlying behaviour. However, performing model selection based on simplicity is not necessarily detrimental, but it provides an intuitive way of model interpretation.

A generalization of the MNL model is the well-known Mixed MNL model [28]. It involves integrating the MNL probability function over the distribution of unobserved random terms:

$$P(i|C;\theta) = \int P(i|C;\beta) f(\beta|\theta) d\beta \tag{2.3}$$

where θ is a vector of unobserved variance parameters specifying the underlying correlation characteristics, β are the model parameters, and f is a density function that the random realizations of β are drawn. The Mixed MNL model allows a flexible substitution method across alternatives in an error components structure [21, 29]. The error components structure divides the utility into two components: one which is specified to be i.i.d. distributed, and another which is endogenous across the alternatives. The structure of the latter term arises because private information (psychometric values, attitudes, habits, etc.) are not fully observed in the data [30].

The choice outcome is often a result of a series of planned or unplanned processes (dynamics), for example, a link-based choice model or choices made repeatedly over a duration [31]. This underlying effect plays a crucial role in developing accurate and realistic choice models. A large number of decisions also involve comparison and evaluating risk between alternatives, for example, Prospect Theory, regret minimization and context processing [20, 32, 33].

Many choice models have been developed recently to capture different effects of these unobserved behaviours. Studies on the impact of psychological factors, e.g. attitudes, perceptions, lifestyle preferences on the systematic utility have paved the way for latent constructs being an integral part of the choice model specification [34, 35]. These classes of models include Latent Class Models (LCM) and Integrated Choice and Latent Variable Models (ICLV) [11].

The LCM is designed to capture taste heterogeneity across population segments or when choice sets vary across individuals [36]. The population segment heterogeneity is defined as latent constructs represented by a class probability:

$$(p(i|C) = \sum_{s\in S} p(i|C,s)p(s) \tag{2.4}$$

where s defines the class segment, and $p(s)$ is the probability that the decision-maker is in the class segment. The ICLV model addresses the latent constructs by incorporating psychological and attitudinal factors into an integrated structural equation model (see Chapter 4 for ICLV formulations). However, two key drawbacks have made the ICLV model limited in their uses in policymaking: First, measurement indicators have to be available, which is often not defined in most data collection – especially in revealed preference surveys. Second, in order for the latent constructs to have a significant effect on the choice model, explanatory variables need to be poor indicators of choice – which means that ICLV models have no added value if the original data structure reflects the population behaviour well enough [35].

2.1.1 Limitations of Random Utility Maximization

While discrete choice modelling has been the cornerstone of behaviour modelling, they can be poorly specified if the observed attributes are highly heterogeneous and individual-specific [37]. A significant limitation is that there are inconsistencies (i.e. irrational behaviour, choice paradoxes, missing data, repeated choices, noisy data) that are usually not considered within the model estimation process. Irrational behaviour reflects the cognitive biases in the decision process that require decision-makers to learn about their environment to minimize the uncertainty of the choices presented to them. Ultimately, the aim is not just to maximize some reward or utility, as in classical economic models, but also to minimize the uncertainty from information-processing costs.

Random utility-based choice models depend not only on the observed attributes of the choice and individuals but also on the unobserved attributes such as psychological factors, irrational behaviour and habits which cannot be easily quantified. As such, the decision maker's internal processes during preference formation are not directly observed and remain unexplained in the choice models and are assumed to be implicitly captured by the error terms [38]. From an information-theoretic perspective, this implies that some internal decisions cannot be observed (latent behaviour), but can be estimated through Bayesian inference methods, i.e. negative log-evidence as the equivalence to model uncertainty [39].

2.1.2 The motivation for information theory

The upper bounds for model evidence of the data are provided by laws of conservation in physics to make sure that the behaviour models are identifiable and can converge to a steady state. This concept is known as the free energy principle that comprises of two terms: the expected utility and the entropy and we borrow this same concept which allows us to model the human behaviour more realistically. Free energy in this context refers to the energy consumption in the human brain during decision making. It posits that the human brain are "computationally" efficient systems. For example, in choice set sampling of alternatives, not all alternatives will be chosen in the choice set, analogous to how human decision making disregard certain options because it might be too energy inefficent to consider all options. The utility term is the standard econometric utility that reflects reward maximization, while the latter term is the average uncertainty sampled from a probability density. It represents an ergodic process that converges the long term average, and it rests upon the fact that self-organizing agents resist a tendency for disorder [16]. This free energy principle has been suggested to provide a unified theory of behaviour, perception and implementation of machine learning which combines insights from the Boltzmann machine for perceptual learning [40, 41, 42]. In particular, the logit model derived by McFadden was also based on the Boltzmann distributions [43, 41].

It can be easily shown that minimizing the free energy can be expressed as a Kullback-Leibler divergence between the recognition model and the generator. By measuring and optimizing for minimum free energy, learning becomes very effective and efficient. Because the divergence is always positive, minimizing the free energy implicitly means that the model converges to minimum uncertainty. This minimization process is performed over time in human behaviour, i.e. formation of habits, increasing knowledge about a decision. Theoretically, this process can also be applied to travel behaviour adaptation and dynamics by applying a stochastic gradient descent on the free energy objective function.

Endogeneity is a core and challenging issue in travel behaviour, as highlighted in [44]. In conventional discrete choice modelling, i.e. MNL, it helps to reason (and to simplify the model) that decision-makers process information rationally, that is to say, that decision-makers do not try to 'discover' underlying biases and change their habits over time. However, information processing cost plays a vital role in an endogenous choice decision, and the behaviour of informed and uninformed decision-makers are very different from each other. In particular, the free energy principle explains the information difference between the two types of decision-makers, leading to stochastic behaviour. Learning from data improves the knowledge about the choice, resulting in more 'deterministic' choices of informed decision-makers.

This dissertation relates to this concept by introducing a generative model that encapsulates the *learning process* of the individual through a deep learning algorithm known as a restricted Boltzmann machine.

2.1.3 Data-driven Travel Behaviour Analysis

The purpose of travel behaviour analysis is to explain the cause-and-effects of certain constraints and attributes on individuals' choice preferences, which are grounded in neoclassical welfare economics [11]. For example, to measure the benefit of time and cost of a travel mode, choice models use willingness-to-pay (WTP) or level-of-service (LoS) metrics. Measuring WTP or LoS is a common practice for analysts and policymakers to decide on improvements or changes to transport systems.

With the emergence of data-driven analytics and new mobility technologies and services, understanding the underlying properties of the data concerning travel behaviour becomes an essential concept for accurate demand forecasting. In recent years, these new services and technologies have changed the way we analyze and evaluate travel behaviour. For example, on-demand ride-hailing services require frequent or even instantaneous information of where and when vehicles are needed to be deployed. App-based travel planners need real-time traffic data which may change by the minute and static model estimated on historical data may not be sufficient. In order to be operationally successful, transport service providers need access to a constant stream of data and a method to transform these data into useful, predictive models in real-time. Furthermore, these models also need to be able to forecast unforeseen events and able to adapt to new modes of travel and evolving travel behaviour. A data-driven machine learning models can be used to complement or replace conventional choice modelling since the underlying learning process emulates the choice behaviour and forms a mechanical representation of the human brain [45].

2.2 A New Approach to Behavioural Modelling

The assumption of "rational decision-maker" is often a convenient way of representing travel behaviour. However, this assumption is often violated in choice situations. The inconsistency may arise due to unknown factors involved that are not captured in the observed data. This is reflected under the term *decision under uncertainty* [22]. In contrast, if the underlying subjective factors are known for all events, then this distribution generates a probability distribution over the choices, which is identified by the utility value. Even then, some decision-makers may not always be utility maximizers, and there has been empirical evidence to suggest that other decision rules or protocols may influence decision making [46]. From the perspective of the analyst – assuming that there is significant enough source of data and a computationally feasible method of model estimation – it is theoretically possible to fully represent an *irrational* (and a more realistic representation of) decision-maker. The hypothesis posits that machine learning and data-driven modelling can account for these constraints and variations arising from data heterogeneity. Data-driven models inspired by neuroscience that form the basis of artificial neural network models can achieve a more realistic representation of human decision-making behaviour. Despite the widespread belief that neural networks are black-boxes and challenging to interpret [15], we argue otherwise: *The main reason being is that with new tools and algorithms that allow efficient computation of deep learning models, the goal of developing an explainable model of an (ir)rational decision-maker using massive datasets is no longer unreachable.*

2.2.1 Machine learning based travel behaviour and decision making

Machine learning is broadly defined as the task of development and analysis of algorithms that can learn from the observed data [47]. Learning corresponds to adjusting and fine-tuning parameters to model certain aspects of the underlying data. There are two primary types of learning methods: supervised and unsupervised

[47]. In supervised learning such as discrete choice models, the problem is given as finding the conditional probability distribution of a dependent output variable given some independent input variables $p(y|x)$. The learning task is to model the relation between the input(s) and the output(s). The learned model can then be used to forecast new output(s) given some new and unseen input value(s).

In unsupervised learning, the learning task is to generate some representation that summarizes and explains the principal features given some inputs $x_1, x_2, ..., x_n$. This category of models may include clustering, principal component analysis or deep generative models [48, 45]. The learned model is used as a generator to synthesize new data that have similar probability distribution as the inputs or as an encoding function to transform data into higher dimension space. To leverage the value of passively collected data with few to no labels, we consider semi-supervised learning which lies between supervised and unsupervised learning. Semi-supervised learning has been used in the past to improve model accuracy in discrete classification tasks with a mixture of labelled and unlabelled data which uses the underlying structure of the data and higher-order correlation to identify important latent features [49, 50].

Deep Learning

Deep learning in neural networks that we know today was initially inspired by neuroscience and sought to represent a human brain by a learning algorithm consisting of neurons, connections and weights. Neurons receive information from input data and manipulate the information to produce an output. When more than one layer of neurons is used, it is generally known as a multi-layer perceptron [45].

Advances in deep learning have enabled neural networks to learn from vast amounts of data and stored in its hundreds, if not thousands of model parameters, by providing a means to compute the gradient of the network efficiently. Deep learning typically uses backpropagation to learn functional parameters from a constant stream of input data. Backpropagation is simply an optimal control problem that solves a complex evaluation function by dividing it into several elementary steps that exploit the derivative used at each step [51]. Also, layers within the neural network employ non-linear transform functions on the input or intermediate layers. There are various types of transform functions used in deep learning. The most common types are (a) linear, (b) logistic (or sigmoid), (c) Rectifier Linear Units (ReLu) and (d) Softplus shown in Fig. 2.1.

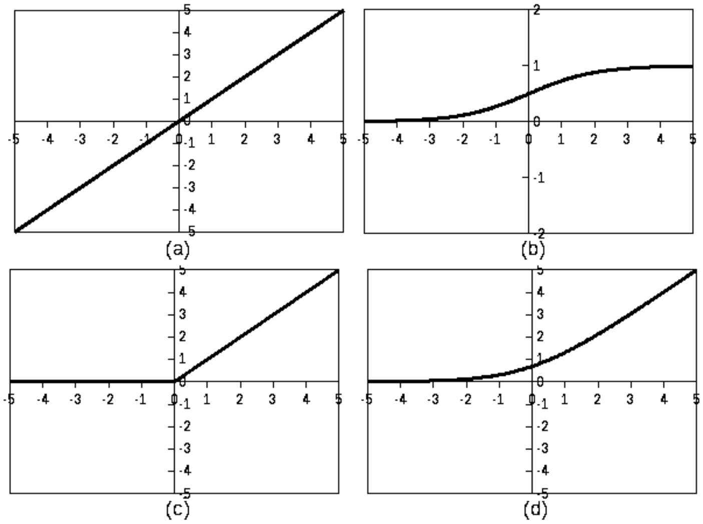


Figure 2.1: Common transform functions used in deep learning

The output y of the neural network is a series of non-linear transformation function applied consecutively with weights and biases. An example of a single-hidden-layer MLP would be the following:

$$h = f(W^{(1)}x + b^{(1)}) \tag{2.5}$$

$$p(y) = g(W^{(2)}h + b^{(2)}) \tag{2.6}$$

where f and g are the transform functions, h is the intermediate layer output/input and W, b are the weights and biases respectively. The output of Section 2.2.1 is a probability if the final non-linearity is a binary or multinomial logit transformation, for example:

$$g = \frac{e^{(W^{(2)}h+b^{(2)})}}{\sum_j e^{(W_j^{(2)}h+b_j^{(2)})}} \tag{2.7}$$

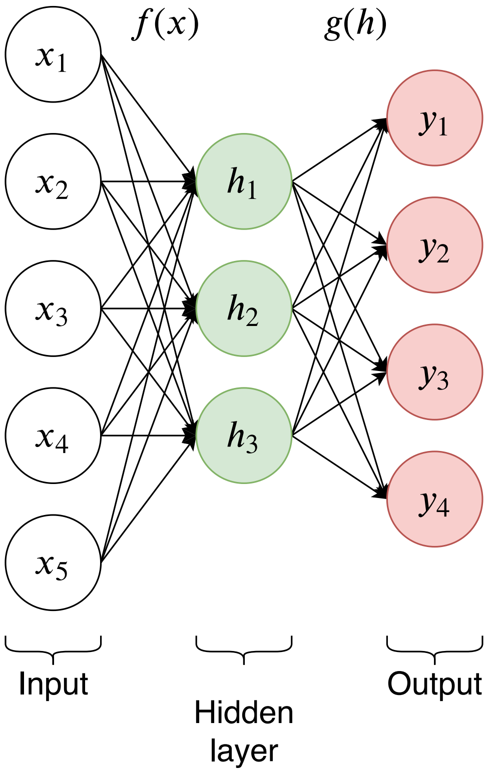


Figure 2.2: Example MLP topology with 1 hidden layer, 3 neurons, 5 inputs and 4 outputs.

As we increase the number of layers in the neural network, the model will recognize more complex decision patterns and thus, increasing the predictive potential. However, too many hidden layers can lead to a reduction in prediction accuracy as the network begins to overfit, and the noise becomes a major factor in deep neural networks [52]. Fig. 2.2 shows a topology of a simple 1 layer MLP. Multiple layers can be stacked within the MLP to increase model complexity.

While simple neural networks such as MLPs are powerful machine learning models, they cannot take advantage of unlabelled training data [45]. There are several methods to account for unlabelled data, including unsupervised and semi-supervised learning. Generative modelling is a form of unsupervised learning that is free from the restriction of having to rely on labelled data. An underlying model is trained to produce samples that are similar in distribution to the training samples.

Generative modelling

Generative machine learning can be defined as a problem where an agent learns to mimic the provided demonstration in the form of sensory data in a neural network [53]. Generative models are particularly appealing in deep learning because they can learn smooth transitions in the latent space, in contrast to fixed decision boundaries [54]. When an unknown probability distribution is placed on the inputs, a latent variable generative model can be defined to discover the underlying correlation in the data. Besides, each latent variable can represent a complex distribution that would otherwise require many discrete components [54]. In estimation, similar to MLPs, a backpropagation learning function can be applied to tweak the structure of the generative model in small steps with each observation of the data.

Estimating a generative model can be performed using simulation, Bayesian inference methods or reinforcement learning [45]. In particular, with Bayesian Inference, a product of learning models (ensemble of possible distributions) is obtained by multiplying the likelihood with this product. Bayesian inference allows the creation of a class of density models that have *components* rather than *categories* (as in discrete choice models) as their latent variable. The model generates a probability distribution over the parameter space, in contrast to traditional statistical approaches (e.g. discrete choice analysis) where a single point estimate is obtained. The Bayesian approach has several advantages: First, by having a probability density, we can take uncertainty into account during the simulation and forecasting, thus improving the quality and realism of the predictions. Viewing the generative model as a "regularizer" by defining a prior over the input data, we can control the complexity or uncertainty of the parameters, thereby reducing the problem of overfitting [55].

One popular class of generative model is the restricted Boltzmann machine (RBM) and the deep Boltzmann machine (DBM) [56] Both models typically involve estimating a joint probability distribution, assuming a $\{0,1\}$-Bernoulli function over the latent variables. Fig. 2.3 is an example of an RBM topology with five observed variables and four latent variables. The RBM is an undirected graphical model containing a set of observed (visible layer) and a set of latent (hidden layer) variables. The two layers are interconnected, and there are no connections between each variable in the visible or hidden layer. The layers are associated with a weight matrix and an optional bias unit.

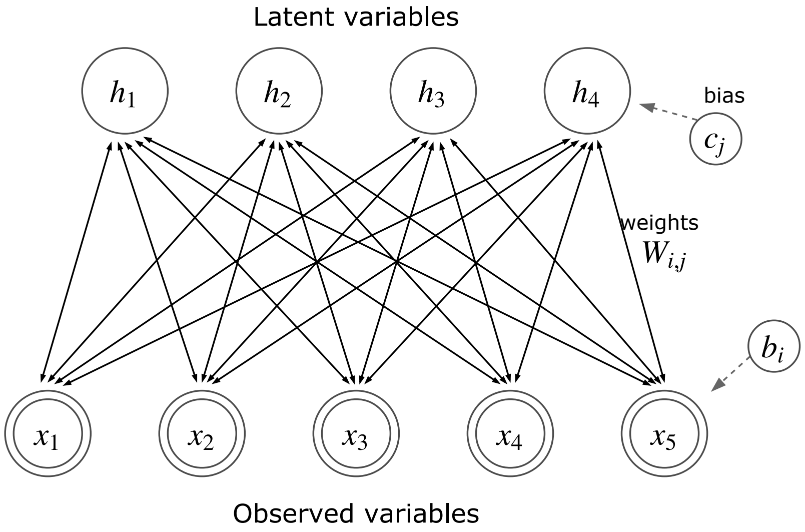


Figure 2.3: Example RBM topology.

Early methods of training RBMs included the use of Markov Chain Monte Carlo (MCMC) methods which is computationally expensive. The recent emergence of variational methods, including Bayesian variational inference, have made generative models much easier to optimize, albeit with a slight increase in bias tradeoff. Training is performed by using a variational lower bound on the free-energy term. It is one of the most efficient methods of learning the underlying structure of data [45].

The RBM can also be extended to model conditional or categorical distributions while keeping the same model structure [57]. For prediction, we are interested in the conditional distribution. Discriminative RBM has been developed to carry out prediction tasks by assuming one subset of observed variables as behavioural attributes and another set as dependent variables (either as continuous or discrete variables are possible). Fig. 2.4 shows various ways to configure an RBM for density estimation, prediction or behaviour modelling.

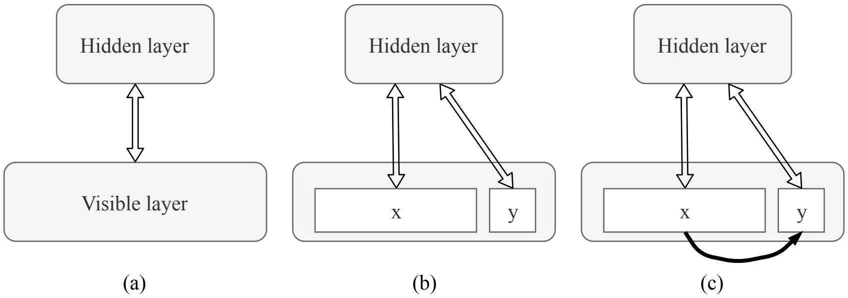


Figure 2.4: RBM framework which models different configurations of input and outputs, (a) generic RBM with 1 hidden and 1 observed layer, (b) the observed layer is divided into input and conditional output, (c) the output is defined as a function of input directly with the hidden layer acting as a regularizer.

2.2.2 Relationship between generative modelling and economic behavioural theory

Generative models explicitly allow learning of relevant features that represent the correlation and unobserved decision process in the data. Understanding and generalizing the underlying structure of the data is of fundamental importance in travel behaviour analysis. Hence, developing generative models and associated learning algorithms could potentially provide rich information and expand the scope of discrete choice modelling further.

Recent advancements in neural processing systems have established the link between utilitarian behaviour with information theory [41]. Specifically, a preference selection can be framed as an information processing constraint that accounts for the natural deviations in econometric behaviour theory [58, 18]. Generative models have the potential to combine individual entropy with a utility that forms the basis for economic interpretation. In simple terms, entropy refers to the underlying uncertainty that accounts for the natural deviations in econometric behaviour [18, 58]. The concept of entropy stems from information theory and can be explained through the same foundations as physical thermodynamics. In thermodynamics, the utility (expected energy) and entropy are related through the Hopfield energy function [59]:

$$F = U + TS \tag{2.8}$$

where U is the expected utility and T is a constant "temperature" term. F is the free energy term that defines how (biological or artificial) systems maintain non-equilibrium steady state within a limited number of system states. It was introduced as an explanation for variational free energy in machine learning by minimizing uncertainty [60]. Under the free energy principle, it has been suggested that human behaviour tends toward entropy – minimizing long term uncertainty [39]. In particular, the multinomial logit model is analogous to the Boltzmann distribution of free energy states:

$$p_i = \frac{e^{-F_i}}{\sum_j e^{-F_j}} \tag{2.9}$$

McFadden has shown that this distribution arises in decision making rules, with an additive noise that follows an extreme value distribution [43].

Individuals pursue a choice task that minimizes the difference between the likely and preferred action, the former represents what the analyst observes, and the latter represents the internal decision of the individual. These two states are often termed as "exploitative" (utility maximizing) and "explorative" (uncertainty minimizing) behaviour in neuroscience [16]. In some cases, individuals seek to minimize uncertainty, while in other cases, they would maximize utility. For example, in a route choice model, an individual traveller may choose a path that satisfies some internal constraints (i.e. habits) over routes that maximize utility (i.e. least travel time) [25]. Satisfying the decision maker's internal constraints reduces the perception of loss over taking an alternative route. Kahneman and Tversky formulated this choice behaviour as *Prospect Theory* [20]. Another example would be decision regret, described as *the inconsistency being brought on by mental shortcuts* [61]. Chorus et al. make the same analysis of this decision paradox and term the problem in discrete choice modelling as *random regret minimization* [32].

In the discrete choice framework, observed utility represents explicit actions while the unobserved component reflects the decision maker's holistic appraisal of the preferred action (which are hidden from the analyst) [20]. RUM theory is often criticized because it assumes a perfectly rational decision-maker and fails to predict behaviour between specific responses that are inconsistent with rational behaviour [61]. A radical idea that was introduced by Sims was to treat behaviour analysis in macroeconomic models as a type of dynamic programming problem with information processing constraints [58]. Sims suggested using information theory to measure *information flow as the rate of uncertainty reduction*. The substantial similarity between information theory in decision making and discrete choice models has been known for some time [62], but there have not been attempts to combine modern deep learning methods with behaviour modelling in a way such that it offers more convincing solutions for handling noisy, large-scale data-driven applications. Perhaps the most crucial question is, how do we formulate behaviour models that adapt to these constraints that would make sense in a data-driven environment?

The core hypothesis of this dissertation is that by framing the optimization between the observed value of choice and preferred (prior) action within a Bayesian framework, we can relate the process of learning through behavioural responses back to the conventional econometric theory of utility. The generative learning algorithm functions by modelling the joint distribution over the observed and latent constructs to describe how an ideal decision maker processes information [63]. In brief, generative modelling allows us to build a model of the internal decision states as *representing* some unobserved beliefs about the choices, expressed as the self-information associated with the choice, plus a divergence between the variational and posterior density.

Furthermore, in scenarios where the objective choice process is motivation driven (i.e. non-utility maximizing behaviour), by a need to satisfy or minimize uncertainties about the task. The computational approach will be as if individuals are actively minimizing uncertainty concerning prior beliefs about future outcomes while still acting on utility-maximizing behaviour. This is characterized as goal-independent *autonomous behaviour* [64]. This dissertation aims to take the classical methods of evaluating discrete choice models and incorporate a learning mechanism through a generative process, improving the consistency of the model estimates that represent autonomous behaviour [64].

2.2.3 How discrete choice analysis can benefit from using a generative modelling framework?

Modelling choice behaviour with latent variables

The use of psychometric data such as perception and attitudinal questions provides the basis for latent behavioural representation. Ben-Akiva et al. emphasized this importance affecting decision making, which leads to a more behaviourally realistic representation of the choice process [38]. Similar to [61], latent variables provide a means to synthesize models with the cognitive workings within the decision maker by including perceptual factors. These perceptual factors may take different forms. For example, in Mixed Logit, this takes the form of an adjustable variance parameter. In Random Regret Minimization, a regret factor is used, which incorporates value loss between competing alternatives [32]. In rational inattention models, time variability constraint that accounts for decision under limited information channel [25].

Even with the prevalence of advanced choice modelling methods such as the ICLV model, the consequences of behaviour modelling remain questionable [35]. Generative modelling provides a goal-driven method of extracting the underlying behaviour from the data observations and without explicitly using psychometric measurement indicators or using structured equation modelling [65]. Furthermore, we can use a generative model to characterize latent variables (alternative as well as individual specific) with semantic meanings and perform sensitivity analysis on the model parameters.

Latent variables in generative models are random variables that are learned from the data through a non-linear transformation function and a stochastic gradient based algorithm. Bayesian inference is used to derive the posterior distribution of the dependent variable, given some observed and latent variables. The generative model learns the underlying joint distribution of the observed and latent variables and optimizes a divergence objective function. This divergence objective function is used because of its plausibility to behaviour learning in a human brain [17, 16].

Modelling multiple discrete continuous outputs

The classical framework of discrete choice modelling assumes an individual selecting one or more alternatives from a slate of choices. In travel behaviour modelling, these alternatives might be mode (train, bus, car etc.), activity (education, employment, leisure, etc.) or location (home, school, work, etc.). The available alternatives may differ over time, and different subsets may or may not have all possible alternatives available to the individual; for example, transit options may not be available to individuals living in areas without access to transit. For example, travel distances are determined by the type of mode of transportation, and it is not plausible that a person chooses walking as their mode while having a long travel distance.

Often, the choice is a subset of available alternatives and not a single choice. A naïve way to consider modelling a subset of choices is to consider each subset as an alternative. However, this would exponentially increase the size of the model and estimation of such models becomes infeasible. Alternatively, one could model each output as an independent selection of choice sets, for example, using a sigmoid function. The downside of this method is that the model loses all correlation between each choice in the subset and does not capture any higher order correlation within the data. Higher-order interactions between observed data are well known to have a significant effect on the decision model. These problems have been addressed in machine learning literature by two approaches: First, by assuming only a restricted set of dependencies between variables [66]. This approach is mainly taken with graphical models such as RBMs. This allows individual tractable partitions and simplifying the estimation process [66]. The second approach is to approximate the form of the joint distribution that takes into account some of the dependencies between the variables, e.g. variational Bayesian inference [67].

Capturing information heterogeneity

The classical assumption about modelling travel behaviour data is that individuals have varying unobserved heterogeneity in their choice preference [68]. However, travel decisions tend to exhibit sensitivity to uncertainty and information processing constraints. Consider a case where an individual is faced with two route options in a choice set when the expected utilities are identical for both options. In utility theory, both options will be chosen at equal probabilities, where in practice, there is a constant stream of evaluations and changes to potential choice strategies as decision-makers' perception and choice process evolve [38].

In the context of data-driven models, behavioural heterogeneity describes the correlation between observed choice attributes and unobserved socio-economic factors using a flexible and tractable model specification. These variations include decision-protocols, choice sets, unobserved taste variations and unobserved attributes [26]. Recent studies on travel behaviour analysis have so far primarily focused on representing heterogeneity in the error correction function and incorporating it into utility based multinomial logit (MNL) models, for instance, mixed MNL models, Latent Class models or the Integrated Choice and Latent Variable (ICLV) model [28]. The use of psychological and perceptual factors in discrete choice modelling has been suggested to complement and extend, rather than replace a choice based perspective of economic theory [38]. A generative model can be developed in the same fashion that complements economic choice by extracting the correlation from observed data. However, the main difference is that it can be performed without an explicit definition of psychometric indicators. These latent behavioural constructs cannot be directly observed; however, by using Bayesian inference, we can quickly approximate the joint distribution through Gibbs sampling.

2.3 Summary

Research on travel behaviour under deep learning and neural networks in the era of Big Data is becoming an important aspect of understanding and anticipating the effects of disruptive mobility as it offers more flexibility and computational efficiencies in handling complex data structures. The supporting hypothesis of this dissertation is that generative modelling provides a new perspective on how analysts can obtain insights into behavioural heterogeneity manifestations by accounting for various heterogeneities from information processing cost, multiple variable dependencies and higher-order correlations. This chapter outlined the background theory and existing research in the context of this thesis. A proposed concept of using generative modelling, in particular, a generative model is developed to able to incorporate latent constructs while retaining economic features that are essential to behavioural analysis. While variational Bayesian inference addresses the shortcomings of MCMC methods, generative modelling methods can also be used to extend Mixed Logit through estimating a flexible underlying latent behaviour model without relying on pre-specified distributions.

Part II

Dissertation Articles

Chapter 3

Modelling Latent Travel Behaviour Characteristics: A Generative Machine Learning Approach

Preamble

This chapter introduces a RBM model on a stated/revealed preference survey study. We compare the effectiveness of the generative modelling approach against a conventional ICLV approach and analyzed the model parameter stability.

This research article appeared in IEEExplore:

Wong, M., Farooq, B., 2018. Modelling latent travel behaviour characteristics with generative machine learning, In: Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC), Maui HI, 2018, pp. 749-754. doi:10.1109/ITSC.2018.8569581

and was presented at the 21$^{\text{st}}$ IEEE International Conference on Intelligent Transportation Systems in November 2018.

Abstract

An information-theoretic approach to travel behaviour analysis is proposed using a generative modelling framework to identify informative latent characteristics in travel decision making. It involves developing a joint tri-partite Bayesian graphical network model using a Restricted Boltzmann Machine (RBM) generative modelling framework. We apply this framework on a mode choice survey data to identify abstract latent variables and compare the performance with a traditional latent variable model with specific latent preferences – safety, comfort, and environmental. Data collected from a joint stated and revealed preference mode choice survey in Québec, Canada were used to calibrate the RBM model. Results show that a significant impact on model likelihood statistics and suggests that machine learning tools are highly suitable for modelling complex networks of conditional independent behaviour interactions.

3.1 Introduction

The increased use of psychological and perceptual variables in travel choice survey has motivated several studies that investigated the explicit effects of latent behaviour in decision-making. Analysis of travel mode choice has focused on the effects of modal travel cost, time or reliability and many recent studies have attributed latent behaviour variables to account for unobservable effects [34, 69]. The Integrated Choice and Latent Variable (ICLV) model is a recent development in structural equation modelling (SEM) to handle hybrid endogenous and exogenous variables in decision-making [38]. The ICLV model has been shown – in some situations – to produce consistent estimates of model parameters, leading to better explanatory solutions [35]. The history of structural modelling dates back to the 1970s and has been initially used in psychology, sociology and market research, and recently it has seen growing applications in travel behaviour involving latent preference "attitudinal" variables and measurement "indicators". The fundamental methodology of SEM assumes prior statistical relevance and prior hypothesis about the subjective variables. Errors in measurement and model structure can be independently estimated, and psychological effects can be directed using measurement indicators. One of the characteristics of latent variable models is that the estimated model parameters are not always unique. The information quality of the underlying data also poses a significant identification problem in SEM. Theoretical analysis of how latent variables can be identified practically is a vital consideration, specifically in the domain of travel behaviour analysis.

Recent studies into some of the insights of the decision-making process with latent variables have investigated the use of machine learning algorithms to enrich limited endogenous variables by learning a generative statistical model explicitly designed to avoid the problems with non-unique parameter estimates. Efficient generative modelling algorithms, e.g., RBMs or Variational Autoencoders, developed for machine learning applications, can be applied to latent travel behaviour models without the need for measurement indicators nor through SEM by incorporating choice posteriors to learn latent variable interactions [70].

In this paper, we aim to develop a novel conditional RBM (C-RBM) for travel survey data that can leverage attitudinal and causal information of choice preference simultaneously. The main contributions of this paper are:

- Propose a C-RBM framework for travel behaviour modelling to incorporate the conditional relationship between observed and latent information.
- Explore the capability and identifiability of the framework for latent behaviour characteristics.
- An empirical comparison with traditional SEM-based discrete choice model.

3.2 Literature Review

3.2.1 Structured Equation Modelling (SEM)

The use of attitudes and perceptions in latent variable modelling have been used in various implementations and approaches in travel behaviour models [71, 72]. The ICLV model is a particularly useful SEM method that incorporates psychometric indicators by constructing a model in terms of a system of unidirectional effects of one variable to another [73]. Within this domain, ICLV models estimate either sequentially or simultaneously on latent variables and indicator manifestation to explain the utility of each alternative. The ICLV model combines consideration for unknown variables with the choice model, offering better explanatory effect.

Early developments of latent behavioural framework are a response to the need for interactions between psychometric data and choice preferences, treating behaviour as an "open black-box" [74]. A distinction in SEM is in the effects between observed and target perceptual variables that are pre-specified graphically. The direct effects of indicator measurements on latent variables are expected to be available. For instance, survey collection does not take into account psychometric factors; latent variables cannot be estimated using the ICLV method. Even when measurement indicators are available, they may be weak predictors of latent variables if the respondents do not understand or inaccurately answer those questions. Indicators may also not provide further useful information and might cause misspecification of the choice model [35]. ICLV explicitly models unobserved (latent) behaviour factors through measurement equations.

3.2.2 General specification of the ICLV model

The ICLV choice model is composed of 3 sub-parts: The choice model, the measurement model and the latent variable model [35]. A maximum likelihood estimation (MLE) function is used to estimate the parameter values. The observable variables consist of generic and alternative specific inputs where all the respondents n gives their stated choice preference i. The set of inputs are referred to as x_m and $x_{im'}$ respectively. In a standard RUM-based multinomial logit utility, the utility is defined by:

$$U_i = V_i + \varepsilon_i = \sum_m \beta_m x_{im'} + \sum_m \beta_{im} x_m + \varepsilon_i, \tag{3.1}$$

where β_m and β_{im} are the parameters of the alternative specific and generic variables respectively. These parameters define the sensitivity of each variable: $x_{im'}$ and x_m. Finally, ε_i is the extreme valued error term that represents the unobservable part of the utility and V_i represents the observed part of the utility. For simplicity, we assume that the alternatives are homogeneous across the population and parameters are estimated without a variance parameter.

The ICLV model extends the utility by adding a latent variable term x_h^* where h notation represents the number of latent variables used. An equation is defined for each latent variable. Random utility with latent variables can be defined as follows:

$$U_i = V_i + \sum_h \beta_{ih} x_h^* + \varepsilon_i. \tag{3.2}$$

Typically, the functions for latent variables are not explicitly defined beforehand. Here we provide several possible ways of how the latent variable can be formulated in terms of observable variables. The measurement model decouples the latent underlying factors from the observed variables and separate representations into discrete, measurable points, e.g. latent attributes such as ‘attitude towards owning a car’. Indicators I define the response of the individual to perceptual questions. For instance, one can ask the question ‘What is the importance of safety when choosing to travel by train?’ A Likert scale usually defines the response, and we assume that all indicators are configured as binary-valued $[0, 1]$ (e.g. not important or important). Similarly, in latent variables, each indicator is defined with one equation. Each equation measures the distribution of indicators conditional on the values of the latent variables, $f(x_h^*)$. For example, indicators can be defined as a conditional probability distribution of latent variables:

$$I_j = \sum_h \beta_{jh} x_h^* + \varsigma_j, \tag{3.3}$$

where ς_j represents the error terms of the indicators and β_{jh} is the parameters defining the weight of the latent variable on the specified indicator. The parameters β_{jh} can be estimated by the probability that the indicator is $I = 1$, using a binary logit model:

$$p(I_j = 1|x_h^*) = \frac{e^{\sum_h \beta_{jh} x_h^*}}{\sum_{j' \in [0,1]} e^{\sum_h \beta_{j'h} x_h^*}} \tag{3.4}$$

In principle, any function for I is possible (including linear when the indicator is a scale, hence a Probit model), but we consider a logit function for simple generalization. Providing indicators may help to capture the systematic response bias not found in observed variables. However, this method cannot be used if psychometric indicators are not available.

3.2.3 Modelling Non-linearity in Latent Variables

Non-linear interaction terms between latent and observed variables allow for cases where latent variables are not monotonically related to observed variables. The difficulty in computing the covariance or correlation matrices among non-linear terms of exogenous latent variables limits the use of non-linear functions, thus requiring non-linear constraints which increases the complexity of the model specification and identification. The linear function used in [35] is described in ICLV models [69]:

$$x_h^* = f(x_m; \hat{\beta}) = \sum_m \hat{\beta}_{hm} x_m + \vartheta_h, \tag{3.5}$$

where $\hat{\beta}_{hm}$ is the parameter describing the linear relation between observed and latent variables, and ϑ_h is a random stochastic term. This form is selected because the function will be linear and continuous and can be easily inferred from. However, there is a risk of overestimation as the value of x_h^* is not bounded ($x \in [-\infty, \infty]$).

There would be potential numerical instability in the gradient estimation procedure (when taking the exponential of a large number of input). A common practice in discrete choice modelling to stabilize parameter identification is to scale the input values to a small (<1.0) number or include a scale estimator [75].

A non-linear formulation is the sigmoid or inverse logit function

$$f(x) = sigmoid(x) = \frac{1}{1+e^{-x}}. \tag{3.6}$$

This is common for latent variables in discrete choice models since the output is continuous and bounded between 0 and 1. Intuitively, the value of the latent variable will represent a probability that the latent variable is included in the choice selection behaviour. The formula is as follows:

$$x_h^* = sigmoid(\beta_{hm}x_m) = \frac{1}{1+e^{-\sum_m \beta_{hm}x_m}} \tag{3.7}$$

Other possible functions of $f(x_m)$ include the rectifier model (commonly referred to as $Relu(x)$ in machine learning literature) is a threshold version of the linear function with $f(x) = Relu(x) = max(0, x)$ and the soft rectifier $f(x) = softplus(x) = \ln(1 + e^x)$ where $x = \sum_m \beta_{hm}x_m + \vartheta_h$ (Glorot et al. 2011). When $x = x_{m_1} - x_{m_2}$, the resulting output becomes a measure of alternative regret.

3.3 Framework and Estimation of Latent Variable Models

Generative models learn the underlying choice distribution $p(y)$ and latent variable distribution $p(x^*|y)$ given some input variables x. A Bayesian inference method is used to derive the posterior distribution of y given some observed and/or latent variable, e.g. $p(y|x^*) = \frac{p(x^*|y)p(y)}{p(x^*)}$. Latent variables are features that perform non-linear generalization of the highly heterogeneous observed data. Intuitively, in terms of econometric analysis, latent variables in generative models are arbitrary variables that depend on observed data, including response choices. In ICLV models measurement functions may be prone to errors. This is not so in the case of generative models, as latent information is inferred from choice data (through a Markov network, for example). The C-RBM is a variant of a Boltzmann machine inference model with an undirected energy-based model (from the basis of information theory and relative entropy) and a tri-partite of variables having symmetric connections.

The RBM framework estimates the amount of information 'bits' required to map the data onto the set of latent variables. Also, each group is conditioned on another set of inputs; in the case of an ICLV, the observed variables can be used as conditional inputs. The latent variables are assumed to be independent of each other and the model has stochastic visible variables $y \in \{0,1\} \forall \mathcal{Y}$ and latent variables $h \in \{0,1\}^J$ conditioned on some known prior distribution x. In discrete choice modelling, one of each constant or alternative specific parameters is fixed to zero. This can be performed in stochastic gradient learning by setting the gradient update to zero of the associated parameter in the computational graph.

The joint distribution of visible and latent variables is given by the Hopfield energy function:

$$\begin{aligned} Energy(y, x^*, x) = \sum_{i \in I} y_i c_i - \sum_{j \in J} x_j^* c_j - \sum_{i,j} x_j^* D_{ij} y_i \\ - \sum_{i \in I} x_{im} B_i - \sum_{j \in J} x_m G_{hm} \end{aligned} \tag{3.8}$$

where c_i and c_j are the constant values associated with the alternatives and latent variables respectively. D_{ij} is the parameter covariance matrix representing the relation between the latent and alternatives. B_i is the parameter vector of the conditional alternative specific inputs x_{im}. G_{hm} is the parameter matrix expressing the relation between latent and observed generic variables. Likewise, one parameter vector row is fixed to zero for model identifiability. We can express the Boltzmann distribution as an energy model with energy function which relates the entropy of the model to a specific state of the machine $F(y)$:

$$p(y) = \frac{1}{Z} \sum_{x^*} exp(-F(y)) \tag{3.9}$$

where Z is the partition function $Z = \sum_{i,j} \exp(-Energy(y, x^*, x))$ over all possible latent vector combinations. $F(y)$ is defined as the *free-energy* function:

$$F(y) = -\ln \sum_{x^*} \exp(-Energy(y, x^*, x)) \tag{3.10}$$

$$F(y) = -y_i c_i - \sum_{j \in J} \ln(1 + \exp(D_{\cdot,j} y + c_j)) \tag{3.11}$$

3.3.1 Objective function and likelihood estimation

To estimate an ICLV model, maximum likelihood (ML) is used most often. ML maximizes the probability that the structural model parameters generate the implied output, and the measurement model maximizes the probability that the underlying latent variables generate the associated indicators. To perform estimation of RBM type models, we need to define the objective that is robust and stable in the biases of the standard errors. A stochastic graph is constructed that incorporates both conditional dependence and the choice model. The C-RBM model learns aspects of an unknown probability distribution based on samples from that distribution. A stochastic gradient descent algorithm iterates across all observations and updates the parameter vectors such that the model best represents the distribution of the choice data (Algorithm 1). To generate latent variables, it is necessary to compute the log-likelihood of the joint distribution $p(y, x^*, x)$. Efficient Markov Chain Monte Carlo algorithm has been developed to deal with such problems using Gibbs chain sampling methods and contrastive divergence (CD). Assuming that individual responses are known, we can model the joint distribution of the responses and latent variables using the Bayesian estimation rule:

$$p(y) = \int_{x^*} p(y, x^*, x) dx^* = \int_{x^*} p(y|x^*, x) p(x^*|x) dx^* \tag{3.12}$$

The probability that the C-RBM model estimates are based on comparing the Kullback-Leibler divergence of the initial probability distribution $p(y)$ and another, final distribution $p(\hat{y})$, where $p(\hat{y})$ is the probability of the reconstructed representation after Gibbs sampling. To find the gradient derivative for the gradient descent training algorithm, we take the derivative of the log probability of the training vector with respect to the model parameters:

$$\begin{aligned}\frac{\delta \log p(y)}{\delta \theta} &=< y_i x_j^* >_{data} - < y_i x_j^* >_{model} \\ &= \phi^+ - \phi^- \qquad (3.13)\end{aligned}$$

where the components of $\langle y_i x_j^* \rangle$ correspond to the expected value under the specified distribution (data or model). The first and second terms are the positive and negative phases of the Gibbs sampling procedure. The update rule from the model parameters can be performed with stochastic gradient descent (SGD) at each iteration t:

$$\Delta\theta = \Phi(< y_i x_j^* >_{data} - < y_i x_j^* >_{model}) \qquad (3.14)$$

$$\theta_t = \theta_{t-1} - \Delta\theta \qquad (3.15)$$

We incorporate a learning factor Φ in the objective function which controls the magnitude of the update parameters. The objective assumes that the marginal $p(x^*|x)$ has a closed form solution and the function generate output samples $\hat{y} \sim p(\hat{y} = 1|x^*, x)$.

3.3.2 Construction of the latent behaviour choice model

The generated parameter vectors of the C-RBM model are then used to estimate a latent behaviour model that contains the utility-maximizing estimator for each observed and latent variables with an indicator model for the latent variable component:

- For the choice model y, the estimator simply calculates the likelihood $\mathcal{L}(\theta)$ under the RUM theory, that is $\mathcal{L}(\theta) = \frac{1}{n}\sum_n p(y|x^*, x; \theta)$. Here θ are the parameters of the generic (β_{im}), alternative specific (β_i) and latent (β_{hi}) variables.

- For the latent variable x^*, we calculate the conditional probability $p(x^*|x)$. A reparameterization boundary condition is placed on latent variables $[0, 1]$. The identified parameter represents the probability that the latent variable is present in the individual.

Algorithm 1: Conditional RBM Gibbs sampling procedure using Contrastive Divergence

```
Input  : Data sample $\mathcal{D}$, batch sample $S_i \subset \mathcal{D}$, $i = 1, ..., s$, iteration steps
         $T$
Output: Model parameters $\theta$.
initialize: $\theta = 0$;
forall $S_i \in \mathcal{D}, \tau = 1, ..., T$ do
    forall $(y, x^*, x) \in S_i$ do
        for $n = 1$ to $N$ do
            iterate over Gibbs chain, CD$_n$
            $< y, x^* >_{data} \leftarrow p(y_n, x^*_n, x_n)$
            Sample: $\hat{y} \sim p(y|x^*, x)$
            $< y, x^* >_{model} \leftarrow p(\hat{y}_n, x^*_n, x_n)$
        end
    end
    parameter update:
    $\Delta\theta \leftarrow \Phi(< y, x^* >_{data} - < y, x^* >_{model})$
    forall $\theta$ do
        $\theta_\tau \leftarrow \theta_{\tau-1} - \Delta\theta$;
    end
end
```

- For the indicator component, statistically significant latent variables are extracted from our C-RBM model estimation.

The choice model be can of any form, e.g. multinomial logit, mixed logit, nested logit, or a combination of different choice mechanisms (for simplicity, we use an MNL in our experiment). Once the choice and measurement model is formulated, the likelihood function is derived to estimate the parameters of the model. The likelihood function is defined as the mixed logit integral of the choice model conditional on the indicator measurement model:

$$P(y|x, x^*_h, I) = \int P(y|x, x^*)P(I|x^*)dx^* \tag{3.16}$$

Assuming that the measurement model follows the logistic sigmoid function with scale and/or translation factor, the integral can be estimated by maximum log-likelihood (MLE), and terms of the resulting densities are:

$$\begin{aligned} L(\theta) &= \log(P(y|x, x_h^*, I)) \\ &= \sum_n (\log(P(y|x, x^*)) + \sum_j \log(P(I|x^*))) \end{aligned} \tag{3.17}$$

The first term is the log-likelihood of the choice model. The second term can be substituted with cross-entropy (CE) maximization:

$$\log(P(I|x^*)) = I * \log(f(x^*|x)) + (1 - I) * \log(1 - f(x^*|x)) \tag{3.18}$$

The CE approach for multinomial logit models is equivalent to the standard log-likelihood for standard discrete-continuous choice models when more than one alternative is selected. In the case of the latent variable and indicator function, the probability of $I_j = 1$ is independent of other $I_{j'}$. This CE expectation-maximization procedure on a multi-attribute logistic function recovers the likelihood of the indicator model efficiently and directly evaluate $P(I|x^*)$ simultaneously with the choice model.

3.4 Case Study: Train Hôtel Mode Choice Dataset

A combined revealed and stated preference travel survey from commuters along the Northeastern USA rail corridor with Montreal in Canada (*Montreal*, *NYC*, *Maine*, *Boston*) is conducted (see Table D.1 for data description). A sleeper train between these cities and tourist destinations (*Train Hôtel*) was proposed to provide an alternative to the regular rail travel mode. The proposed *Train Hôtel* provides overnight sleeper amenities and entertainment for round-trip journeys shown in Fig. 3.1. A joint RP-SP survey design provides multi-attributed and generic variables, resulting in more accurate outcomes. The survey analyses mode choice preference of passengers who travelled between select Canada and USA destinations within 12 months prior from the day of the survey. The data statistics and collection procedure are described in [76].

In the SP choice survey, each respondent was presented with up to 6 alternatives $y_i \in \{Bus, Car\ Rental, Car, Plane, Train\ Hôtel, Train\}$. Each mode alternative

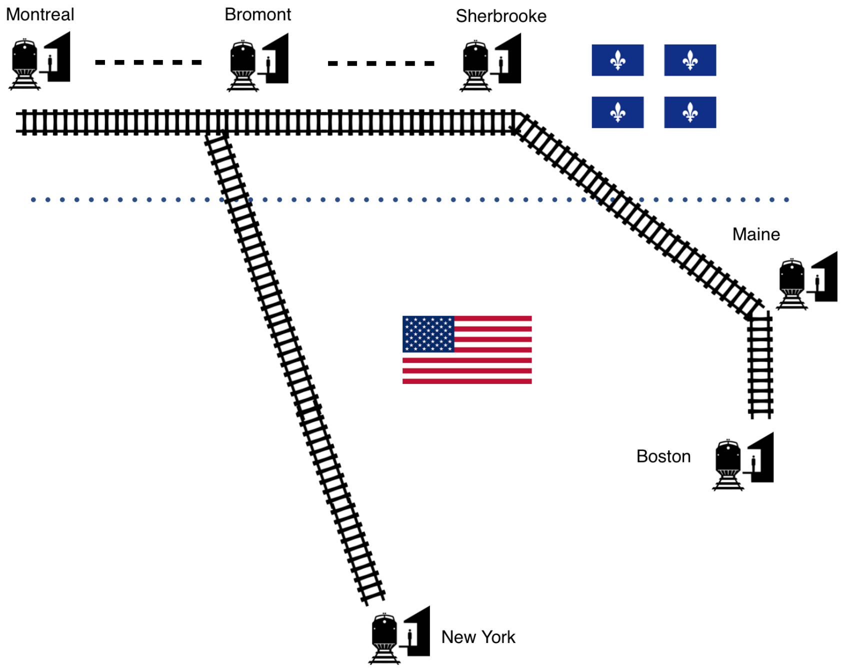


Figure 3.1: Origin Destination nodes modelled for computational experiments.

was characterized by trip duration, trip reliability and trip cost. Each attribute was sampled on different levels for each respondent (e.g. multiple price levels) defined relative to the origin and destination pairs. The level of each quantity was randomized across variables to control for potential ordering bias. However, the choice order between respondents was not varied. The second part of the survey data collected socio-economic and household characteristics of the respondents. The survey data consisted of continuous (e.g. income, age, number of vehicles) and categorical variables (e.g. education, household type). For consistency, all generic variables related to the respondents' characteristics were binary coded (continuous variables are first categorized). The model structure used in the analysis is shown in Fig. 3.2.

For the measurement model, three qualitative indicators were considered for each mode: environmental, comfort, safety, (e.g. safety of car, safety of plane). Respondents indicated their perception of these indicators by the level of importance on a 5-point Likert scale.

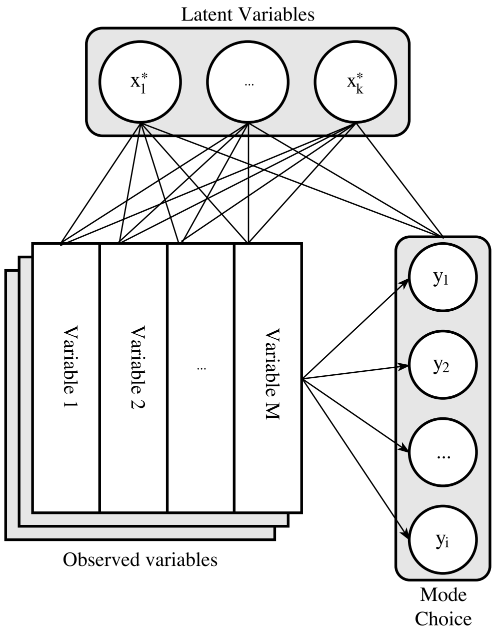


Figure 3.2: A joint tri-partite RBM structure for travel analysis with latent behaviour variables.

3.5 Results

The performance for the mode choice was compared as follows: First, we initialize a set of parameter values using the C-RBM method. Next, we constructed an ICLV model with interaction terms (ICLV) using the significant parameters. By initializing from an optimal non-zero point, we can avoid identifiability problems by having a higher probability of finding the global optima through the gradient estimation parameter search. For estimation using a stochastic gradient method, we fix the gradient for the reference parameter to zero, so updates are not backpropagated to the parameters; therefore, a reference value could be found. Table 3.1 shows the estimated parameters from the model estimation. Table 3.2 shows the optimal ICLV and C-RBM indicator estimates.

3.5.1 Latent Behaviour Model Formulation

We measure the reliability of the latent variable parameters by quantifying perceptual meaning (e.g. quality measure, attitudes towards a particular habit) to each latent variable that could be used as an additional explanatory variable in order to obtain a better fit on the choice model. The latent variables were then evaluated on their consistency through the measurement indicator model. Through this process, latent variables were hypothesized less subjectively since they were learned through the C-RBM model framework. We use significant latent variables as a guide for the construction of the ICLV model, assuming that there should be a relation between the posterior choice and prior distributions. Using the observed distribution of choice data instead of pre-defined latent variables in our estimated model removes assumed causal relation with subjective indicators.

The equations of the ICLV model follows a 6 alternative mode choice model (i) with three latent variables (x_1^*, x_2^*, x_3^*), the three latent variables were measured by specific mode indicator variable (e.g. $I_j \forall J \in \{$ *bus*, *car*, *train*, *plane* $\}$). The measurement indicators I_j were binary coded from a 5-point Likert scale (1, 2, 3 = not important (1), 4, 5 = important (0)) [76]. Latent variable interaction terms, denoted by the observed variables are formulated as:

- Environmental
  Variables: Driving Licence, Age 25-45, FT workers, HS Education, 0 HH Vehicles, 0 or 2 Children, Income ≥60K

- Safety
  Variables: Public Transit Pass, Age 25-45, 1 HH Vehicle, 0 or 1 Children, 20K< Income ≤60K

- Comfort
  Variables: Age ≥45, Male, Tertiary Education, 1 or more HH Vehicles, Income ≤20K, Income ≥60K

Measurement equation of the ICLV model:

$$I_j = f(x_h^*; \beta_{jh}) \tag{3.19}$$

Structural equation of the ICLV model:

$$U_i = \sum_m \beta_i x_{im} + \sum_h \beta_{ih} x_h^* + c_i \tag{3.20}$$

3.5.2 Model analysis

The results of the two-stage approach showed that three attitudinal variables could be included for estimating latent behavioural aspects for travel mode between Montreal and Northeastern USA destinations. The signs of the indicator parameters are as expected in the C-RBM model, and the t-tests show that most parameters are significant. In the ICLV model, the high positive values of the latent variable parameters indicate that individuals are most sensitive to the comfortability of train mode, likewise for environmentally conscious behaviour, improving the environmental impact of train mode also have the highest impact on perception, while car and plane mode had the least effect. For C-RBM, experiments on a different number of latent variables also showed convergence and identification problems as some latent variables were found to be identical or very similar. The results for the estimation of SP variables (cost, time and reliability) are shown in Table 3.3. Both models are consistent in cost, travel time and reliability parameter values.

Assuming similar parameter estimates, the latent constructs observed in the C-RBM models indicated that the effects of choice on latent variables lead to a more accurate representation of behaviour. Comparing the standard ICLV and C-RBM method, there is a higher model fit when the parameters are initialized well before model estimation. Our case study reveals the feasibility of the C-RBM framework on mixed RP and SP data which could account for perception effects related to SP values and attitudinal questions. An essential advantage of this is to be able to estimate the values from the data instead of postulating them. However, we should mention that neither method is sufficiently reliable, but provides a different perspective that is representative of the underlying latent variables. The statistical results show that our method has superior performance in terms of estimated log-likelihood (-1946.872 vs. -2013.685). This shows that alternative specific variables do not have any significant variance when incorporating latent variables; Estimating the

Table 3.1: Optimal Parameters for ICLV and C-RBM Choice Model

	ICLV model			C-RBM latent behavior model		
parameters	comfort	env.	safety	comfort	env.	safety
DrvLicens	-	1.576	-	-0.054	-0.387	2.014
		(5.27)		(-0.24)	(-2.37)	(3.64)
PblcTrst	-	-	-2.327	0.331	0.561	1.597
			(-3.10)	(1.278)	(2.53)	(3.44)
Ag1825	-	-	0.462	-2.087	-2.199	-0.039
			(0.23)	(-3.47)	(-5.36)	(-0.05)
Ag2545	0.691	-1.345	3.976	-1.329	-0.641	2.242
	(1.96)	(-4.41)	(3.00)	(-4.64)	(-3.02)	(3.28)
Ag4565	-	-	-	3.618	1.868	1.463
				(6.97)	(4.44)	(2.20)
Ag65M	-	-	-	0.535	0.07	1.511
				(1.06)	(0.185)	(1.55)
Male	-1.493(-	-	-	-2.453	-1.509	-2.584
	4.53)			(-7.79)	(-7.04)	(-6.00)
Fulltime	-	-0.413	-	-0.756	-0.468	-0.512
		(-1.29)		(-2.85)	(-2.38)	(-1.05)
EduHighschl	-	1.614	0.701	-0.492	0.613	-0.930
		(2.02)	(0.36)	(-0.93)	(1.634)	(-1.49)
EduBSc	1.298	2.062	-	1.683	1.397	-1.470
	(3.82)	(5.76)		(6.45)	(6.90)	(-3.02)
EduMscPhD	-	-	-	0.189	0.788	2.861
				(0.39)	(2.37)	(0.47)
HHVeh0	1.966	1.101	-	-2.392	0.462	-0.404
	(1.67)	(1.36)		(-4.36)	(0.73)	(-0.91)
HHVeh1	-0.531	-	4.648	2.494	0.513	1.786
	(-1.83)		(2.66)	(8.53)	(2.54)	(2.34)
HHVeh2M	1.783	-	-1.129	2.55	0.685	4.237
	(2.04)		(-1.21)	(7.01)	(2.52)	(1.26)
HHChld0	-	2.941	3.514	-0.794	0.534	2.933
		(7.94)	(1.27)	(-3.26)	(3.04)	(5.842)
HHChld1	-	0.219	-4.078	-1.288	-0.056	1.594
		(0.45)	(-5.75)	(-2.81)	(-0.16)	(1.77)
HHChld2M	-	2.508	-	2.786	3.09	-0.184
		(3.23)		(2.78)	(2.26)	(-0.24)
HHInc020K	3.468	-	-	0.312	2.15	-1.852
	(2.66)			(0.61)	(2.57)	(-4.03)
HHInc2060K	-	-	-1.364	3.486	2.701	4.297
			(-1.17)	(6.67)	(4.52)	(1.38)
HHInc60KM	-0.878	-1.562	-	-1.045	-0.384	-0.846
	(-2.75)	(-3.73)		(-3.681)	(-1.99)	(-1.164)

note: for ICLV model, non-significant parameters were removed and re-estimated

* *values in brackets are t-test values at >95% statistical significance*

Table 3.2: Optimal ICLV and C-RBM indicators

	ICLV model			C-RBM latent behavior model		
parameters	comfort	env.	safety	comfort	env.	safety
Bus	1.281	-0.802	0.509	-1.638	2.521	-1.479
	(7.65)	(-3.35)	(2.44)	(-5.72)	(10.83)	(-6.53)
Car Rental	1.778	-2.236	0.597	-2.474	2.686	-3.900
	(6.08)	(-5.37)	(1.78)	(-5.24)	(6.96)	(-10.84)
Car	-1.861	1.627	-0.01	2.904	-4.017	-0.161
	(-17.53)	(12.79)	(-0.09)	(21.05)	(-29.93)	(-1.34)
Plane	-1.635	0.683	0.921	-1.905	0.504	2.655
	(-7.87)	(2.80)	(4.35)	(-6.13)	(1.91)	(12.39)
Train Hôtel	0.646	0.147	1.420	-1.376	1.823	2.448
	(5.83)	(1.34)	(14.35)	(-11.17)	(16.30)	(24.27)
Train	0 (ref.)	0 (ref.)	0 (ref.)	0 (ref.)	0 (ref.)	0 (ref.)
LV1_Comf_Car	0.533	-	-			
	(4.98)					
LV1_Comf_Train	2.16	-	-			
	(13.29)					
LV1_Comf_Bus	-0.88	-	-			
	(-7.08)					
LV1_Comf_Plane	-0.241	-	-			
	(-2.28)					
LV2_Envrn_Car	-	-1.711	-			
		(-11.83)				
LV2_Envrn_Train	-	1.083	-			
		(8.94)				
LV2_Envrn_Bus	-	-0.657	-			
		(-7.81)				
LV2_Envrn_Plane	-	-1.78	-			
		(-12.04)				
LV3_Safe_Car	-	-	0.619			
			(6.14)			
LV3_Safe_Train	-	-	2.215			
			(13.79)			
LV3_Safe_Bus	-	-	0.575			
			(5.74)			
LV3_Safe_Plane	-	-	0.611			
			(6.06)			

note: for ICLV model, non-significant parameters were removed and re-estimated

* *values in brackets are t-test values at >95% statistical significance*

Table 3.3: Alternative Specific Variables and Constants Estimated Within the ICLV and C-RBM Models

	ICLV			C-RBM		
Parameters	value	std. err.	t-test	value	std. err.	t-test
ASC_Bus	-2.485	0.204	-12.179	0.266	0.209	1.273
ASC_CarRental	-2.243	0.33	-6.802	1.319	0.335	3.934
ASC_Car	0.643	0.115	5.619	1.236	0.118	10.507
ASC_Plane	-0.318	0.208	-1.531	-0.779	0.209	-3.732
ASC_TrH	-0.386	0.097	-3.967	-0.452	0.098	-4.609
ASC_Train	0 (ref.)	-	-	0 (ref.)	-	-
cost	-0.609	0.112	-5.447	-0.595	0.114	-5.217
travel time	-0.127	0.023	-5.477	-0.131	0.024	-5.541
reliability	0.249	0.684	0.364	0.42	0.692	0.606
Model statistics						
Null Loglikelihood		- 2917.752			- 2917.752	
Final Log-likelihood		- 2013.685			- 1946.872	
rho square		0.310*			0.332*	
AIC		4273.371			4139.744	
BIC		4948.5			4814.873	

* *note that the functions governing the relationship between* y and x, x^* *are different, so we cannot compare rho square values directly*

model through a joint estimation method does not generally influence the underlying factors of alternative dependent cost, time and reliability variables.

Under the assumption that the latent behaviour function is non-linear and complex, there may be multiple locally optimal solutions, and we did not consider scale and translation effects of the underlying variables and different decision rules in this study. We are currently investigating these effects in our future work. Furthermore, these are important considerations when the structure of the latent variable model changes from a conditional to a joint model.

3.6 Conclusion

In this paper, we develop a new approach to the problem of modelling latent behaviour through the estimation of a joint distribution from its associated choice and auxiliary information. This approach has been studied in different contexts in machine learning models recently. Our C-RBM approach is the first fully developed solution to latent behaviour models. This approach is comparable with previously developed ICLV methods in terms of model fit and does not require additional parameters. The estimation process is straightforward, and convergence is fast for large parameter vectors using stochastic gradient descent with CD objective function.

In a sample of a new travel mode choice, survey respondents were asked to indicate their preference of travel mode given that a hypothetical intercity train service *Train Hôtel* is offered. *Train Hôtel* provides overnight sleeper amenities as an alternative to day trains and other modes such as cars, planes and buses. Results obtained from the C-RBM parameter estimation are compared with results from the ICLV model.

The ICLV approach analytically derives each latent variable under assumptions on the measurement functions. This method is useful when indicators are available, but assumptions may be hard to verify as we are unsure about the interactions of the underlying latent variable generating process. In some cases, theoretical result only gives asymptotic guidance in finite observations. It is likely that the stated indicators may not be reflective of real attitudes and perceptions and heavily influenced by the survey conditions, geographic area, socio-demographics or other revealed information. While these model design choices are data-reliant, having a reliable estimate is a requirement for strong econometric plausibility. In our case study, inference on model performance is a straightforward task of analyzing parameter validity.

Chapter 4

Discriminative Conditional Restricted Boltzmann Machines for Discrete Choice and Latent Variable Models

Preamble

This chapter broadens the idea of RBM based generative modelling towards incorporating behavioural semantics in choice modelling without using subjective psychometric indicators. Rather than using perceptual or attitudinal factors in the measurement equations, we let the generative model automatically tweak and adapt the latent structure to observed variables. The examples we provide in this chapter illustrate the estimation and model development process on a consumer choice preference panel dataset using a Conditional-RBM model that seeks to explain and forecast consumer motivations.

This research article appeared as:

Wong, M., Farooq, B., Bilodeau, G.-A., 2018. Discriminative conditional restricted Boltzmann machine for discrete choice and latent variable modelling, Journal of Choice Modelling, 42, pp. 152-168. doi:10.1016/j.jocm.2017.11.003

Abstract

Conventional methods of estimating latent behaviour generally use attitudinal questions which are subjective and these survey questions may not always be available. We hypothesize that an alternative approach such as non-parametric artificial neural networks can be used for latent variable estimation through an undirected graphical models. In this study, we explore the use of generative non-parametric modelling methods to estimate latent variables from prior choice distribution without the conventional use of measurement indicators. A restricted Boltzmann machine is used to represent latent behaviour factors by analyzing the relationship information between the observed choices and explanatory variables. The algorithm is adapted for latent behaviour analysis in discrete choice scenario and we use a graphical approach to evaluate and understand the semantic meaning from estimated parameter vector values. We illustrate our methodology on a financial instrument choice dataset and perform statistical analysis on parameter sensitivity and stability. Our findings show that through non-parametric statistical tests, we can extract useful latent information on the behaviour of latent constructs through machine learning methods and present strong and significant influence on the choice process. Furthermore, our modelling framework shows robustness in input variability through sampling and validation.

4.1 Introduction

Complex theories of decision-making process provide the basis for latent behaviour representation in statistical models. These processes focus on the use of psychometric data, such as choice perception and attitudinal questions. Although it can provide essential insights into the choice process and underlying heterogeneity, studies have shown limited flexibility and benefits of statistical latent behaviour models, i.e. Integrated Choice and Latent Variable (ICLV) models [46, 35]. Three disadvantages are well known in ICLV models: First, datasets are required to have attitudinal responses, for example, Likert scale questions in product choice surveys. Second, model misspecification may occur when latent variable model equations are poorly defined. Lastly, attitudinal questions are subjective and may change over time.

The objective of this study is to use advanced machine learning (ML) methods to analyze underlying latent behaviour in decision-making based on a set of synthetic ML considerations and hyperparameters without explicitly using attitudinal or perception attributes. A growing body of choice modelling behaviour research focuses on patterns and clusters of behaviour characteristics such as latent attitudes and choice perceptions. Even with the prevalence of advanced choice modelling strategies such as ICLV models, our knowledge of the consequences of latent behaviour in the choice model remains limited [35]. Studies on hidden representations using neural network models may provide more nuanced and potentially new perspectives of latent variables in discrete choice experiments and choice behaviour theory [77]. Given many possible latent variable combinations, it is necessary to use advanced ML techniques to segment the population into groups with similar attitudinal profiles. For this study, we have chosen to use restricted Boltzmann machines (RBM). RBM is a non-parametric generative modelling approach that seeks to find latent representations within a homogeneous group by hypothesizing that posterior outputs can be explained with a reduced number of hidden units [78]. Besides, identifying useful latent variable representations may enable policymakers to better understand the sensitivity and stability of latent behaviour models in surveyed and revealed preference data. We decouple the latent behaviour model underlying the data distribution by estimation on a financial instrument choice behaviour dataset without the need for subjective measurement indicators. The proposed method does not predefine semantic meanings for each latent variable. Instead, we construct a restricted Boltzmann machine to learn the latent relationships and approximate the posterior probability.

We show in our findings that our RBM model approach can characterize latent variables with semantic meaning without additional psychometric data. The parameters estimated through our RBM model present a strong and significant influence in the decision-making process. Furthermore, sensitivity analysis has shown that this method is robust to input data variance and the use of generated latent variables improves sampling stability.

The remainder of the paper is organized as follows: Section 4.2, provides a background literature review on latent behaviour models. Section 4.3 describes the conditional RBM modelling approach and model training methodology, given only observed variables without attitudinal questions. Section 4.4 explains the data and the experimental procedure. Section 4.5 presents the results and performance tests and analyzes the model sensitivity and stability. Finally, Section 4.6 provides a conclusion and future research directions.

4.2 Literature Review

Current practice in choice modelling is targeted at drawing a conclusion on the mechanism of the stochastic model and not so much about the nature of the data itself. This leads to simple assumptions of data relevance and statistical properties of explanatory variables [79]. Several parametric and non-parametric modelling methods are available. Parametric models are regression-based and random utility maximization structural models. Examples of non-parametric methods include latent class and variable models, k-means or hierarchical clustering. These non-parametric methods are often criticized for being too descriptive, theoretical, may result in inconsistent estimates and often not possible to make generalizations [11, 80, 81]. Analysis of data through the statistical properties is generally applied for extracting information about the evolution of the responses associated with stochastic input variables rather than having good prediction capabilities. On the other hand, algorithmic modelling approaches such as artificial neural networks (ANN), decision trees, clustering and factor analysis are based on the ability to predict future

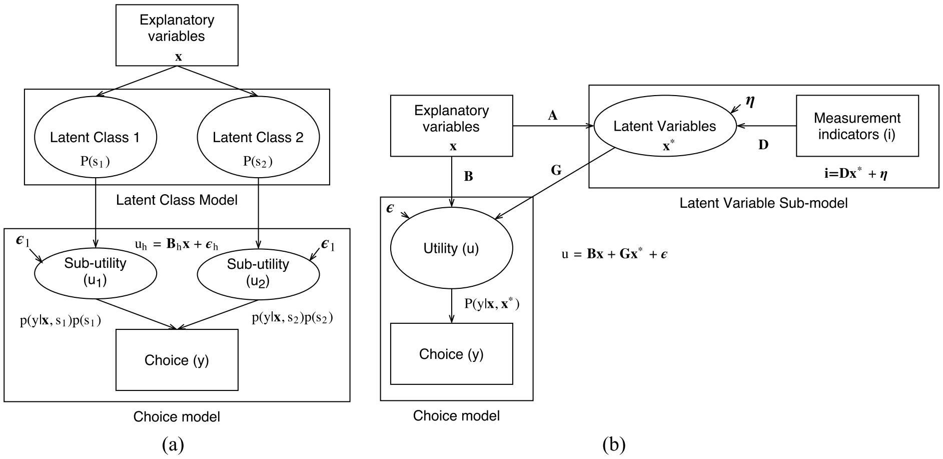


Figure 4.1: Classical structural framework for (a) latent class model and (b) integrated choice and latent variables model

responses accurately given future input variables within a ‘black-box’ framework [15]. Econometric choice models can be estimated by using both parametric and non-parametric methods that incorporate machine learning algorithms into discrete choice analysis to learn mappings from latent variables to posterior distribution [82].

Several different approaches that implement the use of attitudinal variables have been used in existing literature [72, 83, 84]. 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, these assumptions are based on multiple sources of unobserved heterogeneity influencing decisions, e.g. inter- and intra-class variance and ‘agent effect’ [85]. Fig. 4.1 illustrates the Latent Class 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 that assumes a discrete distribution among market segments [86]. 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 the assignment of latent class probabilities. Individuals in the same class share a common joint probability distribution among the observed variables. Under the assumption of class independence, the utilities are generated with a prior hypothesis from several subpopulations, and each subpopulation 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. 4.1). 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 [28].

The use of attitudes and perception latent variables are also particularly attractive and popular in past work [87, 80]. 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 used as principal components [87]. The approach involves the analysis of the relationship between indicators and the choice model. Within this domain, there is the sequential and simultaneous estimation process. The sequential approach estimates a measurement model that 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, which were not taken into account during the initial choice model.

To solve this issue, another approach uses a simultaneous estimation of the 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. 4.1). 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 [88], image, stress and safety attitudes towards cycling [89], and social attitudes towards electric cars [90]. However, the simultaneous approach still relies on a separate measurement model (latent variable model) that 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 [35]. Even when measurement indicators are available, they may not provide any further information that directly influences the choice than through explanatory variables [46]. Consequently, misspecification and other measurement errors may occur when the criteria are not associated with the choice model.

Without measurement indicators to guide the selection of latent variables, we can alternatively use ML for latent variables through data mining. This can be implemented through generative modelling methods used in ML. Generative modelling is a method in ML that uses underlying data to generate latent features or classes through supervised (labelled) or semi-supervised (partially or unlabelled) learning. In our process, generative models estimate the underlying choice distribution $p(y)$ and the latent inference $p(h|y)$, where h is the latent variable. Following which, we implement a Bayesian network that represent a probabilistic conditional relationship between random variables and dependencies to derive the posterior distribution of y given h using $p(y|h) = \frac{p(h|y)p(y)}{p(h)}$. Therefore, efficient algorithms that perform ML and inference, such as RBMs can be used in this method. The denominator is given by $p(h) = \sum_y p(h|y = 1)$ indicating choice y is chosen. The rapid advancement of machine learning research has led to the development of efficient semi-supervised training algorithms such as the conditional restricted Boltzmann machine (C-RBM) [49, 57], a hybrid discriminative-generative model, capable of simultaneously estimating a latent variable model using a priori choice distribution with an latent inference model (Fig. 4.2).

To date, econometric and machine learning models are often studied for its contrasting purposes in decision forecasting by behavioural researchers [15]. Econometric models are based on the classical decision theory that an 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 actual decisions. The generative modelling based approach uses clustering and factor analysis developed through algorithmic modelling of the data. Associations between decision factors can be classified in this method, obtaining latent information without explicit definition of latent constructs [91]. Thus, machine

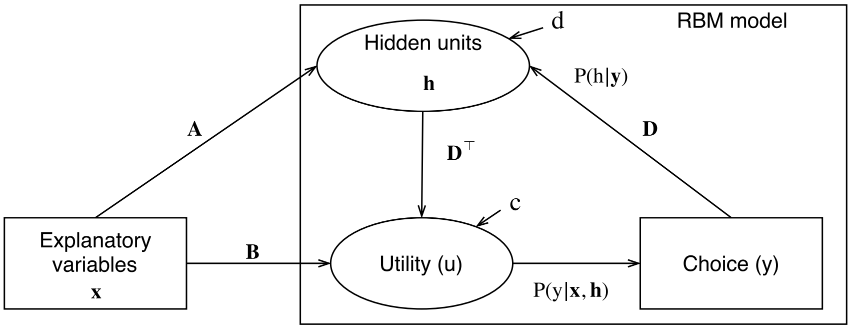


Figure 4.2: Framework for a C-RBM choice model conditional on explanatory variables and choice distribution.

learning algorithms such as ANN that decouple latent information from 'true' distribution generally outperform traditional regression-based models in multidimensional problems [92]. Recent works on latent behaviour modelling on choice analysis agree on the potential of improving behaviour models with machine learning. Examples include combining machine learning to improve complex psychological models [93], representing the phenomena of similarity, attraction and compromise in choice models [94] and inference of priorities and attitudinal characteristics [95].

Despite the many benefits, the interpretation of results is still challenging due to the complexity and number of parameters in ML analysis. As a result, ML models are not often used for general purpose behaviour understanding but created exclusively for a specific purpose for prediction accuracy. Furthermore, with the emphasis on applications and theoretical studies in today's massive data-driven industry, improving analytical techniques with ML is very relevant, although structural modelling, statistical and probability theory will remain the cornerstone of discrete choice analysis.

4.2.1 The basis of latent class and latent variable models

The latent class model shown in Fig. 4.1 is a simple top-down model that imparts generalization properties to the choice model that predefines a discrete number of classes, allowing the parameters to vary with the fixed distribution. Formally, the

LCM choice probability can be expressed as:

$$P(y) = \sum_{n} P(s_n)P(y|x, s_n) \tag{4.1}$$

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. $P(y|x, s_n)$ is the conditional probability of choice y selected given the class s_n and input variable x.

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 [96]. Latent variables can be classified as either attitudinal (individual characteristics) or perceived (personal beliefs towards responses) [11]. 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 useful parameters and predict accurate decision outcomes when there is a lack of a strong significant correlation between explanatory variables and choice outcomes. The functions of the structural and measurement model can be explained in these four equations [35]:

$$\mathbf{x}^* = \mathbf{A}\mathbf{x} + \boldsymbol{\nu} \tag{4.2}$$

$$\mathbf{I}^* = \mathbf{D}\mathbf{x}^* + \boldsymbol{\eta} \tag{4.3}$$

$$\mathbf{u} = \mathbf{B}\mathbf{x} + \mathbf{G}\mathbf{x}^* + \boldsymbol{\epsilon} \tag{4.4}$$

$$y_i = \begin{cases} 1 \text{ if } u_i > u_i' \text{ for } i \in \{1, ..., I\} \\ 0 \text{ otherwise} \end{cases} \tag{4.5}$$

where u_i is the utility of selecting alternative i. $\mathbf{A}$ is a $(k \times j)$ matrix representing the relationship between the k^{th} input explanatory variable in $\mathbf{x}$ and the j^{th} latent variable in $\mathbf{x}^*$. $\mathbf{D}$ is a $(j \times l)$ matrix representing the relationship between the j^{th} latent variable in $\mathbf{x}^*$ and the l^{th} indicator output in $\mathbf{I}^*$. $\mathbf{I}^*$ is the psychometric indicators of the respondent, $\mathbf{I}^* =: [0, 1]$. For instance: "Reliability is important in my travel decision, 0: no, 1: yes". $\mathbf{B}$ and $\mathbf{G}$ represents the model parameter matrices int the utility with sizes $(k \times i)$ and $(j \times i)$ respectively.

$\boldsymbol{\nu}$, $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are the stochastic error terms of the model, assumed to be mutually independent and Gumbel distributed. In a generative model, there are no measurement indicators, however, the indicator parameters $\mathbf{D}$ can still be estimated by setting $\mathbf{G} = \mathbf{D}$, and $l = i$ (Fig. 4.2). Therefore, instead of optimizing the latent variables w.r.t. the indicators, the generative model generates latent variable samples h and optimizes the latent variables distribution w.r.t. the joint distribution $p(y, h)$.

4.2.2 Modelling through generative machine learning methods

In generative machine learning models, hidden units h are the learned features (Fig. 4.2) which performs non-redundant generalization of the data to reduce high dimensional input data [97]. Intuitively, in terms of econometric analysis, hidden units are latent 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 variables. In the case of factor analysis approach, a typical process is to perform feature extraction based on statistical hypothesis testing to determine whether 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 [91, 98]. The learned classes (or clusters) can then be introduced directly into the model via parameterization. In a generative modelling approach, we use the priors directly to learn the distribution of the hidden units. In this process, we extract latent information directly from the observed choice data instead of using measurement functions which may be prone to common misspecification errors.

4.2.3 Balancing model inference and accuracy

One common problem that researchers face when constructing latent behaviour models is specifying the optimal size of latent factors [99]. Since the hypothesis on the number of latent sizes cannot be tested directly, typical statistical evaluation methods such as AIC and BIC are used to guide class selection [99], in the case of ICLV models, through predefinition of measurement functions [77]. However, since the number of latent factors determines the ability of the model to represent the various heterogeneity in the data. It is likely that as we increase h, the choice model becomes more efficient in capturing complex behaviour effects from individual and latent attributes. On the other hand, if we increase the number of latent segments, the number of parameters will also increase at an exponential rate [99]. Therefore, we may gain model accuracy, but we would lose model interpretability.

The trade-off between inference and accuracy is a challenge when dealing with complex data [15]. If the goal of latent behaviour modelling is to leverage data to understand underlying statistical problems, we have to incorporate implicit modelling methods in addition to describing explicit structural utility formulations.

4.3 Methodology

In this section, we provide a summary of restricted Boltzmann machines and how it can be used to generate conditional priors from the choice distributions. We refer readers to [45] for background and details on generative models and deep learning.

4.3.1 Restricted Boltzmann machines

The Restricted Boltzmann Machine (RBM) formalizes the energy-based modelling principle for the development of a computational representation to describe a set of abstract descriptions about the data to its fundamental exogenous properties [49]. The development of RBM models is based on two primary methods: 1) describing a joint probability distribution as an energy function where random variables in this distribution interact with each other, and 2) iterative stochastic search algorithm that minimizes the global energy of the system.

The units of a Boltzmann Machine are divided into two groups: a set of observable variables and a set of latent representational features. The two groups are connected to form a symbolic interaction (Fig. 4.2). Observed variables describe the characteristics of the data and the latent variables describe the undetermined cognitive behaviour(s). The states of the RBM can be viewed as an accept-reject hypothesis of each latent-observable unit pair in the system. The weights of the links describe a pairwise constraint between the accept-reject hypothesis, whereby a positive weight indicates that the connected units support each other. Likewise, assuming all other factors remain the same, a negative weight suggests that the connection pair is not accepted. Therefore, the energy of the RBM is the sum of all the terms weighted by the strength of the connections.

The Boltzmann energy function has mathematical properties that are closely related to probabilistic decision theory, which makes it plausible to represent the systematic connections between latent and observed behaviour. The learning algorithm searches for an optimal solution to the model structure by iteratively updating states of the RBM to represent all possible combinations of individual pairwise hypotheses until it settles to a stationary equilibrium state representing a model of the data. The model has stochastic visible variables $\mathbf{y} \in \{0,1\}^I$ and stochastic hidden variables $\mathbf{h} \in \{0,1\}^J$. The joint configuration $\mathbf{y}, \mathbf{h})$ of visible and hidden variables is given by the Hopfield energy [59]:

$$\mathrm{E}(\mathbf{y}, \mathbf{h}) = -\sum_{i \in vis} y_i c_i - \sum_{j \in hid} h_j d_j - \sum_{i,j} h_j D_{ij} y_i, \tag{4.6}$$

where d_j and c_i represent the vector biases (constants) for the hidden and visible vectors respectively. D_{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 $F(\mathbf{y})$:

$$p(\mathbf{y}, \mathbf{h}) = \frac{1}{Z} \exp(-F(\mathbf{y})), \tag{4.7}$$

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

$$F(\mathbf{y}) = -y_i c_i - \sum_{j \in hid} \ln(1 + \exp(D_{\cdot,j} y + d_j)). \tag{4.8}$$

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

$$p(\mathbf{y}) = \frac{1}{Z} \sum_{\mathbf{h}} \exp(-F(\mathbf{y})). \tag{4.9}$$

The RBM model is used to learn the features 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 represents 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 the Contrastive Divergence (CD) method that starts a sampling chain (Gibbs sampling) from real data points instead of random initialization [100].

4.3.2 Model estimation and 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 $p(\mathbf{y})$, we take its gradient derivative of the log probability of a training vector with respect to the model parameters as follows:

$$\frac{\partial \log p(\mathbf{y})}{\partial \theta} = \langle y_i h_j \rangle_{train} - \langle y_i h_j \rangle_{model} = \phi^+ - \phi^-, \tag{4.10}$$

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

$$\Delta\theta = \Phi(\langle y_i h_j \rangle_{train} - \langle y_i h_j \rangle_{model}). \tag{4.11}$$

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

$$\theta_t = \theta_{t-1} - \Delta\theta. \tag{4.12}$$

To obtain a sample of a hidden unit from $\langle y_i h_j \rangle_{train}$, we take a random training sample $\mathbf{y}$ and sample the state in the hidden layer given by the following function:

$$p(h_j = 1|\mathbf{y}) = \frac{e^{d_j + \sum_i D_{ij} y_i}}{1 + e^{d_j + \sum_i W_{ij} y_i}} = \sigma(d_j + \sum_i D_{ij} y_i), \tag{4.13}$$

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:

$$p(y_i|\mathbf{h}) = \frac{e^{c_i + \sum_j D_{ij} h_j}}{\sum_i e^{c_{i'} + \sum_j D_{i'j} h_j}}. \tag{4.14}$$

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 $p(y) = \sum_h p(h)p(y|h)$ [101]. Due to its bidirectional structure, this framework possesses functional 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.

The model can be inferred from $\langle y_i h_j \rangle_{model}$ by setting the states of the visible variables to a training sample, and then the states of the hidden variables are computed using Eq. (4.13). Once a "state" is chosen for the hidden variables, a "reconstruction" phase produces a new vector $\tilde{\mathbf{y}}$ with a probability given by Eq. (4.14) and the gradient update rule is given by:

$$\Delta\theta = \Phi(\langle y_i h_j \rangle_{train} - \langle y_i h_j \rangle_{reconstruction}). \tag{4.15}$$

We approximate the gradient function by using a CD Gibbs sampler minimizing the divergence between the expected and estimated probability distribution, known as the Kullback-Leibler (KL) divergence [101]. A divergence ratio of 0 indicates that the distributions are entirely similar. The training algorithm ran for a total number of N chain steps and was initialized from a fixed point from the data distribution and then averaged across all examples [102].

4.3.3 Modelling approach

In this paper, the proposed method uses a conditional RBM (C-RBM) training algorithm to include input-output connections that allows for discriminative learning [103]. C-RBM expands the model to include "context input variables", 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. (4.14) does not reconstruct these explanatory variables. The influence factor 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 $\mathbf{y}$, conditional on $\mathbf{x}$ (4.2). In the choice prediction stage, a vector of new input samples $\mathbf{x}$ generate latent variables $\mathbf{h}$. Conditional on the explanatory and latent variables, a probability function describing the choice behaviour is given as:

$$p(y_i|\mathbf{h},\mathbf{x}) = \frac{e^{\sum_k B_{ik}x_k + \sum_j D_{ij}h_j + c_i}}{\sum_{i'} e^{\sum_k B_{i'k}x_k + \sum_j D_{i'j}h_j + c_{i'}}}. \tag{4.16}$$

Likewise, sampling of the hidden state is extended to incorporate $\mathbf{x}$:

$$p(h_j = 1|\mathbf{y}) = \sigma(d_j + \sum_i D_{ij}y_i + \sum_k A_{jk}x_k), \tag{4.17}$$

where the update parameters are $\theta = \{D_{ij}, B_{ik}, A_{jk}, d_j, c_i\}$. During the reconstruction phase, the condition probability (Eq. (4.16)) is equivalent to an 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 are illustrated in Fig. 4.3 and Fig. 4.4. In the positive phase, parameter vectors are adjusted decided by the learning rate σ to learn the transformed latent representation of the training set. In the negative phase, the latent variables are "clamped" or realized, and the parameter vectors are adjusted again by reconstructing the observed variables. From Fig. 4.2, the multinomial (MNL) model estimates the conditional parameter vector $\mathbf{B}$ and bias vector c, while the C-RBM model includes vectors $\mathbf{D}$, $\mathbf{A}$ and d.

4.4 Case Study: Consumer Choice Preference Dataset

In this section, we develop a product choice scenario with explanatory variables using the C-RBM model. The latent variables representing the latent attitudinal variables are simultaneously estimated in conjunction with the interaction with the choice model. First, we construct a structured choice subset from a financial product transaction dataset from the Kaggle database[1]. The data shows a monthly basis record of each financial product purchase by customers of Santander. The period of

[1] Dataset: https://www.kaggle.com/c/santander-product-recommendation/data

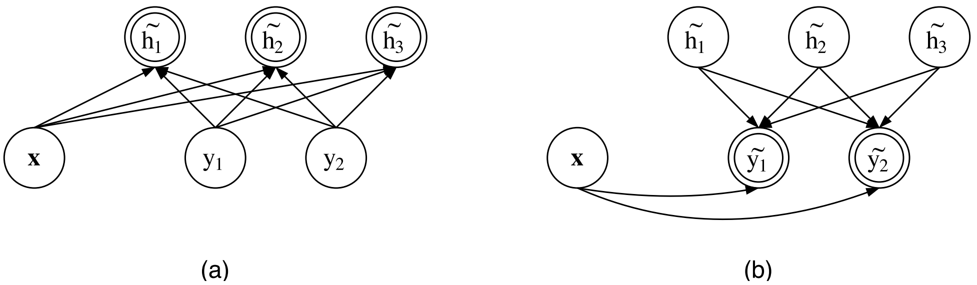


Figure 4.3: C-RBM (a) positive $\phi+$ and (b) negative $\phi-$ phases during semi-supervised discriminative training. Weights (connections) are learned to reduce reconstruction $\tilde{y}$ error.

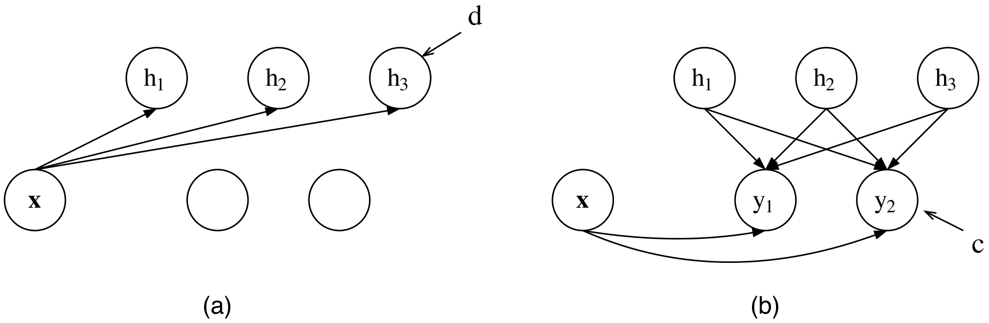


Figure 4.4: During the choice prediction phase, (a) latent variables are sampled using explanatory variables, and (b) the choice model is estimated with variables x and h.

the data ranges from January 2015 to June 2016. Next, we reduced the complexity of the dataset by removing transaction data that contain multiple product choices. To ensure consistency, inputs were scaled and normalized. Overall, the constructed dataset has a total of 13 alternatives (product choice) and 20 explanatory variables. Table 4.1 lists the alternatives and distribution across the dataset. Given the above conditions, a total of 253,803 valid responses were recorded, representing the total population sample with 13 available choices. A descriptive list of mean and standard deviation values of the explanatory variables are shown in Table 4.2. 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 maxi-

Table 4.1: List of choice alternatives ($\mathbf{y}$)

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

mum utility, making an inference about the behaviour of the decision-maker using the predictive model.

4.4.1 Method for assessing C-RBM model performance

We can estimate the weights of 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: ($N_{params} \in \mathbb{R}^{(I\times J)+(K\times I)+(K\times J)+K+I}$). With $J = 4$, we ended up with 409 parameters. To counteract overfitting due to this problem, we trained on 70% of our data and validated the model on the other 30% with a 2-fold bootstrap validation to verify generalization. When the validation error stops decreasing, the optimal estimation is reached [45]. A baseline comparison is set up using a standard multinomial logistic regression model with all explanatory variables and compared to the discriminative C-RBM modelling approach, followed by comparing the log-likelihood, ρ^2 model fit

Table 4.2: Explanatory variable descriptive statistics (**x**)

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

and predictive accuracy across all data models. The criteria for measuring the performance of a categorical based model include ρ^2 model fit and prediction error. The ρ^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 denote the accuracy of the model in predicting the correct choice, and the error is

$$\text{Error}_{valid} = 1 - \sum_{i} P(y_{pred} = 1 | x, h, y_i = 1). \tag{4.18}$$

y_i is the actual choice, and Error_{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.

4.5 Results

We compare the different models based on their generalization performance on the test set. A total of 76,141 observations were used in the test. For this study, we tested both normalized and non-normalized data and found that both data produce a similar result. Model estimation and validation were performed with Theano ML Python libraries[2]. Optimization parameters used were stochastic gradient descent (SGD) on mini-batches of 64 samples for 400 epochs with input normalization. We used an adaptive momentum-based learning rate with an initial rate of $1e^{-}3$ [104]. Training time was approximately 30 minutes for each model, including validation running on an Intel Core i5 workstation. At the given time, computational demand may not be significant to justify the small number of hidden units. However, speed could become a more critical consideration when model estimation and validation increase in data size or using huge parameter vectors with higher dimensionality. The statistical results of the model comparison across the same validation set are shown in Table 4.3. We found that additional latent information about the relationship between explanatory variables and observed decisions was useful and increases model accuracy. Bayesian Information Criterion (BIC) values indicate that 8 hidden units may be the optimal number of latent variables and higher BIC values above 8 hidden units might suggest overfitting. However, when generating semantic class meanings, a smaller number of latent variables may be simpler, therefore, in our example, we use only 2 latent variables for analysis.

To evaluate the efficiency of the models, we used a Hinton diagram [17] to analyze the parameter strengths between independent and dependent variables. We plot the parameter values and significance with the choice on the y-axis and independent

[2] Theano Python library: http://github.com/Theano/Theano

Table 4.3: Model training results

Model	latent variables	Validation error	log-likelihood	ρ^2	no. of params	BIC
MNL	$J = 0$	0.4454	-206808	0.546	273	416915
CRBM	$J = 2$	0.4360	-203558	0.553	341	411237
	$J = 4$	0.4338	-202066	0.556	409	409075
	$J = 8$	0.4323	-200846	0.559	545	408279
	$J = 16$	0.4318	-200223	0.560	817	410321

variables on the x-axis. A Hinton diagram is often used in the model analysis where the dimensionality of the model is high and provides a simple visual way of analyzing each vector. Figs. 4.5 to 4.9 shows the parameter estimates of the completed training stage of the different models. The Hinton matrix shows the influence of each independent variable on each alternative or latent variable. Statistically significant (>95% confidence bound) parameters are highlighted in blue. The values along the x-axis are normalized with zero mean and unit variance. The 13 financial product choices are listed on the y-axis. The estimated parameters and bias of the C-RBM prediction model **B**, **D** and c are projected onto the Hinton diagram (Fig. 4.6a, Fig. 4.7a, Fig. 4.8a and Fig. 4.9a) while parameters **A** and d representing the parameters and bias for the latent variable with respect to the alternatives shown in Fig. 4.6b, Fig. 4.7b, Fig. 4.8b and Fig. 4.9b. c and d are the constants with respect to the observed and hidden layer respectively. The signs and value of each parameter correspond to the size and colour of the patches in the matrix, with white and black representing positive and negative signs, respectively. Statistical significance (t-test) of each parameter is calculated using $\frac{\theta}{\sqrt{\sigma}}$, where σ is the inverse of the Hessian of the log-likelihood with sample size adjustment with respect to the parameters.

4.5.1 Analysis of latent variables

We can characterize each hidden unit with the explained significance and strengths represented by the weights $\mathbf{D}^\top$. $\mathbf{D}^\top$ is the parameter matrix that indicates the linear contribution of each latent variable and a constant d, such that each alternative can

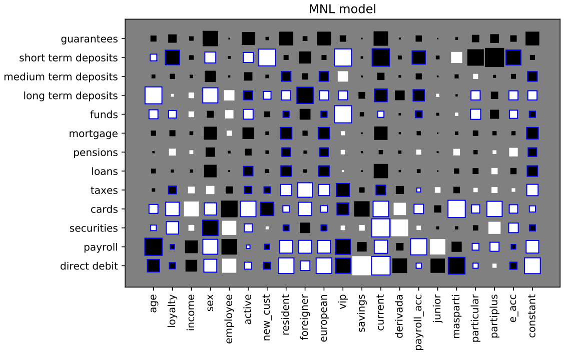
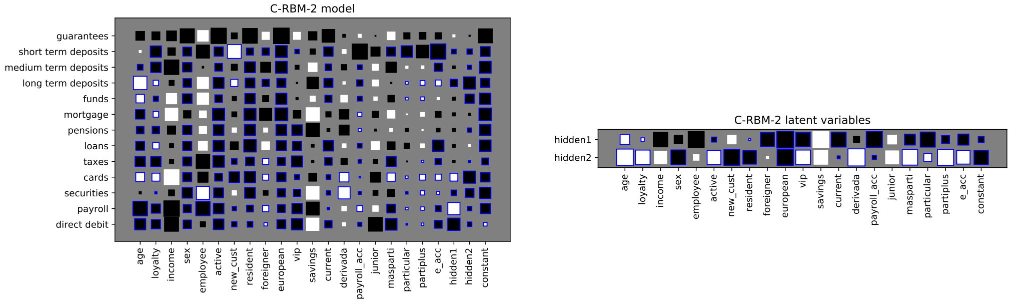


Figure 4.5: MNL model parameters. White: +ve values, Black: -ve values, Blue: >95% significant



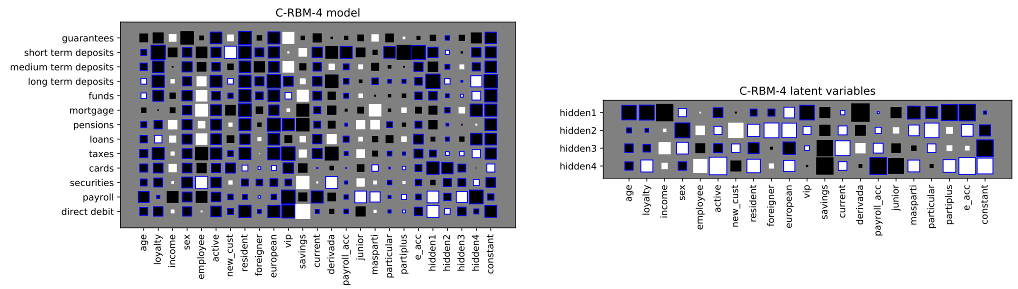
(a)

(b)

Figure 4.6: (a) C-RBM model with 2 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant

be described as a utility function of latent variables: $y = \mathbf{D}h + d$.

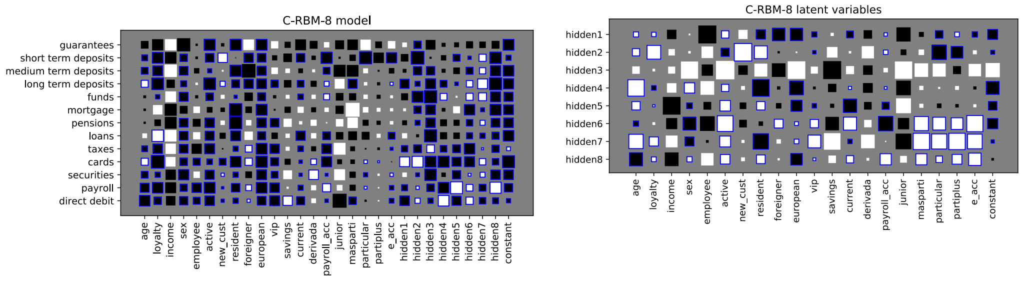
For example, C-RBM-2 latent variable *hidden1* is characterized by individuals who are of working age, non-EU foreign citizens with non-VIP status, and do not own any special accounts. We can infer this latent variable that indicates a 'savings driven attitude' (Fig. 4.6b). From the model results, the population with such characteristics has a favourable preference for purchasing a payroll product and a low motivation of purchasing a (credit/debit) card product as indicated in Fig. 4.6b. Likewise, in latent variable *hidden2*, it is represented by older, loyal customers who


C-RBM-4 model

C-RBM-4 latent variables

(a) (b)

Figure 4.7: (a) C-RBM model with 4 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant

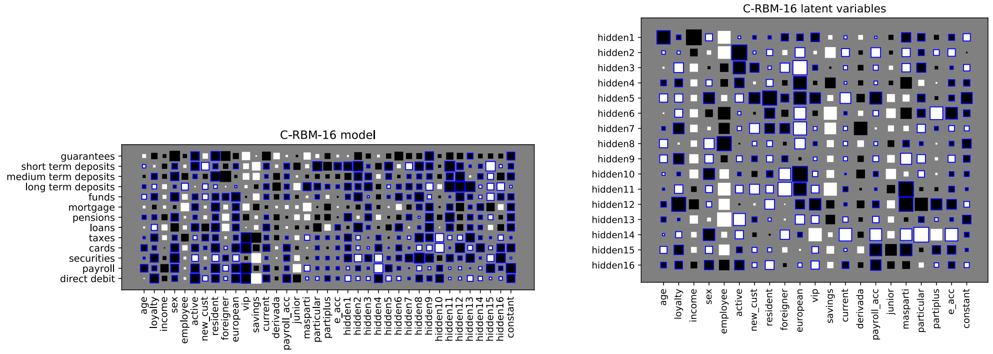

C-RBM-8 model

C-RBM-8 latent variables

(a) (b)

Figure 4.8: (a) C-RBM model with 8 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant

are VIP and have held various account types over their lifetime. This latent variable can be inferred as 'self-reliance attitude' and are an indication of the population who are less likely to purchase long term deposits, funds, securities and card products. The C-RBM 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



(a) (b)

Figure 4.9: (a) C-RBM model with 16 latent variables. (b) Latent variable relationship parameters. White: +ve values, Black: -ve values, Blue: >95% significant

would be sufficient for significant improvement over an MNL structure.

From the presented results, it is clear that the C-RBM models differ significantly from the MNL model in terms of parameters that are reliable and significant. The result seems to be broad-based in the sense that the number of hidden units does not dictate it, and it signifies that the observed distribution has some latent factors that can be further explored. However, we should mention that the training parameter initialization factor may have a small random effect on the model. Note that in the estimate plots, the signs and strength contribution to the choice model differ from model to model, which may indicate that model training may be stuck in some local optima. This also suggests that the hidden and observed layers have different scale [53]. What was suggested in [12] was to increase the learning rate to improve convergence, but that would result in overgeneralization and loss of expressive power in the hidden units. We posit a middle-of-the-road solution should have adequate model accuracy and generalization over a large population.

In an attempt to verify the above discussed C-RBM model, we constructed two latent classes using the obtained latent variables. First, a Latent Class model using BIOGEME was estimated with only the significant variables found in the C-RBM model. Then a (reduced) model was estimated with the significant variables found in the first LC estimate. However, in both cases, we could not identify any significant parameters other than alternative specific constants.

We performed a 2-fold bootstrap validation analysis and determined that the residual from the model fit is not significant. Therefore the model is robust to changes in input data – this is further confirmed with a sensitivity analysis presented in the following section. In the parameter plots, we can see 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 C-RBM 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).

4.5.2 Sensitivity of parameter estimates

The versatility and effectiveness of parameter estimates were determined by a sensitivity analysis of the model output. Methods of sensitivity analysis include variance-based estimator, sampling-based and differential analysis [105, 106]. "Sensitive" parameters are those whose uncertainty contributes substantially to the test results [105]. The model is sensitive to input parameters in the variability associated with the input variable resulting in a large output variability. Sensitivity ranking sorts the input parameters by the amount of influence it has on the model output and the disagreement between rankings measures the parameter sensitivity to changes to the input.

First, we define a list of parameters used in the model by their standard errors calculated over the full dataset. In a large dataset sensitivity analysis, a key concern was the computational cost needed to complete the analysis. Hence we used a sampling-based approach as a cheap estimator to the output percentage difference of the parameter minimum and maximum value. Random sampling (e.g.simple random sample, Monte Carlo, etc.) generates a distribution of inputs and outputs to assess the model under uncertainties [105]. Analyzing the sampling effects may provide information about the overall model performance since parameter sensitivity depends on all parameters in which the model is sensitive and therefore indicating the importance of each parameter [106].

Consider that the C-RBM model is represented by $y = f(x, h)$, where x and h are the input vectors of observed and latent variables respectively, and y is the model

Table 4.4: Parameter sensitivity rank and standard error difference for estimated parameters **B** for sampling-based sensitivity analysis

	RBM 2 LV			C-RBM 4 LV			C-RBM 8 LV			C-RBM 16 LV		
sample size parameter	n	n_s rank	std. err. diff.	n	n_s rank	std. err. diff.	n	n_s rank	std. err. diff.	n	n_s rank	std. err. diff.
βage	15	15	49.30	15	12	0.52	11	11	0.99	11	12	0.64
βloyalty	18	14	59.36	14	15	0.38	15	15	0.82	15	17	0.48
βincome	3	3	3712.99	3	3	26.67	3	3	43.00	3	2	35.82
βsex	12	13	67.51	13	14	0.41	14	13	0.91	14	15	0.52
βemployee	5	2	4267.79	2	4	13.74	5	5	21.27	4	4	33.74
βactive	21	16	47.92	16	19	0.20	19	19	0.34	19	19	0.26
βnew_cust	6	12	53.93	12	7	1.49	8	8	1.34	8	9	0.91
βresident	16	20	16.61	20	20	0.19	20	20	0.31	20	20	0.23
βforeigner	8	17	29.15	17	8	1.43	9	10	0.76	7	7	1.35
βeuropean	17	20	16.62	20	20	0.19	21	20	0.31	21	20	0.23
βvip	20	10	122.66	10	16	0.33	16	12	0.99	16	13	0.68
βsavings	1	1	34177.13	1	1	258.41	2	1	255.12	2	1	181.81
βcurrent	7	11	64.19	11	13	0.41	12	18	0.38	12	16	0.39
βderivada	4	4	3112.38	4	5	4.70	4	4	19.67	5	5	2.82
βpayroll_acc	9	18	24.91	18	18	0.29	18	17	0.52	18	18	0.41
βjunior	2	5	1759.26	5	2	58.29	1	2	45.32	1	3	22.43
βmasparti	11	7	185.94	7	9	1.41	7	6	4.99	9	6	2.29
βparticular	14	8	166.56	8	11	0.61	13	14	0.83	13	14	0.53
βpartiplus	10	6	189.75	6	10	0.65	10	9	1.51	10	10	0.86
βe_acc	19	9	159.38	9	17	0.33	17	16	0.82	17	11	0.91
bias	13	19	19.07	19	6	3.17	6	7	3.35	6	8	0.48

output. We suppose that the model $f(\cdot)$ is a complex, highly non-linear function such that we cannot wholly define the way the C-RBM model responds to changes in input variables. Also, h is dependent on x through a submodel previously shown in Fig. 4.2. Our analysis involves independently and randomly generated sample with size $n_S = 0.1n$ (10% random sample draw), where $n = 76,141$ is the total number of observations. The model performance was considered by sampling stability of the variable parameters. Sensitivities were also assessed for the size of hidden units used in generating the C-RBM models and indicated the number of latent variables (hyperparameter) is required for model identifiability. Since the model was applied using a multinomial logit approach instead of a conditional logit, this resulted in a vast number of parameters. Thus the effect of relative changes to the number of distributed parameters gave the range of variance across each explanatory variable and number of hidden units used. Table 4.4 shows the effects of sampling on the sensitivity and stability of the model observed parameters on the theoretical values and size of latent variables. Notice that the relative difference in standard error between the full and sampled model decreases when the number of latent variables increases. This reveals that the C-RBM models with a high number of synthetic latent variables are robust to changes to input values through sampling. Additionally, the parameter sensitivity rank across variables also becomes more consistent. Hence, the results show that RBM models are efficient in obtaining suitable latent variables with low generalization error. The significant decrease in standard error difference from 2 LV to 4 LV may indicate that the number of latent variables used in the models has a lower bound on the generalization error, which implies that we need careful consideration on h for obtaining efficient but yet accurate exact values of β without losing model interpretability.

4.6 Conclusion

This paper analyzes an alternative method of latent behaviour modelling in the absence of attitudinal indicators. In ICLV models, specialized surveys have to be constructed with attitudinal questions to model latent effects on the decisions. While it has been one of the more popular methods in discrete choice analysis, there are several disadvantages to it. First, attitudinal questions are subjective, and the behaviours 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. Lastly, it can be challenging to collect psychometric indicators from thousands of respondents efficiently. We explore generative modelling of the choice distribution to uncover latent variables using machine learning methods, without measurement indicators. We hypothesized that latent effects could be obtained not only from attitudinal questions but also from the posterior choice distribution. In effect, we are modelling latent components that fit the real choice distribution rather than achieving good statistics on subjective models. For example, there could 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, difficultly 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 current 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, but can be brought in. The underlying RBM is capable of dynamics. 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 [107, 103]. Finally, it is worth noting that as the number of latent variables 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 the efficient use of model parameters. In our observation, performing bootstrap cross-validation and model selection with the lowest validation error was a proper 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 [94].

While the ICLV model is optimized to predict the effects of latent constructs on decision-making behaviour using measurement indicators to guide latent parameters selection, our method uses observed decisions as an information 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 misspecification problems and when explanatory variables are poor predictors, ICLV models can improve latent effects on choice models [35]. Instead, we show that latent effects may not only be present in attitudes and perceptions, but also the direct observation of choices. Our current work explores the use of posterior choice distribution for latent behaviour modelling. Generative modelling in DCA is inspired by state-of-the-art machine learning algorithms that perform unsupervised feature extraction from unlabelled data used in classification problems [59]. In circumstances when attitudinal variables are not available, we have a strong reason to believe that the generation of latent factors is essential and useful in building a discrete choice model.

A future study that would be of interest is to extend this method to datasets with attitudinal questions and surveys. For example, inter-city rail survey [76], and perform an analysis on both RBM and ICLV methods to obtain the generalization error of attitudinal survey models. A comparative study would provide a foundation for analysis of various latent behaviour models through graphical and algorithmic methods and provide guidance not only in selecting the appropriate latent variables, but also direct research effect to more promising directions.

Chapter 5

Information Processing Constraints in Travel Behaviour Modelling: A Generative Learning Approach

Preamble

This chapter provides a theoretical background understanding and interpretation of a generative model with respect to behaviour analysis and choice process. We frame the process of decision making as a learning algorithm that makes use of concepts related to thermodynamics (energy), Prospect Theory (uncertainty) and economic theory (utility). Instead of looking at the model as a “black box”, we explain how behaviour patterns emerge through a stochastic learning process. The key idea in this study is describing the implication of information processing cost and how entropy plays a role in building a model of choice behaviour.
This research article is under first review in *Transportation Research Part B: Methodological.*

Abstract

Travel decisions tend to exhibit sensitivity to uncertainty and information processing constraints. These behavioural conditions can be characterized by a generative learning process. We propose a data-driven generative model version of rational inattention theory to emulate these behavioural representations. We outline the methodology of the generative model and the associated learning process as well as provide an intuitive explanation of how this process captures the value of prior information in the choice utility specification. We demonstrate the effects of information heterogeneity on a travel choice, analyze the econometric interpretation, and explore the properties of our generative model. Our findings indicate a strong correlation with rational inattention behaviour theory, which suggest correlation to prior information in decision making under uncertainty. The principles demonstrated in this study can be formulated as a generalized entropy and utility based multinomial logit model.

5.1 Introduction

The classical assumption about modelling travel behaviour data is that individuals have varying unobserved heterogeneity in their choice preferences [68]. In recent years, the use of data-driven modelling and integration of behavioural and psychological factors in discrete choice and travel behaviour analysis have become active areas of research [37, 108, 109]. In the context of data-driven models, behavioural variations describe the correlation between observed choice attributes and unobserved socio-economic factors using a flexible and tractable model specification. These variations include: *decision-protocols*, *choice sets*, *unobserved taste variations* and *unobserved attributes* [26]. Under these considerations, recent studies on travel behaviour analysis have so far primarily focused on representing heterogeneity in the error correction function and incorporating it into utility based multinomial logit (MNL) models [108]. Models such as mixed multinomial logit (MMNL) or latent class (LC) model offers flexibility in representing heterogeneity and substitution patterns. Also, recent conceptual frameworks such as the integrated choice and latent variable (ICLV) use individuals' psychometric indicators to represent unobserved behavioural and perception heterogeneity [110]. It is also possible to apply a generative machine learning to identify informative latent constructs in travel decision making without subjective behaviour indicators (see Chapters 3 and 4). However, the true underlying behavioural patterns are often unknown and usually approximated by some pre-determined exogenous indicator variables that would often lead to model misspecification due to lack of complete information, or error in data collection [111]. Furthermore, accurate specification of the underlying distribution assumes individuals have access to all available information regarding the travel activity (e.g. travel times of each mode, knowledge of exact traffic status, etc.). This information will not always be available to the individual, and they might also choose not to consider these variables in their decision-making process. Therefore, statistical variations in the observed data may not exhibit the same underlying properties as with the individuals' behaviour.

A different perspective to explain these heterogeneity manifestations is to consider the element of information processing costs based on rational inattention theory [58, 30]. Rational inattention theory is defined as individuals choosing their optimal preference, at the same time considering incomplete information about the choice attributes and relying on their prior beliefs about the choice set. A typical example would be route choice selection: Individuals tend to ignore most path choices and consider only a few prioritized routes in their choice set [112]. These manifestations occur through repeated choice process and prior experiences about the travel routes. As described in [30], information-theoretic approaches do not impose any particular assumptions on what is learned or how they are learned—the structure of the model is estimated through the minimization of decision uncertainty. Under this interpretation, a rational inattention model captures the systematic utility and adjusts for prior knowledge and individuals' internal information processing strategy using an entropy term. Individuals perceive route choices with heterogeneous prior beliefs and allocate different levels of attention to each alternative. Consequently, misspecification in classical econometric model estimation can be interpreted as the systematic error between the data observed by the analyst and the true underlying heterogeneous beliefs of the decision-makers (which are hidden to the analyst).

The objective of this research is to model unobserved variations in travel behaviour data by emulating decisions under uncertainty and information processing constraints as a data-driven generative learning process. We develop a choice model estimation framework with latent constructs that capture information heterogeneity within the data. The key difference between our work and previous literature is that we show how rational inattention can be framed as a flexible and extendable generative learning model that emulates the cognitive processes in human behaviour [113, 25]. We postulate that realistic behavioural patterns can be modelled using a data-driven generative learning process, and we estimate a model to represent the underlying heterogeneity of the data. Lastly, we provide a quantifiable economic interpretation using latent variables by analyzing the model properties and systematic effects from the latent variable parameters. This will provide valuable insights into how modern data-driven and deep learning techniques can be exploited to improve travel behaviour modelling.

The main contributions of this paper are summarized as follows:

- A novel framework for capturing and extracting properties of information heterogeneity in travel behaviour models (Fig. 5.1).

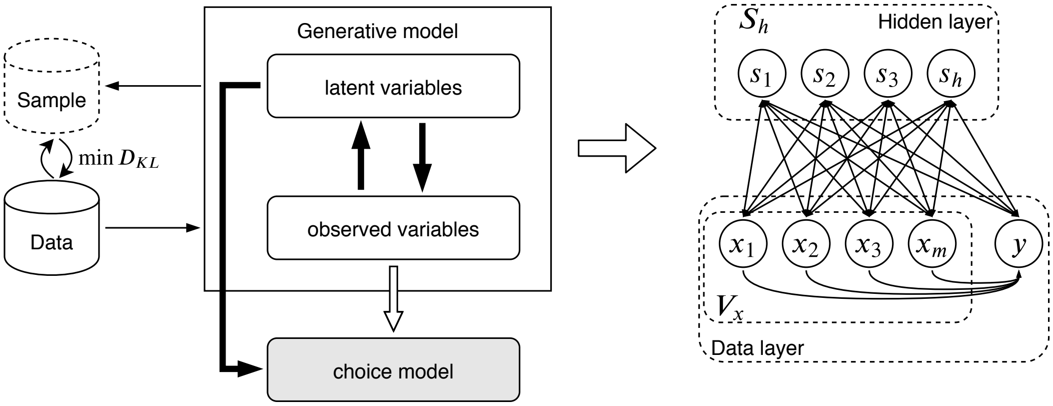


Figure 5.1: Framework for generative modelling.

- We show that generative modelling can be framed as an abstraction of rational inattention theory. Specifically, the learning and optimization process of a generative model emulates the internal information processing constraints of decision making.

- Demonstration of a data-driven modelling approach that exploits start-of-the-art deep learning techniques. A generative model architecture is described in the methodology.

- Discussion on the interpretation of generative learning on discrete choice analysis.

- We provide new insights into the sensitivity analysis of econometric parameters through a travel behaviour case study.

The remainder of the paper is organized as follows: Section 5.2 introduces preliminary concepts related to information theory in choice modelling and discusses existing literature on rational inattention behaviour theory. Section 5.3 describes the generative model framework and estimation methodology. In Section 5.4, a case study example on a trip-based travel behaviour analysis is shown, and we demonstrate how the results explain information heterogeneity in the data. Section 5.5 provides a brief discussion on the results, conclusion and suggestions for future research.

5.2 Information Theory in Behaviour Models

In this section, we introduce several preliminary concepts that relate to our work by beginning with the connection between rational inattention behaviour and information theory in the context of generative modelling.

5.2.1 Rational inattention behaviour

Rational inattention presents a behavioural scenario where individuals' choice influences are based on Shannon's mutual information that measures uncertainties between an exploitative and exploratory choice process. Specifically, it frames the choice problem on observations as well as information processing constraints similar to that of a commutation channel with finite Shannon capacity [18]. By representing information processing constraints, it accounts for the natural deviations in econometric behaviour [18, 58]. This concept stems from the same principles of neuroscience, where behaviour learning and perceptual inference can be explained through information theory and statistical physics [60]. Using modern deep learning techniques, one can construct a rational inattentive learning model using an artificial neural network to provide a principled way of analyzing travel behaviour patterns from large scale datasets.

As a simple generalization, information processing constraints across choice preferences can be represented by an unknown distribution of random utility shocks according to Ellsberg's paradox which showed that individuals systematically violate utility theory by being averse to ambiguity [114]. Consider a case where an individual is faced with two options in a choice set when the expected utilities are identical for both options. In utility theory, both options will be chosen at equal probabilities, whereas in rational inattention, the individual chooses the option that maximizes entropy (attention). This decomposition accounts for the prediction error under different protocols as well as it resembles exploratory choice behaviour (i.e. prospect theory) [20]. For instance, when the differences in utility between two travel modes do not differ, travellers would try new options, with increased risk.

Existing studies on rational inattention in choice modelling research stems from the findings that this behaviour can be generalized in an MNL model [30]. However, they have mostly focused on static models, as dynamic rational inattention models are challenging to solve and may be intractable using conventional methods [115]. The value of adding information processing constraints have suggested well-defined similarities with macroeconomic behaviour theory [18]. Recently, rational inattention has become a particularly appealing approach to modelling choice behaviour. For instance, [30] described the implication of information availability on consumer choice selection behaviour using a rational inattention model. In a combined location and mode choice model, [116] used a method of entropy maximization in a non-linear mixed-integer program subject to available information constraints. [117] investigated consumer inattention correlation with willingness-to-pay for fuel consumption. Recently, rational inattention has been found to work well in time variability problems in travel demand forecasting [25]. The theory of rational inattention seeks to endogenize the imperfect awareness about the circumstances [58]. The decision-maker selects pieces of information that are most relevant for his or her utility and ignores the rest, so long as the information cost can be accounted for in the model.

5.2.2 Information theory

In this section, we explore some key properties of information theory in the context of behavioural modelling. Information theory has been used to provide insights into the non-rational behavioural choice, and it was shown to be equivalent to random utility maximization MNL model [62]. An information-theoretic model can also be used as a tool for generating new predictions beyond MNL restrictions, subject to available information [62]. Recent studies have also shown that this is also functionally equivalent to an additive random utility maximization problem in rational inattention behaviour models and several well-known decision problems can be reasonably represented, e.g. Prospect Theory and Regret Theory [20, 30]. The measure of information heterogeneity is closely related to non-normative representation, involving Shannon entropy [113]. Expected utility representation may not be sufficient in providing the proper specification for these decision problems as individuals may perceive choice probabilities with different levels of uncertainty. Decisions under uncertainty can be interpreted simply by correcting for information processing constraints in the utility specification.

Energy

Assuming a bi-directional system with an observed and an unobserved (latent) states, the level of uncertainty of a state configuration of the system with observed X and latent S random variable is a function of energy $E(x,s)$ of the state proportional to the joint probability $p(X=x,S=s)$ or $p(x,s)$:

$$p(X=x,S=s)=\frac{1}{Z}e^{-E(x,s)}, \tag{5.1}$$

where $Z=\sum_{x,s}e^{-E(x,s)}$ is the normalization function so that $\sum_{x,s}p(x,s)=1$. Due to the logarithmic function, energy decreases monotonically as the probability increases. Imposing monotonicity allows model estimates to be more interpretable and tractable. An event with high energy will have a lower probability of occurrence (individuals will tend to avoid this state). An event with low energy will always be within the expectation of the individual, thus having higher probability [118].

Mutual information

Mutual information allows us to identify general non-linear dependencies by measuring the amount of information processed by the individual, i.e. how far two random variables are from being independent. Given two random variables X and S, let $(X,S)\sim p(x,s)$, the mutual information $I(X,S)$ can be written in the form:

$$I(X,S)=\sum_{x,s}p(x,s)\log\frac{p(x,s)}{p(x)p(s)}=\sum_{x,s}p(x,s)\log\frac{p(x|s)}{p(x)}=H(X)-H(X|S) \tag{5.2}$$

It can be interpreted as the decrease in uncertainty of X given S, where H(X) and H(X|S) are the entropy and the conditional entropy, respectively:

$$H(X)=-\sum_{s}\sum_{x}p(x,s)\log p(x)=-\sum_{s}p(s|x)\sum_{x}p(x)\log p(x)=-\sum_{x}p(x)\log p(x) \tag{5.3}$$

Mutual information is symmetric $I(X,S)=I(S,X)$ and it is non-negative, $I(X,S)\geq 0$ and it is zero if and only if X and S are *independent* (with respect to the model identification process). Hence, the mutual information shown in Eq. (5.2) is equivalent to finding the expected energy difference between the data generating distribution and the exact distribution obtained from the data.

Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence or the relative entropy measures the 'distance' between two distribution, p and q [118]. The KL divergence of q from p is $D_{KL}(q||p)$, when $q = p$ then $D_{KL} = 0$. The mutual information, using the example above, can be interpreted as the divergence of the joint distribution from the product of marginals:

$$I(X,S) = D_{KL}(p(x,s)||p(x)p(s)) \geq 0 \tag{5.4}$$

Thus in practice, we can consider the hypothesis $H_0 : D_{KL} = 0$ against $H_1 : D_{KL} \neq 0$ as a test for *independence* between two random variables [118]. To put it in a different perspective, if we can define a framework where the latent variables interact with the observed variables by a correlation matrix, then the mean and variance of the matrices indicate how much information heterogeneity is present in the data representing the population.

5.3 Methodology

We propose a generative model framework that extends rational inattentive behaviour in discrete choice, interpreting it as an *optimization process* rather than a structural model specification. We differentiate our work from the generalized entropy function described in [113] by framing non-normative behaviour as a learning model – allowing for random perturbations to be data-driven. Under this framework, the estimation of a generative model assumes to emulate information processing constraints in rational inattention behaviour and identifies observed and latent variable interactions through a neural network interface. The estimated latent variable parameters reflect the correlation between random decision and information priors. We use a Restricted Boltzmann Machine (RBM) learning algorithm as an example to estimate the generative model parameters. Other forms of generative model algorithms (e.g. Autoencoders, GANs[1], DBNs[2] [45]) can similarly be used. Another more straightforward form of generative modelling is principal component analysis (PCA). However, PCA has severe limitations as it cannot handle complex non-linear relations in the data [97]. We focus on the RBM learning algorithm as we would show that it is an approximation to a rational inattention information processing with similarities to an error components model. The error components control for the heterogeneity in the observed utility and variances in the unobserved utility, where an entropy function represents the unobserved utility.

5.3.1 Proposed generative model framework

The generative model framework is a tri-partite RBM with a data layer $\mathcal{D}$ representing the set of observed variables $\mathcal{D} = \{x_1, x_2, ...x_m, y\}$ including a dependent variable y and a hidden layer $\mathcal{S}$ representing the set latent variables $\mathcal{S} = \{s_1, s_2, ..., s_h\}$ (see Fig. 5.1). The generative model can be framed as a fully connected tri-partite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{D}, \mathcal{S} \in \mathcal{V}$ is the set of graph nodes and $\mathcal{E}$ are the graph edges. The nodes from $\mathcal{V}_x = \{x_1, x_2, ..., x_M\}$ are connected to $\mathcal{V}_y = \{y\}$ by edge subset $\mathcal{E}_{xy}$, representing the choice model explanatory variable coefficients. The edges between $\mathcal{S}$ and $\mathcal{D}$ are the correlation matrix between the latent and observed variables. The edge subset $\mathcal{E}_{hy}$ represents the decision level heterogeneity. The algorithm focuses on generating synthetic data using a blocked Gibbs sampling protocol, alternating between observed and latent variable samples from the joint distribution conditioned on the previous step. A non-zero valued covariance matrix represents the level of information heterogeneity captured in the data. A zero covariance matrix indicates that the observed explanatory variables captures all the taste variations and assumes a sufficiently homogeneous population. The observed data can be inferred by sampling from the generative model probability distribution. By minimizing the KL divergence between the observed and generated data, we learn the parameters of the correlation matrix between the observed and latent variables. When the generated data have matched the observations, the underlying priors are assumed to have encoded the information heterogeneity of the population and can be represented in the

[1] GAN: Generative Adversarial Networks

[2] DBN: Deep Belief Nets

choice model.

5.3.2 Model specification

The RBM architecture was designed as an efficient feature descriptor that progressively trains a fully connected non-linear model structure [104]. The interactions between the two parallel components capture the information about the heterogeneity present between hypotheses. Each latent variable represents a specific state encoded as distributed binary patterns.[3] The different combinations of latent variables form the intricate behavioural activity patterns and are inferred through sampling from the posterior. Similar to a random utility specification, we start with a scalar energy value describing the joint configuration of observed explanatory variables, dependent choice variable and latent variables:

$$E(\mathbf{x}, \mathbf{s}, y) = -\mathbf{x}\boldsymbol{\beta}y - \mathbf{x}\mathbf{W}\mathbf{s} - \mathbf{s}\mathbf{W}'y - \mathbf{d}\mathbf{x} - \mathbf{c}y - \boldsymbol{\alpha}\mathbf{s} \tag{5.5}$$

The energy function is parameterized by a set of coefficients $\phi = \{\boldsymbol{\beta}, \mathbf{d}, \mathbf{c}, \boldsymbol{\alpha}, \mathbf{W}, \mathbf{W}'\}$, where $\boldsymbol{\beta}$ are the choice model coefficients and $\mathbf{d}, \mathbf{c}, \boldsymbol{\alpha}$ are the constants of the observed explanatory, dependent and latent variables respectively. $\mathbf{W}$ and $\mathbf{W}'$ are the parameters matrices representing the information heterogeneity captured by the latent variables given the observed explanatory and dependent variables. $\mathbf{y}$ is a discrete dependent variable representing the choice alternatives, e.g. $y = \{1, 0, 0\}^\top, \{0, 1, 0\}^\top$ or $\{0, 0, 1\}^\top$ representing a selected alternative. $\mathbf{x}$ is a vector of observed explanatory variables either as discrete or continuous values. Multiple discrete and continuous dependent values can also be used as the output (see Chapter 6). $\mathbf{s}$ is a vector of stochastic binary variables. Given that the non-latent variable terms can be factorized out, the posterior over the latent variables is as follows:

[3] Distributed binary patterns are commonly used in digital signal encoding. For example of a pattern: $\mathbf{s} = \{0, 1, 0, 0\}$ or $\{1, 1, 0, 1\}$. We make the analogy to digital encoding to refer to choice behaviour perceptions. A latent variable model with N elements can represent up to $2^N - 1$ different behaviour perceptions. The Boltzmann architecture uses this representation with a stochastic sampling algorithm to learn the model parameters. Other forms such as multinomial discrete vectors or multivariate normal can also be used as possible encoding patterns, but binary encodings are the most straightforward method to simplify model inference.

$$p(\mathbf{s}|\mathbf{x},y) \propto \prod_h p(s_h|\mathbf{x},y) = \prod_h \exp(\mathbf{x}\mathbf{W}_h s_h + s_h \mathbf{W}'_h y + \alpha_h s_h) \tag{5.6}$$

Using the aforementioned energy function $E(\mathbf{x},\mathbf{s},y)$ allows the conditional to be factorized. Defining the normalizing constant as the sum of the binary configurations, we obtain the *normalized* probability density function for each latent variable s_h:

$$\begin{aligned} p(s_h = 1|\mathbf{x},y) &= \frac{\tilde{p}(s_h = 1|\mathbf{x})}{\tilde{p}(s_h = 0|\mathbf{x}) + \tilde{p}(s_h = 1|\mathbf{x})} && (5.7) \\ &= \frac{1}{1 + \exp(-((\mathbf{x}\mathbf{W})_h + \mathbf{W}'_h y + \alpha_h))} && (5.8) \end{aligned}$$

The objective is to optimize the model parameters such that a sample $\tilde{\mathcal{D}} = \{\tilde{x}_2, \tilde{x}_2, ..., \tilde{x}_m, \tilde{y}\}$ is generated with a distribution as close to the data distribution $\mathcal{D}$. Computing the energy over the data layer $E(\mathcal{D})$ corresponds to the expected energy of the model minus the entropy:

$$E(\mathcal{D}) = \sum_s p(s|\mathcal{D}) E(\mathbf{x},\mathbf{s},y) - H(S) \tag{5.9}$$

which can be simplified into the form:

$$\begin{aligned} E(\mathcal{D}) &= -\log \sum_s e^{-E(\mathbf{x},\mathbf{s},y)} && (5.10) \\ &= -\mathbf{x}\boldsymbol{\beta}y - \mathbf{d}\mathbf{x} - \mathbf{c}y - \log \left(\sum_s \Big(\exp(\mathbf{x}\mathbf{W}\mathbf{s} + \mathbf{s}\mathbf{V}y + \boldsymbol{\alpha}\mathbf{s}) \Big) \right) && (5.11) \\ &= -\mathbf{x}\boldsymbol{\beta}y - \mathbf{d}\mathbf{x} - \mathbf{c}y - \log \left(\prod_h \Big(\sum_{s_h \in \{0,1\}} (\exp((\mathbf{x}\mathbf{W})_h s_h + s_h \mathbf{W}'_h y + \alpha_h s_h)) \Big) \right) && (5.12) \\ &= -(\boldsymbol{\beta}y + \mathbf{d})\mathbf{x} - \mathbf{c}y - \sum_h \log \Big(1 + \exp((\mathbf{x}\mathbf{W})_h + \mathbf{W}'_h y + \boldsymbol{\alpha}) \Big) && (5.13) \end{aligned}$$

Eq. (5.13) is a direct interpretation of the generalized entropy formulation for discrete choice [62, 113]. The coefficients $(\boldsymbol{\beta}y + \mathbf{d})$ stand for the unknown parameters of the explanatory variables for each alternative y and for the generative model respectively. Increasing $\boldsymbol{\beta}$ decreases the energy over the data generating distribution conditioned on a choice alternative, while increasing $\mathbf{d}$ decreasing the energy over all data generating configurations. $\mathbf{c}$ represents the alternative specific constants and $\sum_h \log(1+\exp(\mathbf{x}\mathbf{W}_h + \mathbf{W}'_h y + \boldsymbol{\alpha}))$ is the flexible error component generator given a specific input configuration of observed $\mathbf{x}$ and y with a constant $\boldsymbol{\alpha}$. If this term is near zero, The expected energy function is equivalent to a utility function in a random utility-maximizing (RUM) model. By definition, the probability of $\tilde{\mathcal{D}}$ being generated is the Boltzmann distribution with energy $E(\mathcal{D})$:

$$p(\tilde{\mathcal{D}}) = \frac{1}{Z'} e^{-E(\mathcal{D})} \tag{5.14}$$

The computation of the marginal $Z' = \sum_{\mathcal{D}'} e^{-E(\mathcal{D}')}$, which sums over an exponential number of possible configurations of the data vector, becomes difficult as we increase the number of explanatory variables.

5.3.3 Objective function formulation

Our proposed framework addresses the estimation problem for a highly non-linear and non-closed form function using variational inference. We select from a family of distributions that produce an *approximate* posterior distribution. The specification of the posterior distributions is obtained from data accumulation during the learning phase. If we restrict the family of distributions that are tractable and can be factorized over each variable in Z, the problem of simulation-based estimation becomes significantly simpler. For the sake of clarity, we omit the parameter terms ϕ in the equations below. First, we consider $p(\tilde{\mathcal{D}})$ in terms of energy and the joint probability as follows:

$$p(\tilde{\mathcal{D}}) = \sum_s p(\mathcal{D}, \mathbf{s}) = \frac{e^{-E(\mathcal{D})}}{\sum_{\mathcal{D}'} e^{-E(\mathcal{D}')}} \tag{5.15}$$

We can map the energy of the observed part as a function of the total system energy in a formulation similar to Eq. (5.1) by defining $E(\mathcal{D}) = -\log \sum_s e^{-E(\mathcal{D},\mathbf{s})}$. The posterior over the latent variables as a function of energy using Bayes rule,

$p(a|b) = p(a,b)/p(b)$ results in a Boltzmann probability function over the joint distribution, which reveals the similarities to an MNL model:

$$p(s|\mathcal{D}) = \frac{p(\mathcal{D}, s)}{\sum_{s'} p(\mathcal{D}, s')} = \frac{e^{-E(\mathcal{D},\mathbf{s})}}{\sum_{\mathbf{s}'} e^{-E(\mathcal{D},\mathbf{s}')}} \tag{5.16}$$

If we take the expected values with respect to the posterior on (Eq. 5.15), the uncertainty of choice can be expressed in terms of expected energy and entropy denoted as the evidence lower bound $\mathcal{L}$:

$$\mathcal{L} = \underbrace{-\underbrace{\left[\sum_s p(s|\mathcal{D})\right]}_{=1} \log p(\tilde{\mathcal{D}})}_{\text{uncertainty}} = \underbrace{\sum_s p(s|\mathcal{D}) E(\mathcal{D})}_{\text{expected energy}} - \underbrace{\left(-\sum_s p(s|\mathcal{D}) \log p(s|\mathcal{D})\right)}_{\text{entropy gain}} \tag{5.17}$$

In Eq. (5.17), a rational inattentive based choice can be framed as the information difference between the expected energy and the entropy gain. The first term on the right of Eq. (5.17) denotes the individuals' behaviour towards prior expectations about the choice. The second term is the entropy, and it can be viewed as the information processing constraints in a rational inattentive model or a penalty for low energies. It ensures that the generative model produces low uncertainty values for inputs with high probability in the actual data distribution and high uncertainties for all other inputs [119]. Minimizing uncertainty implies both utility-maximizing and entropy seeking behaviour. Computing the evidence $\log p(\tilde{\mathcal{D}})$ is intractable, but we can use the posterior $p(s|\mathcal{D})$ to evaluate the marginal log likelihood [120].

In many cases, computing the posterior $p(s|\mathcal{D})$ may be difficult when the distribution is complex, as we require an integral over all configurations of latent variables to find the marginal or denominator in Eq. (5.16). The primary motivation of defining the problem as *variational inference* is that we can approximate the posterior distribution using a tractable arbitrary distribution $q(s)$ [67]. In the estimation procedure, we find the parameters that make q as close as possible to the posterior by minimizing $\mathcal{L}$ where q is the approximating distribution, then we have:

$$-\log p(\tilde{\mathcal{D}}) = \mathbb{E}_{q(s)}\left[E(\mathcal{D}) - (-\log p(s|\mathcal{D}))\right] \tag{5.18}$$

To show that the proposed distribution $q(s)$ can be used to approximate $p(s|\mathcal{D})$, we compute the marginal loglikelihood over $q(s)$ to minimize the KL divergence of $q(s)$ from $p(s|\mathcal{D})$:

$$
\begin{aligned}
-\left[\sum_s q(s)\right] \log p(\tilde{\mathcal{D}}) &= \mathbb{E}_{q(s)}\left[E(\mathcal{D})\right] - \left(-\mathbb{E}_{q(s)}\left[\log p(s|\mathcal{D})\right]\right) && (5.19)\\
&= \sum_s q(s)E(\mathcal{D}) + \sum_s q(s)\left(\log p(s|\mathcal{D}) + \log \frac{q(s)}{q(s)}\right) && (5.20)\\
&= \sum_s q(s)E(\mathcal{D}) + \sum_s q(s)\log q(s) - \sum_s q(s) \log \frac{q(s)}{p(s|\mathcal{D})} && (5.21)\\
&= \underbrace{\sum_s q(s)E(\mathcal{D}) - H_q(S)}_{\text{variational free energy } F_q(\mathcal{D})} - D_{KL}(q(s)||p(s|\mathcal{D})) && (5.22)
\end{aligned}
$$

Using the fact that the KL divergence cannot be negative, we get the lower bound on the model evidence, and we define the *variational* free energy $F_q(\mathcal{D})$ as:

$$F_q(\mathcal{D}) = \mathcal{L} + D_{KL}(q(s)||p(s|\mathcal{D})) \geq \mathcal{L} \tag{5.23}$$

The intuition from Eq. (5.23) is that minimizing the variational energy has the same outcome as minimizing $D_{KL}(q(s)||p(s|\mathcal{D}))$. The bound is exact if $D_{KL}(q(s)||p(s|\mathcal{D}))$ term is zero, which would happen if $q(s)$ matches $p(s|x)$ perfectly. Therefore, following the gradient of $F_q(\mathcal{D})$ yields the optimal solution for $q(s)$. Another equivalent form of variational free energy can be derived by transforming the marginal into the conditional likelihood:

$$F_q(\mathcal{D}) = -\log p(\mathcal{D}|s) + D_{KL}(q(s)||p(s)) \tag{5.24}$$

In Eq. (5.24), the objective function can be optimized through assigning specific priors over the generative model, then measuring how well the priors represent the observations. More generally, minimizing $F_q(\mathcal{D})$ together with the KL divergence is a good substitute for minimizing the log-likelihood function [119]. The first and second terms on the right-hand side are known as the fit and complexity respectively in Bayesian statistics. The first term defines the accuracy of the data generating model. If we presume that $p(s)$ is a complex model (real-world representation, intricate correlation between behaviour and choices, etc.), then the complexity tells us how much capacity is required for the (non-trivial) approximator $q(s)$ to match the empirical distribution. The variational energy can be used to determine the strength of non-linear interactions between components in a model. The minimization of variational energy provides consistent and reproducible models, equivalent to maximum likelihood estimation. We can establish the choice model by interpreting the data generating probabilities of a given data vector as the individuals' information heterogeneity by minimizing the variational lower bound. The objective cost function now becomes selecting the model parameters such that:

$$\theta^* = \arg\min_{\theta}\{D_{KL}(q(s)||p(s|\mathcal{D}))\} \tag{5.25}$$

In the proposed generative model, we are interested in evaluating large numbers of non-linear latent variables which belongs to a family of extreme valued distributions parameterized by latent variable parameters $\theta = \{\mathbf{d}, \boldsymbol{\alpha}, \mathbf{W}, \mathbf{W}'\}$. The primary assumption is that the approximating distribution $q(s)$ can be factorized, such that it gives a tractable form:

$$q(s) = \prod_h q(s_h; \theta) \approx \prod_h p(s_h|\mathcal{D}) \tag{5.26}$$

This form allows the generative model to produce distributions with sharper boundaries over conventional mixture models. Using this specification, the model variance can be increased or decreased with the number of activated latent variables.

5.3.4 Parameter estimation

We formalize the model learning as minimizing KL divergence given some observed data $\{\mathcal{D}^n\}_1^\infty$. The important advantage of this is that we can incorporate the differences between an individual's actual behaviour and mean population behaviour effectively in the objective function. The parameter update rule for a generative model is obtained by implementing a stochastic gradient descent on the variational free energy function, updating the weights of the coefficients between latent and observed variables according to the sampling states. Consequently, the gradients with respect to the parameters are as follows:

$$\mathbb{E}_q\left[\frac{\partial}{\partial\theta}\log p(\tilde{\mathcal{D}})\right] = \sum_{n=1}^{\infty}\frac{\partial D_{KL}(q(s)||p(s|\mathcal{D}^n))}{\partial\theta} \approx \frac{\partial}{\partial\theta}E(\mathcal{D}^{(1)}, s) - \mathbb{E}[\frac{\partial}{\partial\theta}E(\mathcal{D}^n, s)], \tag{5.27}$$

where the expectation is over $\tilde{\mathcal{D}} \sim p(\tilde{\mathcal{D}})$. The learning algorithm is based on a Gibbs chain starting at an initial sample $\mathcal{D}^{(1)}$ from the data distribution and converging to the RBM data generating distribution after performing alternating blocked Gibbs sampling between the latent and observed variables. A naive implementation of this learning algorithm would require simulating the Gibbs sampler to equilibrium after every model update before drawing a new set of observations from the data. Sampling from the generative model to produce $\mathcal{D}^{(1)}, ..., \mathcal{D}^n$ with $n \leq 10$ and updating the model parameters between each iteration has been suggested as an optimal tradeoff between fast estimation without loss in generality or stability [104]. The first term on the right-hand side of Eq. (5.27) is the derivative of the energy function w.r.t the initial Gibbs samples and the second term corresponds to the gradient of the energy function after n steps.

Our proposed modification to the RBM learning algorithm uses a hybrid generative learning and maximum utility estimation. Rather than focusing solely on the optimization of the generative component, we also try to maximize the accuracy of our choice model given the data and generative samples. After each generative learning step, we update the choice model coefficients by performing maximum likelihood on the *conditional* using the choice alternative as the dependent variable. Next, we sample latent variables from the generative model using the observed explanatory variables as inputs. These latent variables are assumed to represent the information heterogeneity that is not captured by the explanatory variables. Our modification provides integration with discrete choice modelling methods and allows for other hybrid choice model use cases that can be explored in the future. We specify the conditional logit model using observed and latent variables as follows:

$$p(y_j = 1|\mathbf{x}, \mathbf{s}'; \beta_j, c_j) \tag{5.28}$$

$$= \frac{\exp\Big((\beta_j + \mathbf{d})\mathbf{x} + c_j + \sum_h \log(1 + \exp((\mathbf{x}\mathbf{W})_h + \mathbf{W}'_{hj} + \alpha_h))\Big)}{\sum_{j'} \exp\Big((\beta_j + \mathbf{d})\mathbf{x} + c_{j'} + \sum_h \log(1 + \exp((\mathbf{x}\mathbf{W})_h + \mathbf{W}'_{hj'} + \alpha_h))\Big)}, \forall\{\mathbf{x}, y\} \subseteq \mathcal{D}, \tag{5.29}$$

where there are j alternatives in the choice variable y. In this step, only the β coefficient and c_j alternative specific constants are updated (by maximum likelihood) while keeping the parameters from the generative model unchanged. Given that parameters $\mathbf{W}'_{hj}$ and α_h are estimated from the generative model learning algorithm providing model error correction, the coefficients of the choice model is expected to converge to a non-biased, homogeneous value. This means that as we improve the precision of the data generation protocol, the choice model can be estimated without systematic errors.

5.3.5 Economic interpretation

The basis for the economic interpretation of a generative model is through a combination of individual utility and entropy. Suppose that an individual will be in one of S latent decision states, each state has associated with it a configuration of latent variables: $\{s_1, ..., s_h\}$. These latent variables are related to choice selection strategies, complexity and influence of repeated nature of travel activity choices. Thus they are interpreted as potential decision strategies. If in a particular state S contains all zero elements, then the choice strategy is a pure utility-driven one (since latent variable attributes are ignored). If by contrast, the latent variables are *non-zero*, then one might argue that the individuals used their internal information processing constraints to develop a choice strategy. These interpretations are similar to the rational inattention model, which were identified as decision strategies characterized by continuously optimizing agents [18].

We assume some distribution function to describe G_j, an error generating density function that depends on $\{s_1, ..., s_h\}$ for all j alternatives. The density G_j is the distribution of the unobserved heterogeneity on the individuals with similar utilities for each alternative. It represents the idealistic subjective perception of a particular individual in a specific choice context. We assume that $\{s_1, ..., s_h\} \in [0,1]$ are extreme value distributed across individuals and decisions:

$$G_j(s_1, ..., s_h) = \prod_h (1 + \exp((\mathbf{x}\mathbf{W})_h + \mathbf{W}'_{hj} + \alpha_h))^{-1} \tag{5.30}$$

This specification allows a form of energy-based models to be generated using entropy as a measure without relying entirely on hypothesis-driven utility specifications [27]. As such, from Eq. (5.29), the generative model specification under a generalized extreme valued function can be derived as follows:

$$P(y_j) = \frac{Y_j G_j}{\mu G}, \tag{5.31}$$

where $Y_j = e^{\nu_j}$, $\nu_j = (\beta_j + \mathbf{d})\mathbf{x} + c_j$ and $G(s_1, ..., s_h) = \sum_{j'} Y_{j'} G_{j'}$. $G(s_1, ..., s_h)$ is non-negative, homogeneous of degree μ and function $(s_1, ..., s_h)$ is ≥ 0, $G = \infty$ when $s_\kappa \to \infty$ for $\kappa = 1, ..., h$ and $\partial^r G / \partial(s_1, ..., s_h) \geq 0$ if r is odd and ≤ 0 if r is even. Thus, the level of uncertainty in a choice due to information heterogeneity is described using a function calculated on a set of prior weights and latent variables. The resulting approximate entropy is given as the negative log of the error generating function:

$$H_j = -\log G_j(s_1, ..., s_h) = \sum_h \log(1 + \exp((\mathbf{x}\mathbf{W})_h + \mathbf{W}'_{hj} + \alpha_h)) \tag{5.32}$$

We can expand the model from an MNL specification by substituting $V_j = \nu_j + H_j$:

$$P(y_j) = \frac{e^{V_j}}{\sum_{j'} e^{V_{j'}}} = \frac{e^{\nu_j + H_j}}{\sum_{j'} e^{\nu_{j'} + H_{j'}}} \tag{5.33}$$

where the arguments in V_j are linearly separated into the observed utility ν_j and entropy H_j. Thus the probability of choosing an alternative is a function of the observed utility, corrected by the information processing cost of the set of alternatives *and* its explanatory variables observed by the decision-maker. An interesting consequence is that H_j changes at every instance in the variable space, i.e. individuals with similar utilities may have different choice distributions. Furthermore we can conclude that the changes in the decision making policy are influenced in two ways: first, through the direct correlation with the *observed* attributes and second, indirectly through the information processing capacity of the decision-maker. As a result, even though it is impossible to directly measure the result of economic policy changes on the latent variables, we can obtain the mean and variance of the latent parameter distribution to evaluate the information sensitivity with respect to each explanatory variable.

5.3.6 Statistics for model evaluation and validation

One of the ways to obtain statistics for model evaluation and validation is through simulation and hyperparameter search. Model evaluation can be performed on out-of-sample simulations using adjusted R^2 serves as an equivalent to KL divergence to determine distribution accuracy. For evaluation, we fixed some of the input data and used the generative model to produce new data and compare their distribution accuracy.

There are no exact solutions to the number of latent variables required to create an optimal model. The most commonly used approach is to validate the model by iterative tests on the various number of latent variables. We note that validation is only a crude test of performance, and there are generally no accepted methods to adequately determine the optimal number of variables. Several studies in literature have provided the so-called 'rule of thumb' regarding the number of inputs and layer sizes [121]. However, the optimal number of latent variables used can differ largely between datasets. Too few latent variables and the model cannot capture the intricate structure in the data, too many latent variables may cause overfitting and increases estimation time.

Evaluating the sensitivity of parameters associated with the explanatory variables can be more challenging. In our experiment, we found that monitoring changes to β-parameters as we increase the number of latent variables work well for sensitivity analysis. Theoretically, for variables not influenced by information processing constraints, β-parameters should remain consistent. Otherwise, for variables that are sensitive to information processing constraints, β-parameters would vanish or shrink to a small value as we increase the number of latent variables so that the choice response could not have been derived from that source [18]. From a macroeconomic perspective, the decision making actions should respond smoothly to external factors and any disturbances or randomness should be distinctive and manifest only from an individual's internal information-processing constraints [18].

5.3.7 Comparison with supervised neural networks

The probability distribution in Eq. (5.29) might seem to be equivalent to a single layer neural network (e.g. DNN) with a *softmax* output, we argue that this is not the case. In a DNN, model parameters are optimized to maximize a *predictive* output $p(y|\mathbf{x}, \mathbf{s})$, which may result in significant overfitting if model is mis-specified or too many hidden units are used. Using multiple hidden layers may also potentially degrade the model and result in worse performance [122]. However, in our approach of using generative modelling, parameters are optimized to reduce information loss by minimizing $D_{KL}(q||p)$ in the mapping process between observed and latent states, allowing as much of the original data to be reconstructed. A generative model provides some form of model generalization such that the parameters stay within the range of values that are realistically *representative* of the underlying behaviour, reducing the probability that the model overfits the choice variable.

Since latent variables are stochastic, $\tilde{\mathcal{D}}$ may not always be generated by the same underlying configuration. Likewise, each sample of observed data vector may produce many different configurations of latent variables. The advantage of using unsupervised learning over supervised likelihood learning methods in discrete choice modelling is that it provides a flexible, high-level distributed representation and minimizes optimization inefficiencies caused by random initialization [123]. Model optimization uses a greedy learning algorithm to determine the underlying structure that captures the unobserved heterogeneities without dependency on aggregate choice samples. Similar to rational inattention models, entropy in the variational free energy function is the cost of information from sampling from the generative model.

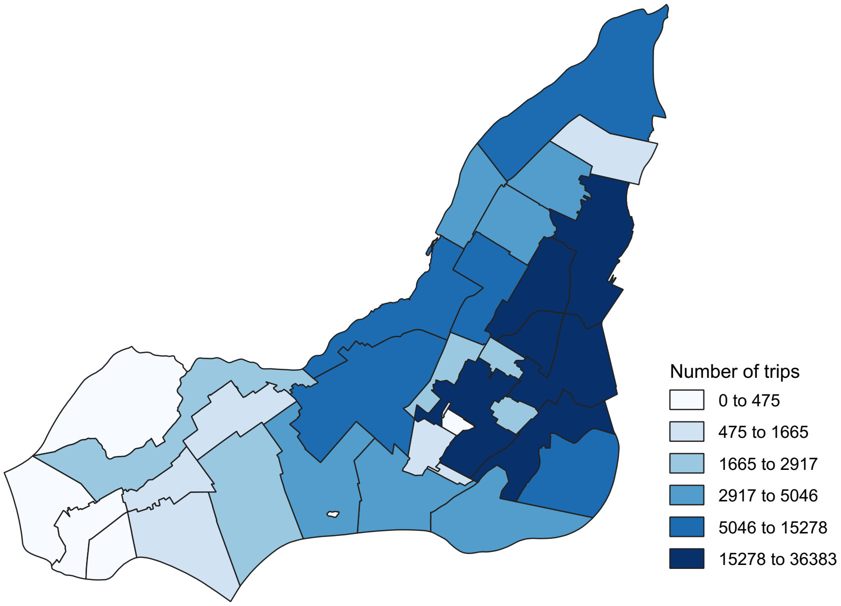


Figure 5.2: Visualization of number of trip trajectory origin points by city district from the dataset.

5.4 Case Study: Montreal Trajet Dataset

5.4.1 Data preparation

We consider a dataset collected from trip trajectories recorded by respondents from the Greater Montreal Metropolitan Area (Fig. 5.2). The data is available as an open dataset provided by the City of Montreal [124] (see Table D.3 for data description). A total of 293,330 trips observations are available in the dataset, and 58,034 trips within these observations have complete travel mode information, purpose and trip characteristics. We divide the data into two partitions: The first dataset ($\mathcal{D}_{lab}$, $N_D = 58,034$) contains complete (labelled) trip data and is used for model training and validation. The second dataset, ($\mathcal{D}_{unlab}$, $N_D = 235,296$) contains incomplete data (unlabelled) and is used for model training, validation, model simulation and analysis.

For model evaluation, we train a generative model using $\mathcal{D}_{unlab}$ and $\mathcal{D}_{lab}$ then we compute the mode choice log likelihood on $\mathcal{D}_{lab}$ for validation. The samples are randomly shuffled and split 70:30 for training and validation. We assume a multinomial extreme valued distribution for categorical observed variables and log-normal distribution for continuous variables. Log-normal is used as the approximation distribution since the continuous data types (speed, distance and duration) follow a positive, right-tailed distribution characteristic. Respective trips of individuals were recorded by self-imputation of their activity for each instance. Routes of individuals are sampled by GPS traces from their smartphones at frequent intervals. Speed, distance, activity type, trip duration and trip start location were used as explanatory variables in the estimation. The alternatives are 1: cycling, 2: driving, 3: driving + transit, 4: transit and 5: walk. Continuous valued variables were normalized to unit standard deviation before model estimation. A *one-of-j* dummy variable encoding was applied to categorical variables. A sine/cosine 2D transformation was applied to cyclical continuous values, e.g. time information.

5.4.2 Choice model validation

We present the results of our model validation by assessing the model training performance and analyze the properties of the estimated parameters. We report the results of our training and validation on model instances with different latent variable sizes: $S = 0$ (standard MNL), $S = 5, S = 20, S = 35$ and $S = 50$. In our experiments, we did not notice any significant improvement over 50 latent variables in our model. To minimize the probability of overfitting in the generative model training, we validate the generative model by monitoring the likelihood loss on the labelled data and select the model parameters at minimum likelihood validation loss.

We used a standard batch stochastic gradient descent (SGD) learning algorithm divided into k data batches and iterate over n blocked-Gibbs sampling steps. We fixed the hyperparameters for all our experiments to be $k = 16$, $n = 10$, and a learning rate of $\lambda = 0.01$ is used, and model parameters are updated in parallel every batch cycle[4].

[4] The problem of identifying optimal hyperparameters is still not fully understood, and it does not provide any useful information with respect to econometric interpretation. In light of this, we selected these hyperparameters as our baseline for the ease of reproducibility in future work.

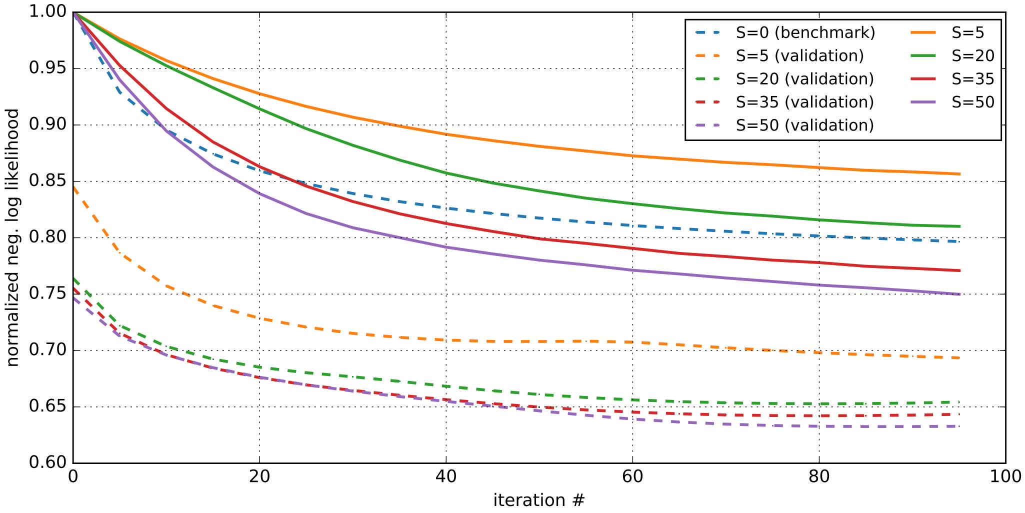


Figure 5.3: Learning curve of the sample negative loglikelihood from the choice model.

We monitor validation error by computing the total negative log-likelihood of the validation data over the choice model at each iteration. As observed in the learning curves (Fig. 5.3), the model estimation process is stable and converges gradually without overfitting. At $S = 50$, the model achieved the best overall performance in terms of the validation log-likelihood. However, the relative gain in performance decreases as we increase the number of latent variables. We hypothesize that there is a maximum bound to the effective possible number of latent variables to represent unobserved variations in the data. This limit can be raised if a more considerable variation in data is used, i.e., data from different sources or over a more extended collection time frame. Note that this analysis is not a test for the 'best' mode – our primary objective is to understand the sensitivity of econometric parameters when a generative learning model is used to account for information heterogeneity. The loglikelihood decreases rapidly for the first 20 iterations, then plateaued as it reached 100 iterations. Estimation time for each model instance was less than a 1 hour running our code on a GPU hardware.

Model performance is evaluated by comparing the adjusted squared correlation $\bar{R}^2$ statistical fit. Fig. 5.4 shows the mode share distribution of the model validation.

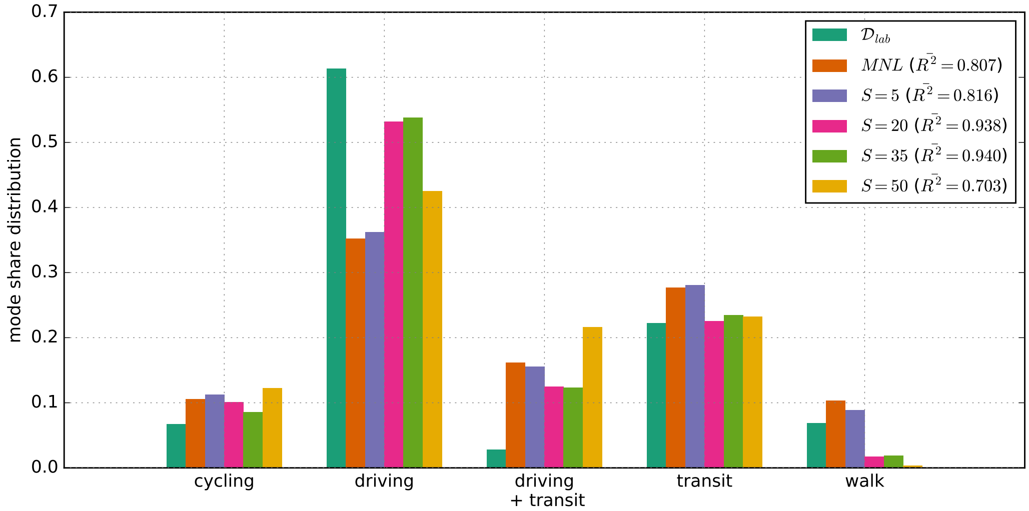


Figure 5.4: Mode share forecast.

For the baseline model, we obtained a $\bar{R}^2$ value of 0.807 We obtained a $\bar{R}^2$ value of 0.940, representing a 15% increase in relative predictive performance. The nominal trend shows that distribution accuracy increases with an increase in the number of latent variables. At $S = 50$, performance drops slightly compared to $S = 35$ indicating that the performance does not increase asymptotically with the number of latent variables. Nevertheless, the results show that the model can be estimated with high accuracy, using KL divergence over maximum likelihood as the objective function. In this example, the models do not consistently predict the *driving+transit* and *walking* alternatives probabilities. One explanation can be attributed to the low observation counts of these two alternatives. Another possible explanation is that *driving+transit* and *walking* trips have a low correlation with the observed explanatory variables.

5.4.3 Latent variable analysis

To understand the representational value of latent variables, we analyze their sparse-overcomplete properties [119]. Sparse-overcomplete representation a situation when a large number of latent variables are estimated while only a small number of them are non-zero [119]. It is a practical constraint that allows for more efficient use of latent variables and more flexibility in handling complex correlations which results in a better approximation of the statistical distribution of the data. Sparse representation has two main advantages in generative modelling [125, 52]. The first advantage is that the model will be able to control the dimensionality of representation, given a set of inputs, avoiding the overfitting problem. The second advantage in the context of travel behaviour model inference is that the resulting representation is more likely to be linearly separable, decreasing the complexity in the model even though more parameters are estimated. This means that even with a large number of latent variables, sparse distribution of parameters would constraint the model to learn distributions which are most statistically significant in reproducing the original data.

The plots in Fig. 5.5 show the mean and variance of estimated latent variable parameters $\mathbf{W}'_{hj}$ given the choice outputs. Since we use binary coding for latent variables, the parameters offer insights into how many latent variables are utilized at any one time. Parameter vectors with mean values close to zero and low variance indicate that the latent components are sparsely distributed. We assume that overcomplete representation ($S \geq X$) does not cause model overfitting as not all latent variables are active. The figure shown below illustrates that our generative modelling approach is an efficient method of capturing the underlying heterogeneity across different mode choice decisions. The mean converges to zero, and standard deviation decreases as the number of latent variables increase, indicating that the generative model 'suppresses' the influence of less relevant latent variables on the behaviour model.

The results suggest that the RBM learning algorithm inhibits weight connections between the observed and latent variables in order to produce sparse representation. At ($S = 50$), the mean parameter activation is near zero with small standard deviation ($\mu \leq 0.02, \sigma \leq 0.17$) for *cycling*, *driving*, *driving + transit* modes with an average latent variable activation rate of 85.4%, 84.4% and 87.7% respectively. For *transit* and *walk* modes, the average activation rates are 90.6% and 92.6% respectively, indicating that these modes have a higher level of information heterogeneity and less correlated with the observed explanatory variables.

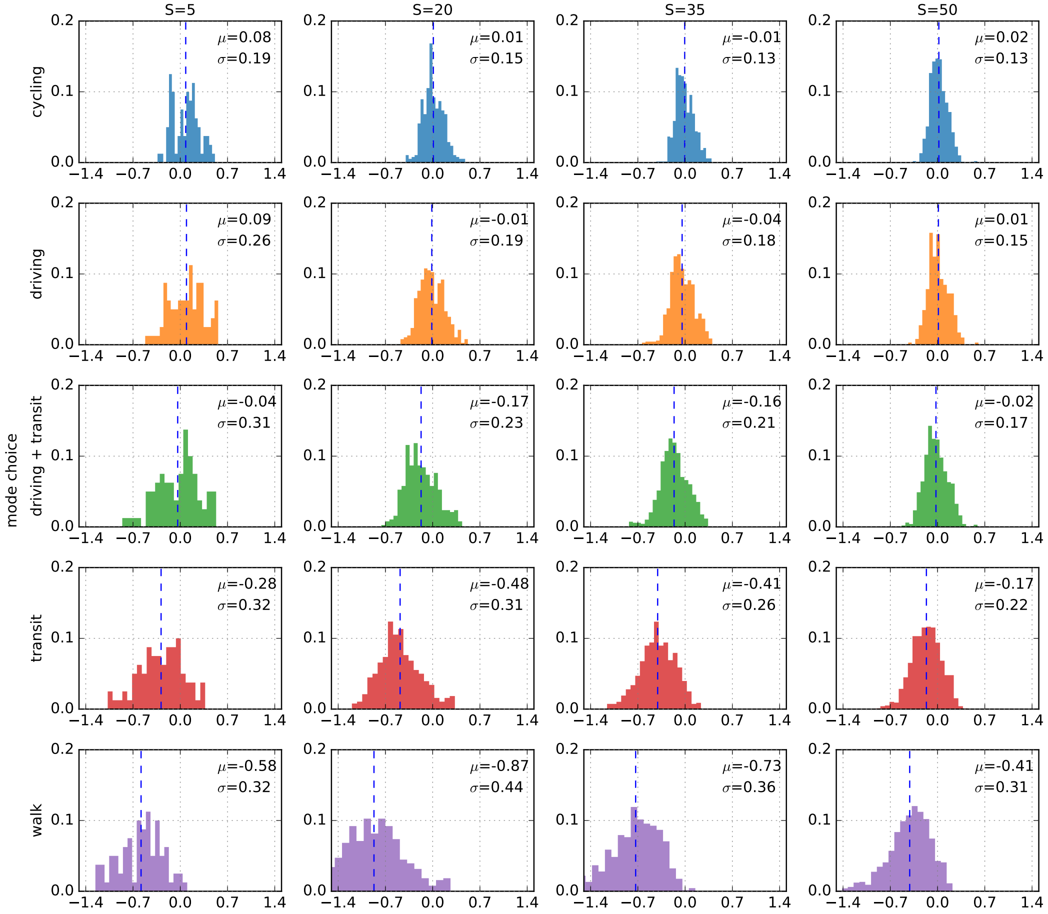


Figure 5.5: Distribution of data generating parameters.

5.4.4 Generative model evaluation

To evaluate generative model performance, we measure the statistical fit of the reconstructed distribution. Simulated reproduction of population data has been used previously to analyze the efficiency of model-based fitting [126]. Simulation experiments allow evaluation of the model on limited data knowledge, reproducing accurate data distribution while having partial information shows flexibility in capturing decision heterogeneity due to information constraints. Therefore, the performance re-

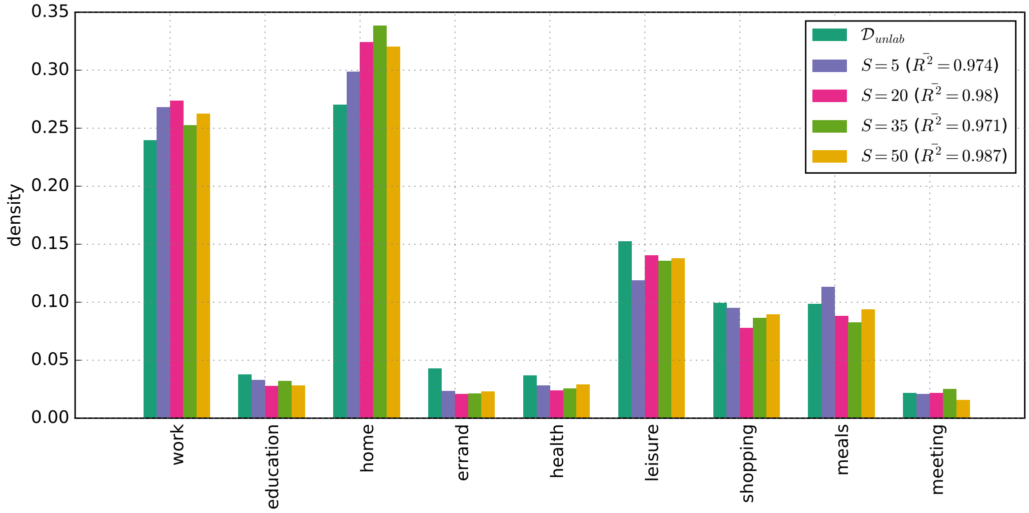


Figure 5.6: Comparison of data generating output on activity type data.

sults of these simulation experiments can be used to calibrate large scale data-driven models where complex data correlation is present and accounts for the presumption that individuals have limited information processing capacity in choice selection. We use Gibbs sampling to obtain data from the generative model. First, evaluate the data generating distribution accuracy using the unlabelled dataset $\mathcal{D}_{unlab}$. Fig. 5.6, Fig. 5.7 and Fig. 5.8 shows the data generation results for activity, distance and trip duration variables respectively.

Next, for the data generating process, we draw an initial sample from the dataset and fix the observed variable to that data vector and perform Gibbs sampling, alternating between the latent and observed sample conditional probabilities. Lastly, we clamp the non-target variables to the data vector and update the simulated values of the target observed variable. For instance, we generate activity type data using the following steps:

$$\{\tilde{s}_1, ..., \tilde{s}_h\} \sim p(s_1, ..., s_h|\text{speed,duration,dist,origin,destination}),$$
$$\tilde{x} \sim p(\text{activity}|\{\tilde{s}_1, ..., \tilde{s}_h\})$$

The simulation results show the effects of increasing latent variables on the

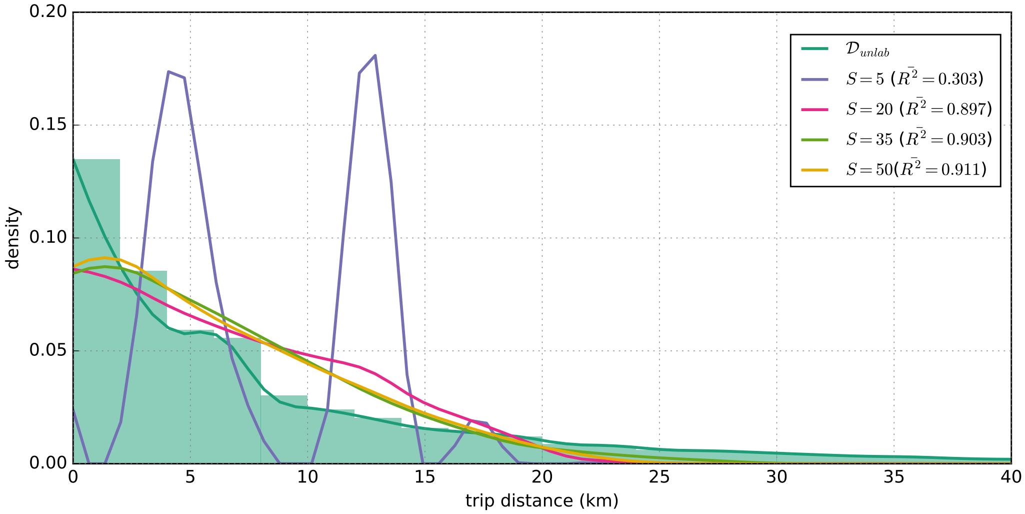


Figure 5.7: Comparison of data generating output on trip distance data.

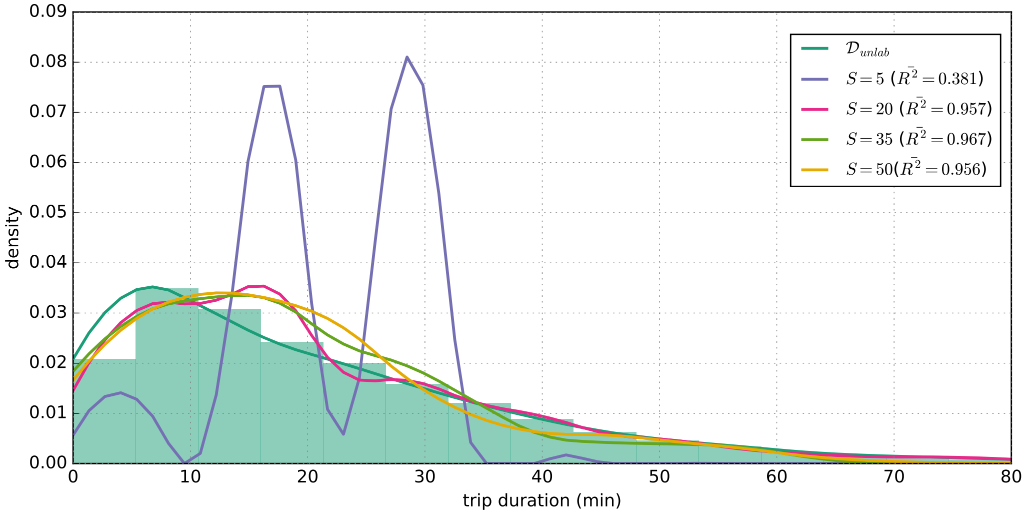


Figure 5.8: Comparison of data generating output on trip duration data.

performance of the data generating model. $S = 35$ and $S = 50$ achieved high similarities in recovering the original data distribution with $\bar{R}^2$ value well above 0.9. At $S = 5$, there was an insufficient number of latent variables to capture the structure of the data, shown by the low $\bar{R}^2$ value. Increasing to $S = 20$ significantly improves the result as it increases the non-linear information capacity.

5.4.5 Sensitivity analysis of model parameters

Finally, in this section, we investigate the systematic effects if the generative framework on β-parameters in the mode choice model. In practice, bias and variances are subject to independent processes. As such, each individual may have vastly different underlying error correction function for the same utility and each configuration of explanatory variables. Mixed Logit specification has been used previously to account for this problem, but unfortunately, any variability or noise in the dataset (e.g. through different collection techniques, missing information etc.) will be added to the β-parameter model predictors. This is less of a problem if one is only interested in the relative variance given the model parameters. To account for the systematic effects of information heterogeneity, the net utility of each alternative should remain homogeneous across the population (e.g. zero noise level), such that the latent constructs can compensate for the degree of uncertainty.

Fig. 5.9 shows the estimated β-parameters of the choice models with different number of latent variables. The β-parameters identify the systematic effects of each explanatory variable on each choice alternative. The values on the left edge of each plot show the β-parameters estimated with a standard MNL model. As we increase the generative model capacity (by increasing the number of latent variables), β-parameters converge to a stable predictor. This is an interesting finding as it may indicate that an ordinary utility-based choice model may not take into account the systematic effect of information heterogeneity.

We perform a test on the identification of the β-parameters by computing the maximum entropy (maxent) estimate on the observed choice probability in the dataset shown in Table 5.1. The maxent estimate value quantifies the degree of uncertainty within the underlying model accounting for the complexity as well as to determine whether the variance can be attributed to information heterogeneity. Analysis of the maxent can provide information about the uncertainty of the predictors across choice probabilities [127]. We compute maxent of the explanatory variable parameters using the formula:

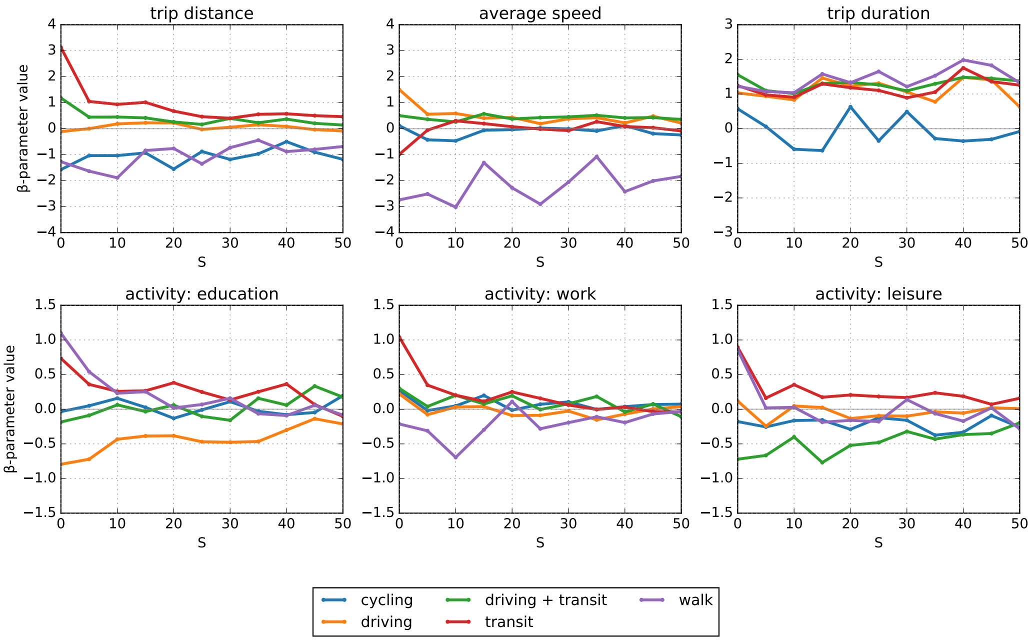


Figure 5.9: β-parameter estimates using mode choice as the dependent variable, horizontal axis represent number of latent variables.

$$maxent(\beta_j) = -\sum_j p(y_j) \log p(\hat{y_j}) = -\sum_j p(y_j) \log \left(\frac{e^{\beta_j}}{\sum_{j'} e^{\beta_{j'}}} \right) \tag{5.34}$$

where the population class share for each alternatives $p(y_j)$ are: cycling=0.068, driving=0.613, driving + transit=0.028, transit=0.222 and walking=0.069 from the labelled dataset. The resulting $maxent(\beta_j)$ may, therefore, be interpreted as the maxent estimate of β_j as the proportion of the sample population in alternative j. Likewise, a high maxent value indicates a high degree of stochasticity in the decision-making process. We find $p(\hat{y_j})$ by computing $(e^{\beta_j} / \sum_{j'} e^{\beta_{j'}})$. As the negative entropy increases, e.g. $maxent(\beta_j) \to 0$, the correlation between the β-parameter and choice probability converges to the true value, e.g. $p(\hat{y_j}) \to p(y_j)$.

The maxent estimate indicates the level of correlation between the set of β-parameters and the output dependent choice variable. Table 5.1 shows that the β-parameters for distance (2.833) and education activity (2.234) variables in the

Table 5.1: Result of maxent estimates on β-parameters.

Parameters β_j	$maxent(\beta_j)$				
	S=0	S=5	S=20	S=35	S=50
Distance	2.833	1.721	1.511	1.518	1.568
Trip duration	1.706	1.600	1.591	1.807	1.847
Speed	1.532	1.456	1.513	1.503	1.538
Activity: Edu.	2.234	2.038	1.781	1.834	1.756
Activity: Work	1.640	1.619	1.693	1.696	1.584
Activity: Leisure	1.677	1.596	1.538	1.517	1.512
mean (std. dev.)	1.94 (0.502)	1.67 (0.199)	1.61 (0.11)	1.65 (0.153)	1.63 (0.135)

benchmark model are less likely to influence decisions relative to the other predictors and becomes an indicator of model misspecification. However, as we increase the number of latent variables in the generative model, maxent decreases, and as such, the β-parameters becomes a better predictor of the behaviour. This suggests that the mode choice decision behaviour of individuals is less sensitive trip distance and education-related activities.

The econometric interpretation of this result implies that individuals seek to use their prior information (e.g. past experiences, habits, choice dynamics) for mode choice decisions rather than driven by exogenous variables. The significance of the distortion effect of information heterogeneity on the β-parameters decreases as we include a larger correction in the utility function. This apparent correlation provides evidence that in order to maximize utility and therefore better model prediction accuracy, latent variables can be incorporated in the framework to model information heterogeneity – The generative model accounts for the variational effects from information heterogeneity, increasing regularity in the utility specification.

Consequently, the estimated β-parameters would reflect the true underlying predictors. As observed earlier that expected utility can be modelled by the individual's decision strategy shown by evaluating entropy (by a function of latent state vectors) of the choice model.

5.5 Discussions and Conclusion

5.5.1 Discussions

Our findings have several important policy implications. First, we have shown that by optimizing a set of internal latent variables to represent distinctive decision strategies of each individual, we can emulate information processing and learning-based decision-making behaviour incorporated into a choice model. We tested the framework and learning algorithm on the dataset to emulate information processing constraints in travel behaviour and decision making. Our methodology consists of applying an entropy-based error component that used latent constructs in a generative learning model to optimize a set of parameters that minimizes a divergence between the observed and simulated data.

Second, following behaviour theory in discrete choice analysis, our generative model showed that individuals may not always be utility maximizers and therefore MNL models alone may not be sufficient in modelling travel behaviour in large scale datasets. We have shown that maxent estimates of β-parameters can be reduced by having a learning model component that captures information heterogeneity, population and decision level variance and incorporating the entropy function into choice utilities. Our analysis and simulating experiments have shown that β-parameter estimates in Fig. 5.9 scales according to the number of latent variables in the model and it shows significant improvements to choice probability predictions. The learning framework was able to extract useful information from the dataset, with the assumption that information heterogeneity is present in the data. The changes in maxent shown in Table 5.1 indicated that the β-parameter has a high level of information heterogeneity, and the misspecification is minimized by incorporating latent variables through a learning process emulated by a generative model. Information theory motivates the explanation for this phenomenon: breaking down the processing costs of information related to the choice into a linearly separable component serves as a regularization term in the utility specification.

Lastly, it would suggest that distance-based trip planning is more strongly correlated to long term individual habits and perception of the travel route and less likely due to specific change in trip distance. Our experiments showed how some explanatory variables could contain a more significant source of information heterogeneity and increasing the generative model capacity increases the choice probability accuracy more robustly. The results indicated that the improved model fit could be attributed to more efficient use of the generative model, which suggests that stochastic choice selection in decision making can be associated with the availability individual's prior information.

5.5.2 Conclusion

Generative modelling presents a new perspective on how analysts can obtain insights into behavioural heterogeneity manifestations by accounting for information processing constraints in the model learning process. Based on rational inattention behaviour and information theory, we develop a systematic approach to identify information heterogeneity, and we propose a data-driven generative learning process to emulate decision making under uncertainty and information processing constraints. It explains why not all exogenous information is used in the decision-making process, as discussed in [58].

The impact of this study on travel demand modelling is that we can take advantage of noisy data (e.g. GPS, Wi-Fi, cellular networks) to develop a flexible, operational, and adaptive model framework. Our underlying assumption is that large and unstructured data from passive information sources that contain behavioural information not captured in explanatory variables can be exploited with the proper learning models and optimization algorithms. This study demonstrates the properties and expressive power of the generative modelling framework to emulate decisions under uncertainty and information processing constraints. We define the source of heterogeneity to be the inherent nature of the data itself, and by updating the model using an iterative KL divergence minimization process, we can synthetically reproduce the unobserved variations using latent constructs in a generative model. The latent constructs provide additional error correction for information heterogeneity in the utility specification, allowing the model to simulate decision making and choice actions with internal information processing components. It also allows a convenient representation of entropy by incorporating an error generating function into the framework. Our results indicate a strong correlation with rational inattention behaviour theory, which shows that individuals may tend to ignore certain explanatory variables or rely on prior information for discrete choice decision making. The experiments identify several vital components of the generative model, which are more sensitive to information heterogeneity and apply an automatic correction for this variation by representing the heterogeneity as an entropy measure in the utility specification. More generally, principles from generative modelling demonstrated in this paper can be applied to existing travel behaviour analysis to benefit from using large data sources, where latent behaviour information are not directly captured in the explanatory variables.

5.5.3 Future work

The scope of this paper focuses on the implementation and basic methodology of developing a machine learning-based generative model for discrete choice analysis. There are several extensions to this study which can be addressed in future work:

(i) Exploring the use of variation inference techniques in Mixed Logit models to address estimation tractability, allowing for a comparative analysis between discrete choice and machine learning-based methods.

(ii) Several other variants of generative model learning algorithms (e.g. GANs, Autoencoders) could be tested to obtain insights into how they would emulate distinctive social and cognitive behavioural concepts. Additionally, generative modelling can be extended to other constraints beyond information processing costs, for example, budget and time constraints.

Chapter 6

A Bi-partite Generative Model Framework for Analyzing and Simulating Large Scale Multiple Discrete-Continuous Travel Behaviour Data

Preamble

This chapter features an application of generative machine learning in multiple discrete-continuous data modelling. It shows how forecasting and simulation can be implemented using a RBM framework. In addition, evaluation is performed using conventional methods of behaviour analysis – elasticities of model parameters, parameter stability and moment analysis. Finally, the experiment and methodology highlighted in this chapter connects to the broader scope of the transport an mobility market by enabling the use of interpretable machine learning in demand forecasting systems.

This research article is under review in *Transportation Research Part C: Emerging Technologies*, special issue on Emerging Methods for Data-driven Urban Transportation and Mobility Modelling: Machine Learning and Complexity Approaches.

Abstract

The emergence of data-driven demand analysis have led to the increased use of generative modelling to learn the probabilistic dependencies between random variables. Although their apparent use has largely been limited to image recognition and classification in recent years, generative machine learning algorithms can be a powerful tool for travel behaviour research by replicating travel behaviour by the underlying properties of data structures. In this paper, we examine the use of generative machine learning approach for analyzing multiple discrete-continuous (MDC) travel behaviour data. We provide a plausible perspective of how we can exploit the use of machine learning techniques to interpret the underlying heterogeneities in the data. We show that generative models are conceptually similar to choice selection behaviour process through information entropy and variational Bayesian inference. Without loss of generality, we consider a restricted Boltzmann machine (RBM) based algorithm with multiple discrete-continuous layer, formulated as a variational Bayesian inference optimization problem. We systematically describe the proposed machine learning algorithm and develop a process of analyzing travel behaviour data from a generative learning perspective. We show parameter stability from model analysis and simulation tests on an open dataset with multiple discrete-continuous dimensions from a data size of 293,330 observations. For interpretability, we derive the conditional probabilities, elasticities and perform statistical analysis on the latent variables. We show that our model can generate statistically similar data distributions for travel forecasting and prediction and performs better than purely discriminative methods in validation. Our results indicate that latent constructs in generative models can accurately represent the joint distribution consistently on MDC data.

6.1 Introduction

Large scale ubiquitous multidimensional travel data sources such as smartcard data or on-demand ride-sharing services provide enormous potential for travel behaviour analysts to implement new and innovative methods and algorithms for travel behaviour pattern forecasting [128, 129]. In addition to size, these abstract data are also increasing in complexity, which necessitates data pruning or sub-sampling techniques to extract useful information and to improve estimation time at the cost of model accuracy. Until recently, the most popular approach for travel behaviour modelling applications was hypothesis-driven discrete choice models (DCM). At the core, DCMs consist of defining a set of rules for Random Utility Maximization (RUM) [27]. For instance, RUM have been been used in estimating route choice models with traffic network and socio-demographic information, including regret minimization [32], prospect theory [23] and the rational inattention model [113]. Generative modelling proposes an alternative approach to analyzing travel behaviour data by constructing a model of the underlying distribution using unsupervised learning to generate new data with similar stochastic variations as the population. In contrast, DCM is optimized from the maximum utility by estimating conditional probability distributions through a hypothesis-driven process with assumptions on the prior distributions. Generative modelling also relates to classical statistical methods, i.e. Information Theory and Shannon entropy [130]. When applied to travel behaviour datasets, the generative model behaves as an information processing constraint of the individuals as part of their decision process. Individuals may weigh the information cost of changing travel habits, e.g. mode choice or route choice, given some known characteristics of the competing alternatives and this decision process is assumed to be continuous and simultaneous.

The benefits of using generative modelling are tied to behaviour theory and information processing cost in macroeconomic problems – generative models provide a more plausible framework for understanding selective and dynamic responses [60]. Previous work has provided a theoretical explanation to these interactions using artificial neural networks and how sensory information is reconstructed through generative modelling [60, 39]. The goal of this study is to present generative models as a behaviourally intuitive representation of travel decision making with an endogenous learning process. We argue that the main advantage of generative machine learning is that we can rely less on hypothesis-driven behaviour assumptions and representing decision perturbations beyond unobserved utility terms [131]. Recent developments in artificial neural network and learning algorithms have made it possible to estimate complex and non-rational behaviour (relaxation of IID assumptions) models that generalize better to various decision-making strategies [30]. This paper offers a plausible perspective of how we can exploit the use of emerging machine learning techniques to model the behavioural processes prior to decision making actions.

We propose an extension for generative machine learning to accurately model multiple discrete-continuous (MDC) large-scale travel behaviour data. We show that our proposed model can generate reasonably accurate data reconstructions, given suitable data observations and capacity for training. Our proposed generative model provides a simple and intuitive mechanism for understanding the trade-offs between entropy and utility-maximizing behaviour by resolving uncertainty using variational Bayesian inference methods.

The main contributions of this paper are summarized as follows:

- We propose a bi-partite generative model to handle large travel behaviour datasets with MDC data types using an RBM learning algorithm;
- Systematically describe the machine learning framework used to train the generative model using a variational Bayesian inference objective function;
- Show how an information-theoretic model leads to economic behaviour compatibility that can be understood as: (a) lower evidence bound that depends on a variational free energy function, and (b) a measure of risk minimization that approximates the posterior distribution;
- Develop analytical methods to generate conditional probabilities, elasticities and latent variable distributions that can be used for interpretation and economic analysis.

With the emergence of data-driven demand and services that use abstract forms of data, for example, social media data, there is a need to understand the underlying properties and correlation between 'Big Data' sources and choice actions to model travel behaviour using the potential of modern generative and deep learning techniques. This paper aims to bridge the gap between traditional means of travel behaviour analysis dependent on identifiable variables and using abstract data that require machine learning techniques to extract useful information. The novel approach tackles the problem of representing information heterogeneity in data-driven behaviour models using a joint distribution of discrete and continuous data.

This paper is organized as follows: In Section 6.2, we explain the background of the generative model and the variational Bayesian inference method. In Section 6.3, we describe our adaptations of generative machine learning methods, implementation on discrete and continuous travel behaviour datasets and optimization using variational Bayesian inference. In Section 6.4, we present the case study. Results on large scale travel data are in Section 6.5. Finally, discussions and conclusions are in Section 6.6.

6.2 Literature Review

Conventional DCM are used to estimate travel behaviour models from large scale multidimensional geospatial datasets e.g. GPS systems [132, 133]. However, missing or noisy data could lead to inaccuracy in model estimation and may require the incorporation of latent variables. In transportation, obtaining useful information from these datasets may be difficult because important trip details (mode choice, pricing, number of passengers, etc.) cannot be recorded directly from GPS data points [134]. Another obstacle is defining a generalized framework for incorporating latent variables or missing data points into multidimensional choice models. Latent variables are essential in travel behaviour modelling as they capture behavioural perceptions related to uncertainty and describes the underlying mechanism of the choice selection process [72]. However, model specifications with complex distributions may not produce an identifiable closed-form solution for maximum likelihood estimation. For the above reasons, researchers have implemented Monte Carlo methods and variational Bayesian inference for analytical approximations to incorporate mixed distributions and choice dynamics into the model estimation process [135, 136].

Variational Bayesian inference combines prior knowledge and empirical evidence to resolve uncertainty and adapt to noisy datasets through data-driven algorithms such as neural networks and generative models [137]. Variational Bayesian inference methods are widely used in machine learning with successful applications in data mining and sentiment analysis [47, 138]. In classical Bayesian modelling, the posterior distributions are usually estimated by simulation or sampling-based methods. A commonly employed sampling-based algorithm for travel behaviour datasets is the Markov Chain Monte Carlo (MCMC) algorithm where the posterior distribution is simulated by drawing repeated samples from a Markov Chain until convergence [139]. The stationary distribution of the Markov chain represents the posterior distribution.

In recent years, MCMC algorithm has played an important role in travel behaviour modelling problems in transportation, with successful applications in agent-based simulations [140], hybrid choice models [141, 142], and population synthesis [126, 143, 144]. However, in order to match the asymptotic efficiency of maximum likelihood, MCMC draws must grow at a rate faster than the square root of the number of agents [27, 145]. With complex mixing distributions, convergence may not be guaranteed in a reasonable time, resulting in poor estimation. This makes sampling-based estimation methods infeasible beyond relatively simple models and small datasets for obtaining accurate results. This challenge has led to the development of convergence testing methods to assess model precision [145]. Another viable approach is the iterative Expectation-Maximization (EM) algorithm for posterior estimation [146]. Although the EM algorithm may be useful in small datasets and for incomplete data, the rate of mixing is also known to be extremely slow in some cases [147, 148].

6.2.1 Conventional MDC model estimation approaches

The conventional hypothesis-driven approach for MDC modelling is primarily by the multiple discrete-continuous extreme value (MDCEV) model [149]. It incorporates a non-linear function in the utility structure to account for choice substitutions, continuous consumption and multiple alternatives. In the MDCEV model, multiple constraints are pre-defined, hypothesis-driven based utility function. There is the assumption on MDCEV that a single baseline utility influences both discrete and continuous consumption. Although this has been expanded recently by incorporating different utility functions for discrete and continuous options [29]. Other models for estimating MDC include the *translated quadratic non-linear additive model* which provides corner solutions and diminishing marginal utility. This has been used in modelling consumer choices with multiple purchase variety [150].

Large sources of travel behaviour datasets are becoming available via new sources like social media, smartphone apps, and communication networks. There is a need for new approaches that are specifically designed for these large datasets. Our current work differs from hypothesis-driven approaches in which we develop a generative model with a joint distribution accounting for latent correlation effects in large datasets. The result is a data-driven generative model described by the underlying latent behavioural distribution, and the solution entails finding the model parameters that best replicate the outcomes. We develop the estimation procedure using a Gibbs sampling based gradient descent method, typically used in machine learning.

6.2.2 Existing developments of generative modelling in transportation

One of the key issues in discrete choice model design is the assumption that observations are drawn independently, although this assumption of often always violated in real-world problems. Alternatively, this problem can be handled by considering a more flexible model with a richer set of random variables with data-driven distributions that allow practitioners to describe a model that best represents the behaviour of the population.

In transport modelling, several studies have been conducted that investigate how probabilistic models can be effectively leveraged to model spatial-temporal data through Bayesian inference techniques. Probabilistic models have been described to be a form of ‘transfer learning scheme’ instead of traditional learning where calibration is done on a single source of labelled data [151]. Transfer learning enables relaxation of various assumptions in the modelling process and being able to reconstruct new and unseen observations from the joint probability which is useful for exploiting and extracting non-survey based data, e.g. social media data, that has little direct correlation with travel behaviour. In transport studies, model-based machine learning approaches such as generative modelling are primarily used for classification of unseen observation by identifying the latent variables that describe some contextual information not captured in the data [152]. Latent Dirichlet Allocation (LDA) [153] is another popular variation of generative modelling that is commonly used to analyze structure in the data without prior labels, for example, the discovery of activity patterns in trip modelling [154, 155].

Probabilistic Graphical Models (PGMs) describes the representation and structure of probability distributions compactly and intuitively by encoding the independence assumptions and causality between random variables in the factorized graph edges [156]. Each edge connection corresponds to the strength of direct dependence between the random variables, and each random variable can be constructed as a conditional model given the other variables and the corresponding edges. PGMs have been used for traffic simulation by representing traffic links as the graph edges and estimating the model using a first-order spatial Markov model [157]. [158] developed a PGM for realistic highway scenes by modelling vehicles as nodes and interactions between vehicles as factor graph edges. By generating novel ‘path’ probabilities between random variables, PGMs can model all types of interactions and correlations that can best represent the underlying properties of discrete and continuous data.

6.2.3 Generative modelling using artificial neural networks

Generative models are used to learn a representation of a dataset as a joint distribution over the observed variables. The joint distribution analyzes the extracted information without relating it to the observers’ prior knowledge, and these subjective measures are based on so-called information criteria, e.g., Akaike’s information criterion or Shannon entropy [130]. Subjective measures consider additional knowledge about the observation such as novelty, counter-intuitive behaviour or familiarity. Existing discrete choice models are based on such measures to represent latent behavioural information about the traveller’s behaviour such as latent class (LC) models, Mixed logit (ML) and integrated choice and latent variable (ICLV) models [36].

Early statistical methods used generative modelling for dimensional reduction such as principal component analysis (PCA), k-means clustering and linear discriminant analysis (LDA). PCA can be used as a simple dimensional reduction tool that relies on linear assumptions where each dimension (PCA latent variable) is highly correlated to each other. However, abstract data sources may not possess these properties and are more likely to be noisy, complex, and have multiple non-linear correlations. In order to sufficiently capture non-linear variations in the data, deep learning techniques can be applied.

Recently, more powerful forms of generative models are based on neural networks and have been widely used in applications such as population synthesis, semantic analysis and recommendation systems. Some of these generative models include restricted Boltzmann machines (RBM), generative adversarial nets (GAN) and variational autoencoders (VAE) [45].

RBMs are the earliest and most simple form of parametric generative models that perform representation learning by fitting the neural network model to the data. RBMs are utilized as building blocks for constructing deep artificial neural nets such as Deep Belief Nets (DBN) [50]. Inference in RBM generative models is difficult. Thus efficient training algorithms were introduced to approximate the inference procedure [100]. The general training process for RBM is a pairwise contrastive divergence algorithm which is bi-directional to allow up and downstream propagation of network weights. Synthetic data can be sampled from the trained generative model that have similar statistical properties as the input dataset. Compared with PCA or clustering based modelling approaches, RBMs have shown a strong capacity to model joint distributions and have been successfully applied to capture spatial-temporal patterns [107]. The RBM generative model restricts lateral connections within layers, which provides independent and identically distributed (IID) assumptions about the observed and latent variables. For prediction and forecasting, RBMs are typically used for learning latent features followed by either a generative simulation-based classifier or directly as a multi-layer neural network classifier [65].

Other generative models such as VAEs are used to perform non-linear mapping of the input variables to 'encodings' by compression and marginalizing out noisy data as part of the training process [159]. The 'encodings' capture the most meaningful information of the data, similar to a clustering algorithm. Estimation of VAE requires layer-wise training by optimizing the lower bound of a variational Bayesian inference objective function by applying a gradient-based updating rule. GANs are another type of generative model that trains a generator and discriminator in the neural network simultaneously. The discriminator attempts to distinguish between the real data and the generated data and minimizes the error of differentiating real from synthetic data. This method is designed to be used for semi-supervised learning and was commonly implemented on computer vision and image classification tasks [45].

6.2.4 Model optimization algorithms

The approach to solving the optimization problem in neural networks is to apply gradient descent via a backpropagation learning algorithm to calculate the gradients w.r.t. the likelihood function [56]. This formula for gradient descent is applied to the variational inference algorithm in a generative model based on the principle of energy minimization [160]. A symmetric parameterized model such as the RBM uses a Gibbs sampler starting at some random data point that would allow the neural network to update the parameters until convergence is reached. The procedure is known as *blocked Gibbs sampling* by alternating updates between 'visible' and 'hidden' neurons. However, the sampling approach requires running a Markov chain until convergence. An approach using *contrastive divergence* approximates the optimization problem by replacing the energy minimization gradient function by a fast approximate [100].

The objective of generative models is to learn meaningful ways to represent the input data through a subset of underlying latent variables. This information processing architecture was suggested as a representation of behavioural stimuli [60]. It treats choice behaviour the same way as the rational inattention model, which depends on the context formed by prior beliefs [30]. Several studies have shown the superior performance of the generative model in solving challenging decision-making problems over typical discrete choice and discriminative neural networks. To the best of our knowledge, the use of generative learning is limited to image and video data to capture motion and dynamics. Here, we extend our previous work on RBM based single discrete choice and latent variable models [65] to incorporate multiple discrete-continuous choices. We also propose a generic algorithm for estimating MDC models using generative machine learning. The trained model is used to generate conditional samples and then used to perform classification tasks as well as travel behaviour prediction.

6.3 Proposed generative machine learning approach

In this section, we describe our adaptations of current machine learning methods, introduce our generative bi-partite framework for modelling MDC data and the associated model optimization algorithm. A list of notations used throughout this paper is given in Table 6.1.

Table 6.1: Notations.

Notations	Description
$\mathbf{x}$	set of input variables $x_1, x_2, ..., x_K$
$\mathbf{s}$	set of latent variables $s_1, s_2, ..., s_J$
$\mathcal{H}[x]$	entropy of x
$D_{KL}[a\|\|b]$	Kullback-Leibler divergence of a from b
$\mathcal{F}$	variational free energy
$E(\mathbf{x})$	energy of $\mathbf{x}$
$\langle x \rangle_q$	expected value of x over distribution q
$\sigma(x)$	sigmoid function operator $(1+e^{-x})^{-1}$
$\mathcal{N}(W, \Sigma^2)$	Gaussian distribution with mean W and variance Σ^2
$\nabla_\theta(f)$	gradient of function f w.r.t. θ
η	stochastic gradient descent rate. Note: $\eta < 1$

6.3.1 Generative bi-partite model

Conventional DCM methods often face difficulties in estimating large datasets with MDC choice outputs due to exponentially increasing choice set selection [161]. Furthermore, the complexity of estimating DCM increases when incorporating hidden variables, requiring additional variational parameters while making model inference intractable and impractical. One approach we can use is to approximate each unobserved component with a point estimate. However, we cannot quantify the uncertainty or confidence interval of these hidden variables. The other approach is to find a joint distribution of the hidden and observed components and perform Bayesian analysis – this usually results in an intractable integral. The core function of generative machine learning solves the two problems by computing the integral through optimization of a variational free energy objective function and uses probabilistic Bayesian techniques to obtain the parameters of the model.

Our proposed solution is a generative bi-partite graph framework that models the underlying processes that are likely to generate the data. The assumption is that large amounts of data are available that can represent the true population behaviour. See Fig. 6.1 for an illustration of the model. First, we consider the joint distribution given as $p(\mathbf{x}, \mathbf{s})$ over the set of *binary* hidden random $\mathbf{s} = s_{1:J} \in \{0, 1\}$ and observed $\mathbf{x} = x_{1:K} \in \mathbb{R}^{\mathcal{D}}$ variables. We specify a prior distribution $p(\mathbf{s})$ about the hidden variables and quantify how $\mathbf{x}$ relates to $\mathbf{s}$ with the likelihood function $p(\mathbf{x}|\mathbf{s})$. Applying the Bayes' rule, we obtain the posterior distribution:

$$p(\mathbf{s}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{s})}{p(\mathbf{x})} \propto p(\mathbf{x}|\mathbf{s})p(\mathbf{s}) \tag{6.1}$$

where $p(\mathbf{s})$ is the hidden layer distribution, e.g., Bernoulli, multinomial or normal, that are the latent priors, and conditional densities $p(\mathbf{x}|\mathbf{s})$ are the likelihood components of the Bayesian model. If the latent priors are tractable, the likelihood component may have D_{cont} continuous and D_{cat} discrete categorical components such that $\mathbf{x}$ can take the following dimensions:

$$\mathbf{x}_D = (\underbrace{x_1, ..., x_{D_{\text{cont}}}}_{\text{continuous}}, \underbrace{x_{D_{\text{cont}+1}}, ..., x_{D_{\text{cont}}+D_{\text{cat}}}}_{\text{discrete}}) \tag{6.2}$$

For categorical dimensions, we can apply a multinomial logistic distribution of k possible alternatives represented by the vector $x_{D_{\text{cat}}} = (x_{D_{\text{cat}_1}}, ..., x_{D_{\text{cat}_k}})$ with $x_{D_{\text{cat}_k}} = 1$ if the k alternative for variable $x_{D_{\text{cat}}}$ is chosen. The multinomial distribution is defined by:

$$p(x_{D_{\text{cat}_k}} = 1) = \frac{e^{f_k(\mathbf{s};\theta)}}{\sum_{k'} e^{f_{k'}(\mathbf{s};\theta)}} \tag{6.3}$$

The continuous multivariate component of this vector can be modelled with a normal distribution where $x_{D_{\text{cont}}}$ is drawn from a Gaussian $\mathcal{N}(W, \Sigma^2)$. If W is not lower bound, the resulting function may generate negative values. To distinguish between positive only values in travel behaviour data, e.g. speed, distance, a *stepped sigmoidal* function can be used for generating positive real valued data:

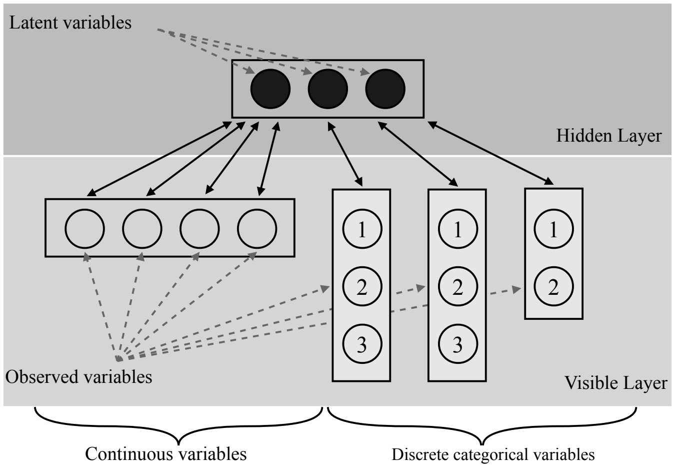


Figure 6.1: The generative bi-partite framework. The visible layer represents the input discrete and continuous data. The hidden layer represents the stochastic latent variables derived from the RBM learning algorithm. Bi-directional arrows indicate information passing in both directions. The hidden layer can be used to generate new data with similar statistical properties as the input.

$$\sum_{i=1}^{\infty} \sigma(\mathbf{s} - i) \approx \ln(1 + e^{s}) \tag{6.4}$$

The sum of $\sigma(\mathbf{s} - i)$ components represents an infinite set of binary logistic models with shared weights and fixed constant offsets. Applying this formulation increases the capacity of the logistic model to express a broader range of positive linear values but retains the same closed-form derivative and the same number of parameters. It can also be further approximated with the function $\ln(1 + e^{s})$. This method has been used in the past to develop models such as the Infinite RBM and Rate-coded RBM in generative machine learning [162, 163].

As the hidden layer represents a fully distributed mixture model, the model can be considered a mixture model with 2^{J} components with $K + J + KJ$ parameters. This representation of travel behaviour data makes it attractive because the complex correlations between observed variables and events as a result of interaction can be captured by a one or combination of multiple latent variables in the least number of additional parameters, as opposed to conventional mixed logit or latent class model.

We refer readers to Appendix B.1 for detailed mathematical explanation on variable correlations among MDC choices and conditional probability generation.

6.3.2 Variational Bayesian inference

The marginal distribution of $\mathbf{x}$ can be obtained by integrating the joint distribution: $p(\mathbf{x}) = \int_{\mathbf{s}} p(\mathbf{x}|\mathbf{s})p(\mathbf{s})d\mathbf{s}$. We are interested in obtaining the posterior belief $p(\mathbf{s}|\mathbf{x})$ that depends on the data to know how $p(\mathbf{x})$ are distributed. Assuming that the data are conditional upon the hidden variables, the maximum likelihood of the data, i.e. $\arg\max_\theta \ln p(\mathbf{x})$ may be difficult as we require the integral to be tractable. In most cases, it is difficult to compute in closed form and approximations are required. A popular method of approximating the posterior is through the MCMC algorithms [164]. However, such algorithms have a high computational cost and are more suited for well-structured small samples. By starting from some arbitrary initial distribution $q(\mathbf{s}_0)$, a stochastic transitional distribution $\mathbf{s}_t \sim q(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{x})$ is applied iteratively and the outcome $\mathbf{s}_T$ converges asymptotically to the exact posterior $p(\mathbf{s}|\mathbf{x}) \approx q(\mathbf{s}_0|\mathbf{x}) \prod_{t=1}^{T} q(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{x})$. The downside of this is that with MCMC, we do not know how many iterations are sufficient and finding the right posterior approximation may be difficult with large datasets and complex distributions.

Alternatively, it has been shown that the contrastive divergence algorithm works well on large datasets that may not be well-structured (see Section 6.2.4). Variational Bayesian inference provides a better alternative to such problems by optimizing a more straightforward function that approximates the posterior faster than conventional sampling methods. It has also been shown that for random utility based choice models, the variational error is negligible and variational inference shows asymptotic behaviour [165]. First, we posit that there is a tractable distribution $q(\mathbf{s})$ that approximates the exact posterior $p(\mathbf{s}|\mathbf{x})$. To find $q(\mathbf{s})$, we search over the set of distributions that minimizes the Kullback-Leibler (KL) divergence objective function:

$$\begin{aligned}
\arg\min \quad & D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{x})] \\
s.t. \quad & \frac{p(\mathbf{s}|\mathbf{x})}{q(\mathbf{s})} > 0, \\
& D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{x})] = 0 \iff q(\mathbf{s}) = p(\mathbf{s}|\mathbf{x})
\end{aligned} \tag{6.5}$$

where $D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{x})] = -\int_{\mathbf{s}} q(\mathbf{s}) \ln \frac{p(\mathbf{s}|\mathbf{x})}{q(\mathbf{s})} d\mathbf{s}$. If no assumptions are made, then the equation is minimized when $q(\mathbf{s}) = p(\mathbf{s}|\mathbf{x})$. The key benefit for variational Bayesian inference is that we can choose a restricted class of density distributions (partitions) for $q(\mathbf{s})$ which are simple enough for computational efficiency but flexible enough to capture the posterior distribution.

A simplifying assumption of $q(\mathbf{s})$ is that each of the partitions is independent and we can find a formula that computes $q(s_1, s_2, ..., s_J)$ using the values of the observed input data. This assumption means that the probabilities form an intersection of densities, which is an efficient way of modelling high-dimensional data while satisfying low-dimensional constraints [166]. In comparison to latent class models, this translates adding contributions in the log domain, rather than in the probability domain. The model can accommodate for a ‘no option’ edge case in the probability density where a component has zero contribution (negative infinite energy) [100]. We factorize $q(\mathbf{s})$ by taking the product over independent latent variable densities:

$$q(\mathbf{s}) = \prod_{j=1}^{J} q(s_j) \approx \prod_{j=1}^{J} p(s_j|\mathbf{x}), \quad \mathbf{s} = \{s_1, s_2, ..., s_J\} \tag{6.6}$$

Each latent variable density $p(s_j|\mathbf{x})$ is a product of expert (PoE) model. The PoE distribution produces a model with marginal independent hidden states by specifying independent expert priors [56]. If we assume each expert is a tractable distribution with a closed form solution (e.g., logit or exponential), the generative model can be computed efficiently. However, the objective function in Eq. (6.5) requires the computation of the partition function $p(\mathbf{x})$ and $\ln Z = \ln p(\mathbf{x})$. By applying a change-of-measure technique to the objective function and using Bayesian inference, we obtain:

$$\begin{aligned} D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{x})] &= \int q(\mathbf{s}) \ln q(\mathbf{s}) d\mathbf{s} - \int q(\mathbf{s}) \ln p(\mathbf{s}|\mathbf{x}) d\mathbf{s} \quad (6.7) \\ &= \int q(\mathbf{s}) \ln q(\mathbf{s}) d\mathbf{s} - \int q(\mathbf{s}) \ln p(\mathbf{x}, \mathbf{s}) d\mathbf{s} + \ln p(\mathbf{x}) \int q(\mathbf{s}) d\mathbf{s} \quad (6.8) \\ &= -\mathcal{F} + \ln p(\mathbf{x}) \quad (6.9) \end{aligned}$$

where $\int q(\mathbf{s}) d\mathbf{s} = 1$, the expectation $\langle f(x) \rangle_q = \int f(x) q(x) dx$ and $\mathcal{F}$ is the variational free energy and can be expressed as:

$$\mathcal{F} = \langle \ln p(\mathbf{x}, \mathbf{s}) \rangle_q - \langle \ln q(\mathbf{s}) \rangle_q = \langle \ln p(\mathbf{x}, \mathbf{s}) \rangle_q + \mathcal{H}[q] \tag{6.10}$$

In practice, the variational free energy is used to optimize the solution by a Gibbs sampling algorithm. The variational free energy lower bounds the partition function $\ln Z \geq \mathcal{F}$ for any $q(z)$. This bound is true since $D_{KL} \geq 0$ holds, which can be derived through Jensen's inequality [167]. We also note that $-\langle \ln q(\mathbf{s}) \rangle_q = \mathcal{H}[q]$ is the entropy of the approximating distribution q and $\langle \ln p(\mathbf{x}, \mathbf{s}) \rangle_q$ is the expected energy of the joint distribution. Therefore, minimizing the KL divergence implies maximizing the variational free energy: $\arg\min D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{x})] = \arg\max F$.

The variational free energy implies that decision makers are compelled to maximize both expected utility and information (entropy) gain. In purely econometric (utilitarian) choice models, independence of irrelevant alternatives holds and a rational decision maker would always choose the alternative with the highest utility. However, it is generally known that irrational behaviour plays a significant role in choice selection [30, 113]. In this context, incorporating KL divergence as a generalized measure of uncertainty in the model accounts for the variance over the utilities of the choices. This is also known in some literature as risk-seeking or risk-avoiding behaviour [32, 39]. Next, we develop the parameter estimation procedure for the proposed generative model.

6.3.3 Learning algorithm

Standard learning algorithms for generative models utilize a stochastic gradient descent method for optimizing the objective function. Assume that an arbitrary Gibbs-Boltzmann energy function is given by $E(\mathbf{x}, \mathbf{s}; \theta)$ where θ represents the model parameters. The energy in this context describes a value that is assigned to a state of the system. The energy curve is continuous, and the state(s) with the lowest energy corresponds to the highest probability. We relate the RBM energy function to utility, where the inverse of utility is the energy, but states have both independent observed and latent variables. Then the generative model is a joint probability distribution over the observed and latent variables in a configuration given by the Boltzmann probability distribution:

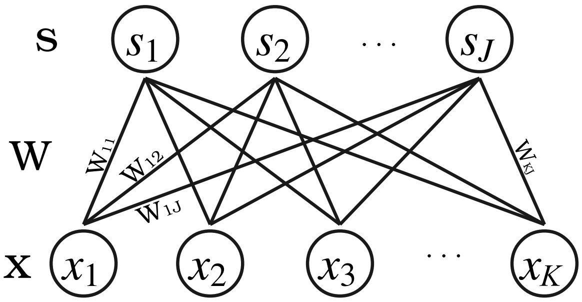


Figure 6.2: Graphical illustration of an RBM with connections represented by $\mathbf{W}$ between hidden $\mathbf{s} = (s_1, s_2, ..., s_J)$ and visible layer $\mathbf{x} = (x_1, x_2, .., x_K)$. The connections are undirected, and the weights are the strength of the connections. Weight updates are performed bi-directionally in every batch step.

$$p(\mathbf{x}, \mathbf{s}) = \frac{e^{-E(\mathbf{x},\mathbf{s})}}{\sum_{x,s} e^{-E(\mathbf{x},\mathbf{s})}} \tag{6.11}$$

Illustrated in Fig. 6.2, we express the RBM as a bipartite graph of a visible and a hidden layer connected by a weight matrix. These are considered as unsupervised learning methods, whereby there are no category labels or output values for model optimization. RBM models are stochastic rather than deterministic: latent variables are randomly sampled according to a joint distribution specified by the model. Let $\mathbf{W} \in \mathbb{R}^{K \times J}$ be the weight matrix connecting the hidden layer $\mathbf{s} = (s_1, s_2, ..., s_J)$ and visible layer $\mathbf{x} = (x_1, x_2, .., x_K)$. The magnitude of $\mathbf{W}$ measures the strength of the connection between two units. The interaction between the two layers defines the energy function:

$$E(\mathbf{x}, \mathbf{s}) = -\mathbf{x}^\top \mathbf{W} \mathbf{s} - \mathbf{b}^\top \mathbf{x} - \mathbf{c}^\top \mathbf{s} \tag{6.12}$$

The marginal of the visible layer is $p(\mathbf{x}) = \sum_s p(\mathbf{x}, \mathbf{s})$. $\mathbf{b}$ and $\mathbf{c}$ are the parameters for the visible and hidden layer respectively towards the joint distribution density (Appendix B.1). The variational free energy objective is the lower bound approximation to the marginal log likelihood since the KL divergence is always positive:

$$\ln p(\mathbf{x}) \geq F + D_{KL}[q(\mathbf{s})||p(\mathbf{s}|\mathbf{x})] \tag{6.13}$$

The objective is to compute $q(\mathbf{s})$ that maximizes F with respect to q, yielding the variational density as an approximate to the posterior $q(\mathbf{s}) \approx p(\mathbf{s}|\mathbf{x})$:

$$\begin{aligned} & q(\mathbf{s}) := \max_{q(\mathbf{s})} F \iff \\ & \nabla_{q(\mathbf{s};\theta)} F = 0, \text{ for any } \theta^* \in \arg\max_{x \in \mathbb{D}} \ln p(\mathbf{x}; \theta^*) \end{aligned} \tag{6.14}$$

at which point, (-F) is equivalent to the log likelihood $\ln p(\mathbf{x})$ of the RBM model. Using the definition of thermodynamic free energy in bounded rational decision making process $F = U - T\mathcal{H}$, where U is the expected utility (energy), T is the temperature constant ($T = 1$) and $\mathcal{H}$ is the entropy [168, 41], we obtain the following derivative of $(-F)$:

$$\begin{aligned} \nabla_{q(\mathbf{s};\theta)}(-F) &= \nabla_{q(\mathbf{s};\theta)} \ln \sum_s p(\mathbf{x}, \mathbf{s}; \theta) && (6.15) \\ &= \nabla_{q(\mathbf{s};\theta)} \ln \frac{\sum_s e^{-E(\mathbf{x},\mathbf{s};\theta)}}{\sum_{x,s} e^{-E(\mathbf{x},\mathbf{s}\theta)}} && (6.16) \\ &= \nabla_{q(\mathbf{s};\theta)} \Big(\underbrace{\ln \sum_s e^{-E(\mathbf{x},\mathbf{s};\theta)}}_{\text{utility } U} - \underbrace{\ln \sum_{x,s} e^{-E(\mathbf{x},\mathbf{s};\theta)}}_{\text{entropy } \mathcal{H}} \Big) && (6.17) \end{aligned}$$

To find $q(\mathbf{s})$, we take the derivative of *negative* F w.r.t. the RBM parameters $\theta = (\mathbf{W}, \mathbf{b}, \mathbf{c})$. We arrive at the stochastic gradient descent (SGD) learning update on the negative variational energy objective function:

$$\theta_t \leftarrow \theta_{t-1} - \frac{1}{A_\tau} \eta \sum_{A_\tau} \nabla_{q(\mathbf{s};\theta)} (-\mathcal{F})_{A_\tau} \qquad \forall A_\tau \in \mathcal{D}, \tau = 1, ...T \tag{6.18}$$

where η is the learning rate and derivative of $(-\mathcal{F})$ represents the convergence step towards a locally optimal variational approximation: $q(\mathbf{s}) = \prod_j q(s_j)$. Depending on the form of the distribution (we used binary logistic distribution, i.e. $q(s_j) = (1 + e^{-Wx-c})^{-1}$ in our example), the optimization can be solved analytically. Since the derivative can be inferred as the average energy change over A_τ, the gradient yields the difference between the expected utility $\ln \sum_s e^{-E(\mathbf{x},\mathbf{s};\theta)} = U$ and the entropy $\ln \sum_{x,s} e^{-E(\mathbf{x},\mathbf{s}\theta)} = \mathcal{H}$ gradients.

The utility $\ln \sum_s e^{-E(\mathbf{x},\mathbf{s};\theta)}$ is expressed as the energy Eq. (6.12) over all possible configurations of s. We can associate the first and second term as the expected energy value obtained from the conditional $p(\mathbf{s}|\mathbf{x})$ and joint distribution $p(\mathbf{x},\mathbf{s})$ respectively (Appendix A.4), using the gradient $\mathbf{W}$ as an example:

$$\begin{aligned} \nabla_{q(\mathbf{s};\mathbf{W})} U &= \langle \mathbf{xs} \rangle_{p(\mathbf{s}|\mathbf{x})} \\ \nabla_{q(\mathbf{s};\mathbf{W})} \mathcal{H} &= \langle \mathbf{xs} \rangle_{p(\mathbf{x},\mathbf{s})} \end{aligned} \tag{6.19}$$

The contrastive divergence (CD) algorithm takes a point estimate from one or more Gibbs sampling steps drawn to approximate the equilibrium energy:

$$\begin{aligned} \langle \mathbf{xs} \rangle_{p(\mathbf{s}|\mathbf{x})} &\sim \langle \mathbf{x}^0 \mathbf{s}^0 \rangle \\ \langle \mathbf{xs} \rangle_{p(\mathbf{x},\mathbf{s})} &\sim \langle \mathbf{x}^t \mathbf{s}^t \rangle \end{aligned} \tag{6.20}$$

where $\langle \mathbf{x}^t \mathbf{s}^t \rangle$ is the average over product of the generated input samples multiplied and the generated latent variable samples from the Gibbs chain and $\langle \mathbf{x}^0 \mathbf{s}^0 \rangle$ is the initial sample (see pseudocode in Appendix A.4). Typically, a 1-step Gibbs sample chain $(CD_N; N = 1)$ is sufficient for fast learning gradient estimation [102]. The gradient estimators can be used to minimize the objective function using a suitable learning rate. The free energy is representative of the relative fit of the generative model with respect to the data distribution. If the gap between the utility and entropy increases, it represents model overfitting [100].

6.4 Case Study: Montreal Trajet Dataset

In this section, we describe the generative modelling process focusing on the data generation and inferring from the estimated latent variable component. We describe how we pre-process the data and how the learning algorithm is used to optimize and generate statistically similar synthetic data for comparison.

6.4.1 Case study

We evaluate our proposed methodology on a trip trajectory dataset: the MTL Trajet GPS data from the Greater Montréal Region [124]. The open dataset consists of a total of 293,330 trip observations. The data were collected from respondents living in the Greater Montréal region (Fig. 6.3). Trip trajectories were recorded in an application that runs in the background of participants' smartphones. Participants were also prompted to report their travel mode and trip characteristics in addition to the GPS trajectories. We consider the following revealed characteristics for our model: mode choice, trip purpose, trip distance, origin-destination point and departure/arrival time.

6.4.2 Data pre-processing

The GPS data from the mobile app are sampled at 4 to 10 second intervals. From multiple users' GPS trajectories, we detect points at the origin and destination and matched to one of 34 boroughs of Greater Montréal. First, we verified each observation D_n contains valid trajectory points, and we removed all corrupted data points outside the city boundary. Next, we calculated the total trip distance between the start point and end point by the total sum of all point-to-point raw GPS coordinates. Alternatively, open map data can also be used to find map matched travel distances. Travel time was calculated by taking the time differential between the first and last coordinates. Input time data x_t were reparameterized into linear cyclic encoding features using sin/cosine transform: $x_{t_{sin}} = \sin(2\pi x_t)$, $x_{t_{cos}} = \cos(2\pi x_t)$. Cyclic encoding features allow time data to be represented consistently and can be used as linear input. Continuous data (trip distance, trip time) were normalized to unit variance. Discrete categorical data (mode choice, purpose, origin-destination) were encoded as one-of-k vector: $x_{mode} = 2 \in \mathbb{R}^4 \to x_{mode_v} = \{0, 0, 1, 0\}$. We selected trip candidates with a simple constraint of minimum 10-minute travel time and users had reported their travel mode and trip purpose. Once all the valid trip observations were selected, we used this processed dataset for training and validation. Since our methodology is an unsupervised learning algorithm, we did not consider any output data for cross-validation. For model validation and data generation, we used the full training dataset to compare our results.

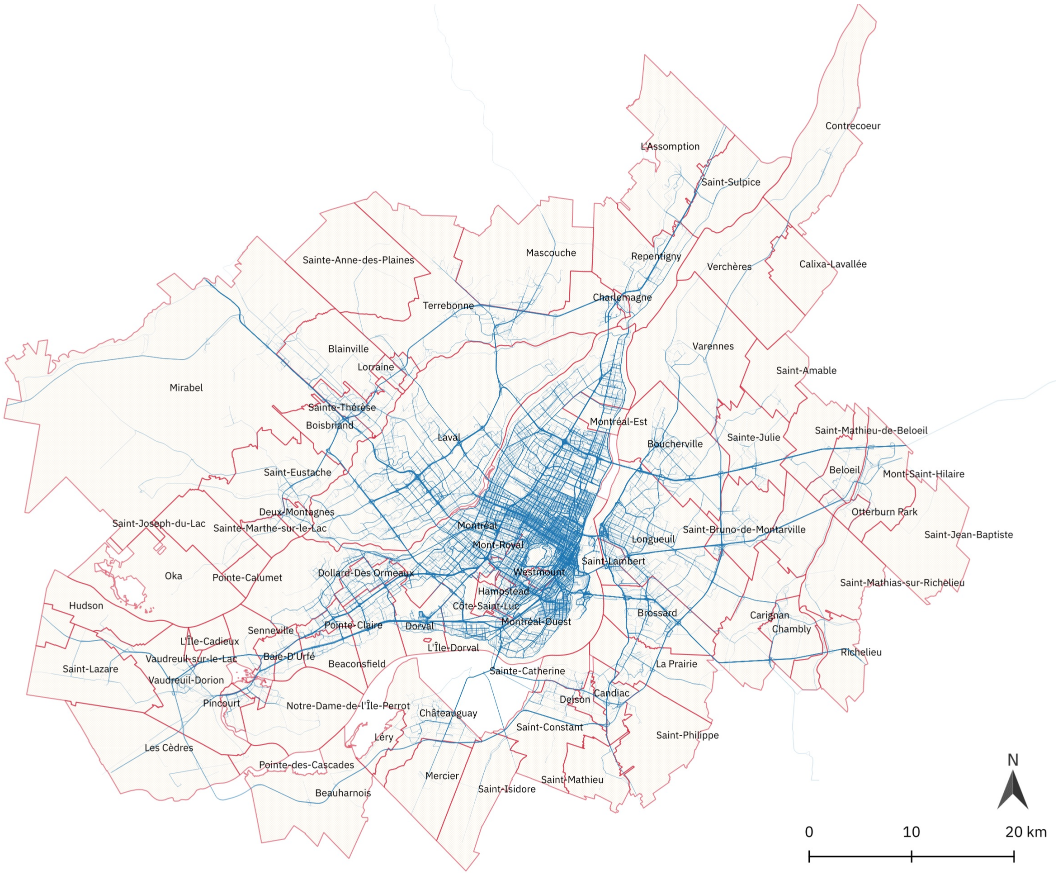


Figure 6.3: Visualization of the trip trajectories across the Greater Montréal Region from the pilot study.

6.4.3 Training

We used a standard batch stochastic gradient descent learning algorithm for model estimation implemented using Theano Python machine learning libraries[1]. The model parameters were updated after every batch sample. We bootstrap iterations over mini-batches of observations, randomly sampled from the input data $\mathbf{x}_D$. We defined a decaying learning rate η, starting at $10e^{-2}$ at the first iteration and decay at a rate of 0.1% per batch. The objective function is calculated as the difference in the first-order derivative of expected free energy of the input and the sampled

[1] Theano Python library: http://github.com/Theano/Theano

data. In this paper, we did not explore other novelty regularization methods such as dropout or model ensemble, which could be future work for implementation.

6.4.4 Data validation

A typical estimation procedure would be to divide the data into training and validation sets. The full dataset consists of a labelled subset ($N = 58,034$) and an unlabelled subset ($N = 235,296$). The labelled subset consists of trips with full information availability and the unlabelled subset consist of trips with missing variables. Using the labelled subset, we divide training and validation in a 70:30 ratio for model benchmarking against a comparable feedforward neural network (NN). Accuracy validation is often misleading when a model is tested on a biased or imbalanced dataset. In our case study, the dataset we obtained cannot fully represent the whole population of the area due to physical limitations, e.g. availability of all transport modes, the use of smartphone applications, etc. Therefore, we address this shortcoming by implementing a likelihood validation as a proxy to determine the model predictive accuracy.

For evaluating generative model performance, we simulate the model on the unlabelled data (with missing data) and compare the statistical properties of the generated output against the labelled dataset. This is equivalent to testing the 'unsupervised' learning performance. The accuracy of these predictive probability distributions depends on whether the 'correct' priors lead to reasonable predictive accuracy. We estimated a series of models with different latent variable sizes and reported the model fit. Ideally, increasing the size of latent variables would improve the fit for each variable dimension if input variables are assumed to be independent and identically distributed. Our proposed method of variational Bayesian inference satisfies the likelihood principle where the inference depends on the distribution of the data [63].

Next, we analyze the mean and variance effects of latent variables on the generative model. Deep learning NN models are prone to overfitting when model parameters have a large bias and low variance, which results in poor predictors beyond the training data. Such networks are naturally viewed as black-box functions and challenging to analyze. By contrast, variational Bayesian inference allows the analyst to infer how flexible a model is warranted by the data [169]. Likewise, when parameters have low bias and high variance, it will result in low statistical confidence and makes the model harder to fit the data. The consequence of the parameter uncertainty is that we cannot differentiate between good predictors and sampling error in our model. Well-calibrated models should have flexibility in accounting for sampling error as well as robustness to avoid misspecification.

We also performed analysis over the elasticity of the choice probabilities w.r.t. to changes in the independent variables. In our result, we show the direct elasticity of *mode choice* with respect to *travel distance*. which can be calculated directly from the optimization step using the Jacobian function.

6.4.5 Benchmarking

We benchmark our results against a comparable single hidden layer feedforward NN with the number number of latent variables and mode choice as the output. This is equivalent to partitioning the generative model into a hidden layer $h(x)$ and computing the conditional output of the mode choice probability $f(h(x))$. The NN hidden and output layer equations are given by the following:

$$h(x) = (1 + e^{-(-\mathbf{xW}-\mathbf{c})})^{-1} = \sigma(-\mathbf{xW} - \mathbf{c}) \tag{6.21}$$

$$f(h(x)) = \frac{e^{W_k h(x) + b_k}}{\sum_{k'} e^{W_{k'} h(x) + b_{k'}}} \tag{6.22}$$

The first difference between this approach and a discriminative-generative modelling approach (Appendix B.1) is the direct estimation of the likelihood given the inputs, rather than an auxiliary step in generating latent variable samples, then using these samples to generate the output mode choice data. The second difference in the feedforward NN model is that the individual's observed utility is drawn from a non-linear deterministic component. In contrast, the observed utility in the generative model is drawn from a linearly separable entropy term as described in Appendix B.1.

We benchmark our model against the NN and compared the normalized log likelihood shown in Figs. 6.4 to 6.6. As expected, the *training* curves converge asymptotically, which indicates that the gradient estimation reached a local optimum. The *validation* curves show the model fit on the validation data subset.

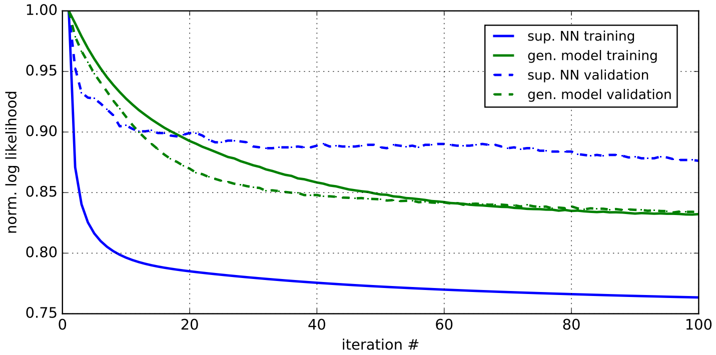


Figure 6.4: Training and validation likelihood curve (H=5)

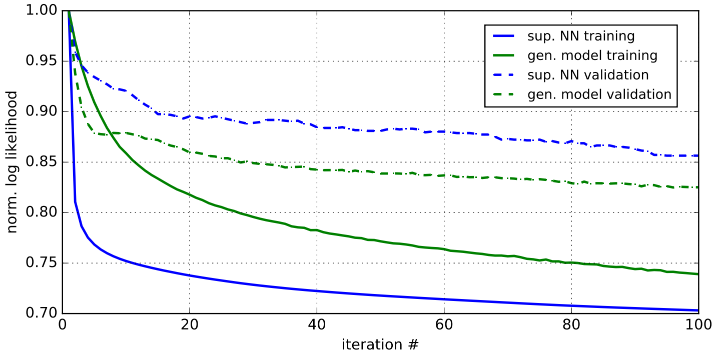


Figure 6.5: Training and validation likelihood curve (H=25)

While the supervised NN training curve shows better model fit than the generative model in all 3 model instances (which is normal as the supervised NN model optimizes the model likelihood), it also points to higher overfitting shown by the more significant disparity between the training and validation likelihood. Even though the generative model produces a weaker model fit on the training curve, the validation curve is better than the supervised NN and less likely to be overfitting.

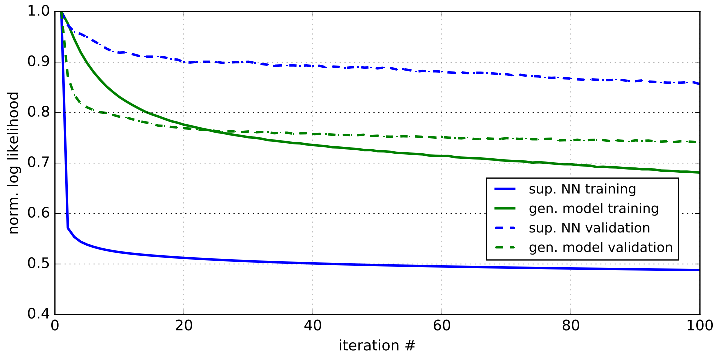


Figure 6.6: Training and validation likelihood curve (H=100)

6.5 Results

6.5.1 Latent constructs parameter analysis

For model analysis, we trained the model on a single layer fully connected network with $H = 5$, 25 and 100 latent variables for 100 iterations over the dataset using our generative learning algorithm. To verify if generative modelling provides better model generalization, we plot the distribution of the model parameters connecting the latent variables and mode choice data and compute the magnitude of mean and variance of the weight matrix. The results are shown in Fig. 6.7. We observed that with 5 latent variables H_5, the model parameters do not fit well to the input data. The mean and variance parameter values are $H_5 = \mathcal{N}(5.237, 9.33)$. Increasing the number of latent variables substantially improves the model, where the mean and variance converges to zero mean and unit variance at H_{25} and increasing to H_{100} improve the model further. The estimated mean and variance are $H_{25} = \mathcal{N}(0.45, 7.603)$ and $H_{100} = \mathcal{N}(-0.102, 1.624)$.

One reason for the improvement is the concept of sparse overcomplete representation of weights and activations in deep learning. It has been shown that sparsity can be an important factor in explaining and capturing the variations in the data by reducing the number of activated parameters [52]. The parameter distribution indicates the mean activation and utilization rate of latent variables. When estimated parameters have low mean and variance, we can determine which subset of latent

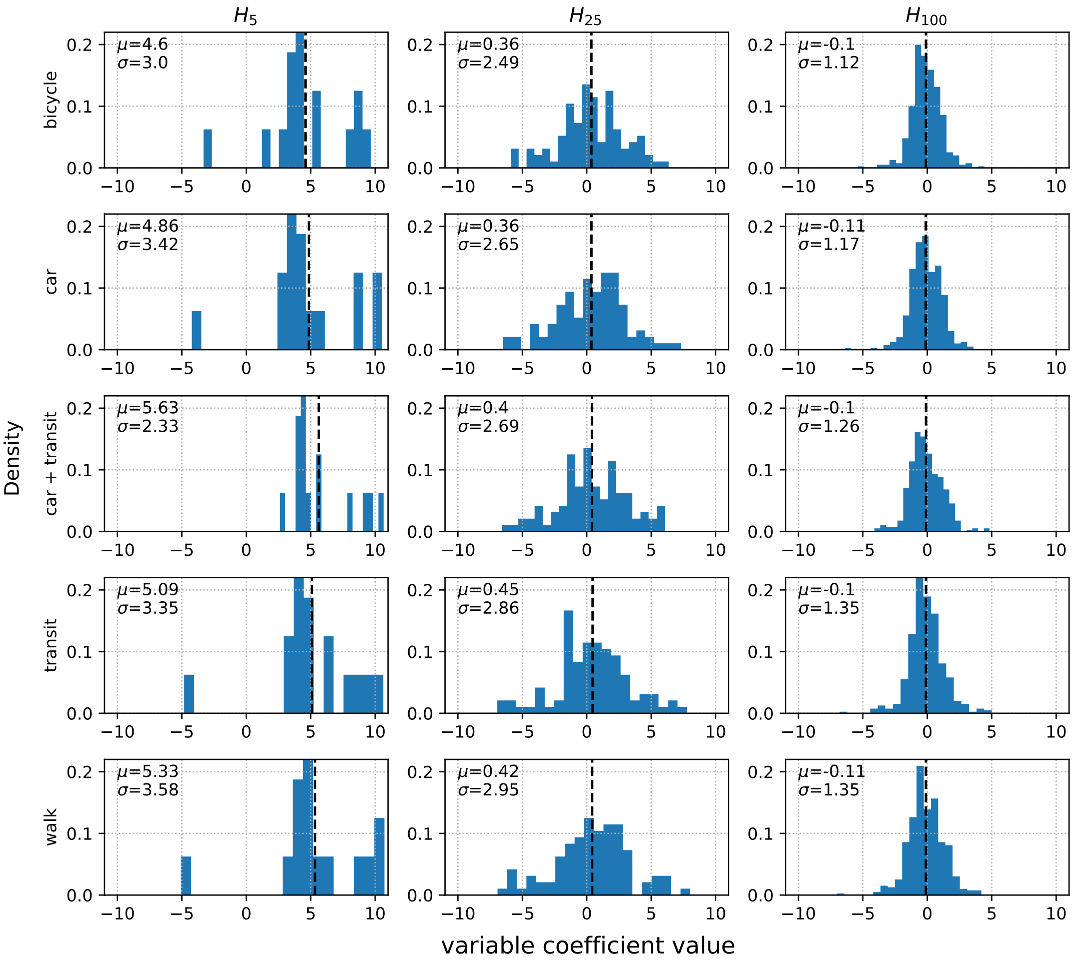


Figure 6.7: Histogram of parameter value distribution by mode choice and number of latent variables. Vertical dashed line represent distribution mean.

variables are ‘activated’ and which are ‘inhibited’ – when parameters are zero or near zero, their contributions in the log domain is negative across the distribution. This result suggests that the latent variables provide a strong indication of model identifiability by producing sparse parameter representation.

6.5.2 Interpretation of latent constructs

In conventional choice modelling latent variable interpretation are justified by explicitly introducing indicator variables to correspond to different latent variable states [72]. For example, a useful indicator might allocate attitudinal variables: safety, comfortability or eco-friendliness [35]. However, for this method of latent variable classification to be effective, the indicators must be free from outliers and assumed to be uncorrelated to other events or error terms. In generative modelling, we can emulate the travel decision process as a learning algorithm, to provide an underlying explanation for sensory information inputs. We can think of the latent variables as an interpretation of the observed data (e.g. how individuals consider their distance, mode choice, location choice, etc., simultaneously).

The earlier models that did not explicitly use psychometric indicators to capture the latent variables were alternative specific only and did not vary over the individual market segments [38]. Our proposed model has no restrictions on latent variables being alternative specific. It imposes a generalized logical structure (probabilistic graphical model) and accounts for uncertainty and variance from observed data (explanatory variables and observed choices) through Bayesian probability theory. However, it is flexible enough that it can also be formulated as a model structure that only captures alternative specific variations by removing the connections between the latent constructs and the explanatory variables and any other setup is also possible. We modify the estimation step so that the connection strength conditions on the between the observed and latent variables. This representation can be more useful when attitudinal variables are not IID and have a high correlation with each other and thus, require knowledge of the underlying distribution. The latent variable parameters define an entropy term which can be interpreted as a structure for capturing unobserved correlations between variables. This can be framed as an entropy generalization to the linear MNL model structure where the latent variables form an error correction function. This structure also represents a simplistic model of how decisions are simulated not just by random utility, but also the dynamical effects of information availability, habits and perceptions. This reflects the role and importance of neural networks in capturing realistic behavioural responses beyond direct cause-and-effect maximum utility-based observations.

6.5.3 Model elasticity

In econometric analysis, elasticity is an important metric to measure the effects of changes in the value of the explanatory variables (e.g. cost, distance) on choice probabilities. This test is an indicator of the variation in elasticities of the unobserved heterogeneity of the population w.r.t to the choice decision. In the context of generative models, we can use the Jacobian determinant to compute the elasticity (Appendix B.2). The direct elasticities of *mode choice* with respect to *travel distance* are shown in Fig. 6.8. As expected, the elasticities are all negative. Distance is most strongly correlated with driving with an average elasticity of -0.635 and a standard deviation of 0.535. Since walking trips are for relatively short distances, our results show that walking mode choice is inelastic w.r.t. distance with an average and standard deviation of -0.084 and 0.336 respectively. Moreover, the average elasticity for driving mode is larger than transit or driving+transit mode, meaning that as the distance increases, the probability of driving decreases faster. This is verified by the collected GPS data, where individuals used public transit (commuter trains) more for long-distance trips – especially for commuting.

Elasticity and latent variable parameter inspection can be regarded as measures of posterior and prior heterogeneity, respectively. It puts forward a plausible model that assumes the generative model emulates an individual's prior information about the choice with respect to the latent variables, and the elasticity of demand for each mode choice in this context quantifies how much the individual would react to the decision having formed some prior beliefs generated from the model.

6.5.4 Data simulation

Generative models can be used to represent the underlying distribution of the data; thus, they can be beneficial in forecasting. We used the trained model to generate samples from $p(\mathbf{x})$, which have similar statistical properties as the input data. First, we consider a single observation where we observed only part of the data vector and the other part of the vector is unknown. The unknown vector can be a single or multi-variable vector. We denote this as $\mathbf{x}_D = (x_1, ..., x_{D-1}, x_{D_{unk}})$, where $x_{D_{unk}}$ is fixed as the unknown variable. The objective is to predict $x_{D_{unk}}$ using the remainder of the 'known' data vector by sampling from the distribution $p(x_{D_{unk}}|x_1, ..., x_{D-1})$.

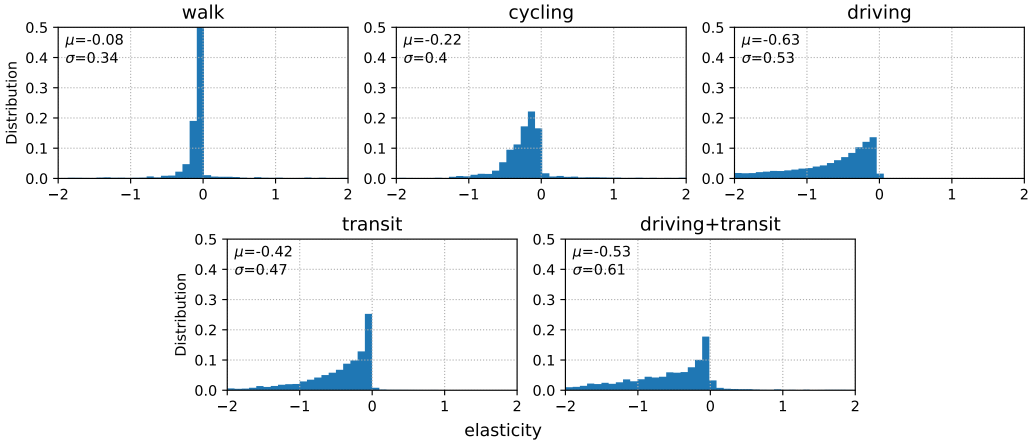


Figure 6.8: Distance elasticities on mode choice

We should note that conventional likelihood tests are not suitable in this instance because the outputs of the generative model are stochastic data-driven probability distributions, rather than a deterministic probability distribution of a dependent variable. Appendix B.1 describes how these distributions can be computed.

Next, we clamp the known variables to the input data and then sample the states of the hidden layer. We use the sampled states of the hidden layer to generate the remaining state of the unknown variable, completing a full Gibbs sampling step. This process is not limited to a single unknown variable. If more unknown variables are used, it reduces the ability of the model to capture the data representation (this is analogous to adding noise to the input, we can fix $x_{D_{unk}} = 0$ for the variables we want to forecast). Therefore, the robustness of the model can be quantified by the information loss when adding noise to the input and how well it recovers this lost information.

We show that as we increase the model capacity and complexity, the model can generate synthetic samples that emulate the original distribution of the data. The output generated samples are evaluated against the inputs, and we compute the R^2 distribution fit. The results are shown in Fig. 6.9. In particular, we observed that discrete categorical variables (trip mode and trip purpose) are easily represented with small model capacity ($R^2 > 0.937$), but continuous variables, e.g. trip distance ($H_{25} : R^2 = 0.759$) and time ($H_{100} : R^2 = 0.639$) require more latent variables to

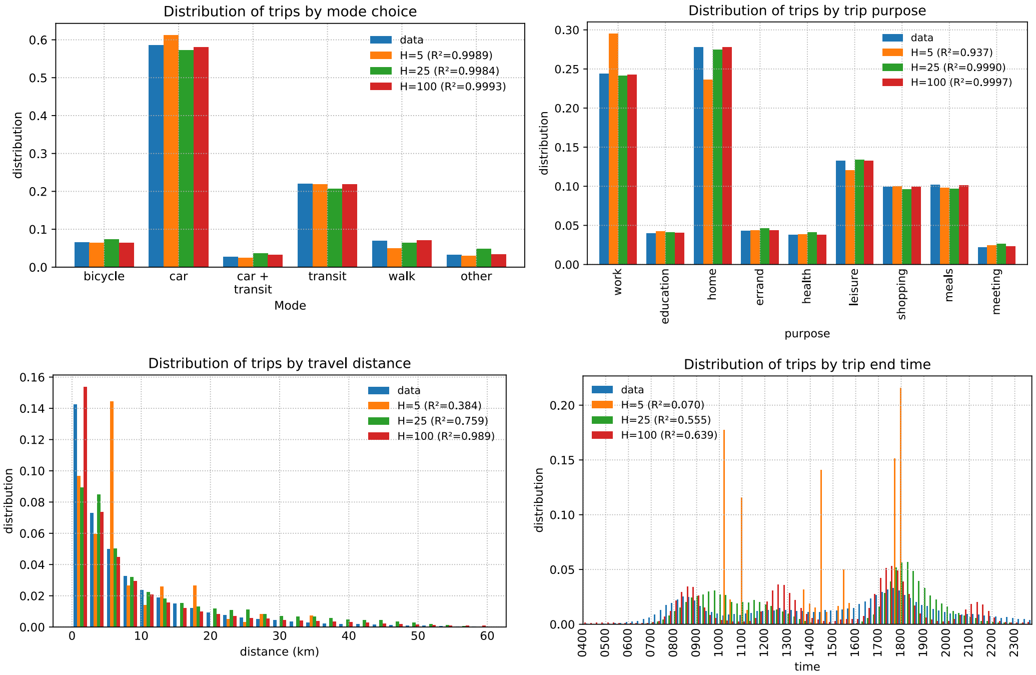


Figure 6.9: Discrete and continuous data generated from the model.

capture the underlying distribution accurately.

The generative model is also able to learn the multi-modal cyclical nature of trip arrivals, which is significantly challenging for a standard logit model to estimate. In our simulation, the latent variables can generate a statistical distribution with modes in the morning and evening peak hours as well as a smaller peak around mid-afternoon. Surprisingly, even with no indication of how the distribution is supposed to be or using any pre-defined measurement indicators, the generative model can capture the underlying properties of a complex distribution, demonstrating a level of understanding of the semantic variations in the dataset. Finally, multiple discrete-continuous data can be generated from the conditional probability densities – an example would be combining mode choice with distance, as shown in Fig. 6.10.

We analyze the model results using the Kruskal-Wallis statistical test and report the k-th central moments up to k=4 shown in Tables 6.2 to 6.5. The results indicate that the samples generated from $H = 100$ are similar to the original data based on the

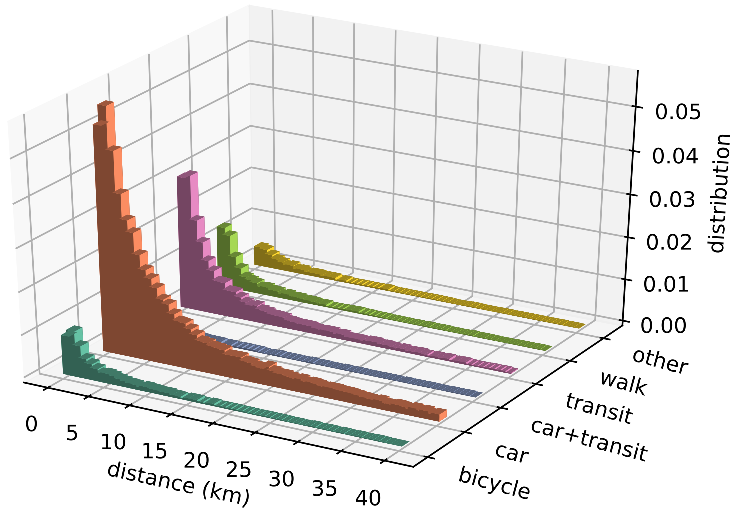


Figure 6.10: Data generation for a joint MDC output

test statistics and the p-values. The k-th moments of the generated data converge to the k-th moments of the original data indicating that the generative model is well representative of the underlying behaviour. We also report the variable pair correlation shown in Table 6.6. The correlation pair also confirms that the higher-order generative models can emulate the distribution of the original data with high accuracy.

The sample statistics of the generative model are shown in Table 6.7. For each of these models, we report the two-way likelihood Chi-square test, mean squared distance and p-value of the generative models on discrete variables mode and trip purpose. For trip distance and trip arrival counts, we report the RMSE of the samples against the original data. RMSE for trip distance are ($H_5 : 4.171, H_{25} : 4.721$ and $H_{100} : 1.852$). RMSE for trip arrival counts are ($H_5 : 128.7, H_{25} : 26.8$ and $H_{100} : 23.5$) Model significance is computed as the p-values for χ^2 at 5% sample size. The analysis shows that the models with a larger number of latent variables are more consistent and statistically significant ($H = 100 : \chi^2 = 2.3308, p \leq 0.115$), even though R^2 values indicate that the generative model well represents the data.

Table 6.2: k-th central moment of the original data samples

Moment	mode	purpose	distance	time
1	0	0	0	0
2	1.56	6.28	109.62	2.518e+05
3	1.78	5.32	2961.60	9.848e+07
4	6.80	70.23	1.62e+05	1.923e+11
Kruskal-Wallis	-	-	-	-
p-value	-	-	-	-

As shown in the results, the generative model can represent both discrete and continuous data types simultaneously. This relates to the sparsity concept mentioned in the previous subsection – the model is robust to corrupted data and information retrieval from truncated data is possible. This experiment shows how we can use a generative model for model prediction and forecasting for various input variable types. In terms of latent variables, this is not an exhaustive analysis, and we can increase the size of latent variables to increase the representational power, but with diminishing returns. However, it has been shown that for a neural network with T input dimensions and $T-1$ latent variables, it is globally stable and satisfies the necessary conditions for optimality with no local minima in the error surface [170].

While these tests may serve as useful benchmarks, we note that the choice of latent variable size is still arbitrary and dependent on many various factors including data size, number of variables, complexity and amount of 'missing' information in the data collection. However, as we have shown that in general, generative modelling may serve as a useful additional tool for travel behaviour analysts to estimate MDC data using variational Bayesian inference techniques. Collectively our analysis of the generative modelling provides empirical support that unobserved information in the data plays an important role in the model estimation, which has previously shown to be plausible in discrete choice theory [30].

Table 6.3: k-th central moment of the generated data samples (H=5)

Moment	mode	purpose	distance	time
1	0	0	0	0
2	1.23	6.21	47.35	4.931e+06
3	0.95	12.13	642.88	4.497e+10
4	2.70	86.86	1.598e+04	4.702e+14
Kruskal-Wallis	513.94	1165.31	510.98	85.13
p-value	≤ 0.05	≤ 0.05	≤ 0.05	≤ 0.05

Table 6.4: k-th central moment of the generated data samples (H=25)

Moment	mode	purpose	distance	time
1	0	0	0	0
2	1.62	6.23	152.98	7.058e+05
3	1.84	5.25	2765.74	9.100e+08
4	7.25	68.85	1.019e+05	2.435e+12
Kruskal-Wallis	1.01	1.31	326.58	6.74
p-value	≤ 0.315	≤ 0.253	≤ 0.05	≤ 0.05

Table 6.5: k-th central moment of the generated data samples (H=100)

Moment	mode	purpose	distance	time
1	0	0	0	0
2	1.58	6.28	131.94	5.983e+05
3	1.81	5.33	3280.01	7.743e+08
4	6.94	70.02	1.356e+05	1.767e+12
Kruskal-Wallis	0.393	0.092	20.88	1.63
p-value	≤ 0.53	≤ 0.76	≤ 0.05	≤ 0.2

Table 6.6: Variable pair correlation

Variable pair	Original data	H=5	H=25	H=100
mode-purpose	-0.0961	-0.1149	-0.1002	-0.0954
mode-distance	-0.1884	-0.4439	-0.2269	-0.1919
mode-time	0.0349	-0.0244	0.0667	0.0382
purpose-distance	-0.1396	-0.2846	-0.1549	-0.1453
purpose-time	-0.1039	-0.3866	-0.1504	-0.1052
distance-time	0.4777	0.8247	0.5715	0.4907
mean difference	-	0.07	-0.004	-0.001

Table 6.7: 5% sample size analysis of the generative choice model outputs

Model	χ^2	dist.	R^2	p-value
		mode choice		
H=5	23.658	4.6791	0.9989	1.0
H=25	33.8029	6.1074	0.9984	1.0
H=100	2.3308	1.5569	0.9993	$p \leq 0.115$
		trip purpose		
H=5	55.535	7.5163	0.937	1.0
H=25	6.041	2.499	0.9994	$p \leq 0.36$
H=100	0.334	0.582	0.9997	$p \leq 0.01$

6.6 Conclusion

As the use of machine learning models and algorithms becomes increasingly significant and essential in travel behaviour research, more emphasis has to be put on model interpretability rather than pure forecast accuracy. Our work focuses on methods and tools for analyzing and interpreting complex travel behaviour data and the estimation of MDC models. Notably, we introduced a generative machine learning approach for analyzing and estimating large scale MDC travel behaviour model that uses variational Bayesian inference for model training. We proposed an RBM-based learning algorithm to model behaviour data, accounting for heterogeneity and variable correlations. The estimation results indicated that while supervised learning performed better on the training set, it performed worse than the generative model on validation. This makes generative models less prone to overfitting and more relevant in the context of accurate travel behaviour modelling and forecasting. We showed how the proposed model could be used to compute the conditional probability distribution of the dependent variables as well as the associated elasticities. This concept can be expressed in terms of information gain to quantify their contribution to utilitarian behaviour by measuring the KL divergence between the observed and simulated data.

For the case study, we implemented the algorithm on an open large travel behaviour dataset. We were able to estimate model parameters to fit the underlying distribution of the data while retaining identifiability and sparsity. The sparse distribution of parameters enabled the generative model to capture the correlation effects between input variables for both discrete and continuous variable types. To ensure that latent variables capture data heterogeneity, we perform simulation tests, and we showed that the generative model was able to recover the original data with a similar statistical distribution. For model interpretability, we show that elasticities can be obtained for economic analysis. Also, we report model statistics, correlation and sample analysis, which indicate that the shape of the distribution converges to the original data samples.

We note that the additional complexity due to an increase in the number of hidden units is minimal when using the first-order stochastic gradient optimization. The increased size of model parameters did not constitute a significant increase in computational time, due to fast and efficient tensor-based operations using Theano machine learning libraries. We also observed that increasing the number of hidden units of an order of magnitude does not correspond to the same increase in computational time. On the other hand, increasing the number of observations in the training data will be the main bottleneck in each iteration of the dataset. In future work, we can also look at regularization, e.g. L1/L2 penalty or drop out techniques, to reduce overfitting from the effects of using latent variables.

With the development of ubiquitous data collection methods for travel behaviour analysis, there are potentials for generative machine learning to be used for modelling these large multidimensional travel information datasets. Overall, integrating probabilistic variational Bayesian inference methods can improve model tractability and interpretability. Adopting this framework into dynamic road pricing, route choice recommendation and traffic network simulation are some interesting applications for future work.

Part III

Concluding Remarks

Chapter 7

Conclusion

This dissertation presents a series of interconnected research works that address behaviour modelling under the new data-driven paradigm of machine learning and generative modelling. This work is highly applicable to emerging use of machine learning in econometric welfare analysis, understanding value of time through deep learning models and interpretability of neural networks. The first work concerns the use of machine learning in discrete choice modelling. A generative model-based behaviour analysis concept is put forward as a new class of tractable and fully interpretable form of choice model. Chapter 2 presents the fundamental theory behind the use of generative modelling in discrete choice analysis and how it relates back to the original Boltzmann distribution derived by McFadden [43, 41]. The chapter highlights the strengths and weaknesses of current and proposed methods of travel behaviour modelling. Chapter 3 introduces the Conditional Restricted Boltzmann Machine (RBM) algorithm and its use on a stated as well as revealed preference travel survey with measurement indicators.

The second work concerns the methodological development of behaviour models by framing the choice process as an optimization algorithm that conforms to behaviour theory in literature, in particular, the notion of information processing in decision making and disentangling unobserved behaviour through modelling the higher-order correlation. Chapter 5 analyzed the properties of generative modelling, estimation process and statistical analysis of information processing costs.

The third work focuses on the applications to discrete choice modelling and travel behaviour research. Chapter 6 addresses the complexity of modelling choice behaviour with multiple, varied outputs that have hidden correlation which is not specified in the observed data. Furthermore, the experiments shown in Chapter 4 highlights how behaviour modelling is not restricted to subjective psychometric indicators, and machine learning can be exploited to uncover processes and contexts in choice modelling.

7.1 Research Contributions

This dissertation put together four research articles, each making a different contribution to the state-of-the-art in behaviour modelling. The following summarizes the main results from the articles based chapters of Part II:

Modelling Latent Travel Behaviour Characteristics: A Generative Machine Learning Approach

The third chapter of this dissertation opens with the concept of RBM generative modelling algorithm for modelling latent travel behaviour characteristics. We developed a novel framework by estimating the joint distribution of the choice variable and auxiliary information. This approach is useful when indicators are available, but assumptions may be hard to verify as we are unsure about the interactions of the latent variable generating process. This method is the first fully developed solution in the context of travel behaviour modelling research. Comparable with previously developed ICLV methods in terms of model fit, the generative learning algorithm does not require any additional parameters over the ICLV model with an equivalent number of latent variables. The estimation process is performed using stochastic gradient descent with a contrastive divergence optimizer. The experiment showed that it is possible to derive latent variables analytically when indicators are available, but assumptions are difficult to verify. It is most effective when the explanatory data are noisy and have little direct correlation with the choice output, while there is a strong underlying correlation between unobserved variables. It shows that stated indicators may not be reflective of behavioural attitudes and may be highly influenced by survey conditions, geographic region and socio-demographic variables.

Discriminative Conditional Restricted Boltzmann Machines for Discrete Choice and Latent Variable Models

The fourth chapter begins with a hypothesis that latent effects can be obtained not only from attitudinal questions but also from the data distribution of the observed preference. We consider that not all datasets can capture psychometric indicator data. Furthermore, when psychometric data are available, they may be subjective, and they may change over time. We provide an adaptation of an existing class of generative models known as Conditional Restricted Boltzmann Machines that are used to learn a latent representation from the data. We performed a bootstrap cross-validation and model selection for estimation consistency to account for overfitting in our estimation. We found that the generative learning algorithm has substantial improvement in model fit without the use of subjective measurement indicators. Thus increasing model accuracy in predicting choice behaviour.

Information Processing Constraints in Travel Behaviour Modelling: A Generative Learning Approach

In the fifth chapter, we posit that generative models can be used to emulate information processing and learning-based decision-making behaviour in a discrete choice model. Our methodology consists of applying an entropy error component using latent constructs in a generative model based on rational inattention behaviour and information theory since decision-making rules are not always consistent with rational behaviour and this methodology accounts for behaviour learning and unobserved decision motivations. We optimize the model parameters using a Kullback–Leibler (KL) divergence minimization between observed and simulated data.

The study demonstrated the theoretical and practical properties of the generative model in three ways: First, the latent constructs provide error correction for information heterogeneity in the utility specification, which allows the model to simulate decision making with information processing constraints. Next, the experiment uncovered a strong correlation with rational inattention behaviour theory through a generative model. It shows that individuals may ignore certain explanatory variables and rely on prior (latent) information in their decision-making process. We looked at the changes to the econometric parameters and showed how latent constructs could change the model specification. Finally, estimating maximum entropy reveals that decision uncertainty is reduced by incorporating latent constructs accounting for information processing costs. The impact of this study is that we can use noisy data that contains latent behavioural information to improve our model analytical properties.

A Bi-partite Generative Model Framework for Analyzing and Simulating Large Scale Multiple Discrete-Continuous Travel Behaviour Data

In the sixth chapter, a generative model solution is developed to model discrete-continuous travel behaviour from a large scale dataset. This study connects to the broader scope of the transport an mobility market by enabling the use of interpretable machine learning in demand forecasting systems. With the development of ubiquitous data collection methods for travel behaviour analysis, there is a benefit for generative machine learning to be used for modelling these large multidimensional travel information datasets.

This methodological framework allows for the interpretation of complex travel data by analyzing its underlying data structure and unobserved heterogeneity. The incorporation of an RBM-based learning algorithm to model travel behaviour data captures the underlying correlation effectively through latent variables. In addition to that, the learning algorithm is robust to data overfitting it also allows for the interpretation of latent variables and elasticities, which enables it to be work with discrete choice analysis. The model is analyzed through the inspection of model sparsity and correlation effects of model parameters. The results showed that it retains parameter stability which offers better interpretability over MLP models. For model interpretability, elasticities can be obtained for the econometric analysis, thus providing a novel way for the analyst to look into the "black-box."

7.2 Limitations

Every modelling approach has its own set of advantages and drawbacks, and it is the analyst's job to be aware of them and choose the right model in the right situation. A linear in utility, a discrete choice model can focus on highly informative explanatory variables, but it may not be able to model complex and noisy data easily and may lead to model misspecification when the explanatory variables are noisy and highly heterogeneous. Generative models, on the other hand, can interpret and extract latent behavioural information from complex and noisy data without explicit labels or indicators through unsupervised learning. However, the performance of the generative model depends significantly on the size and variation of the dataset in capturing the underlying latent behaviour effectively. A way to account for these problems is by increasing the depth of the network model, but it also comes at the cost of modelling time complexity. The learning algorithm may also lead to inconsistencies in parameter estimates if the model is too small and several major drawbacks in deep learning methods need to be addressed [45]. Another limitation is the estimation and convergence of the gradient function as there may be multiple local optima points which require a good initialization point. However, initialization of parameters can also be tricky and techniques to find proper initialization points are still an open problem in research [53].

Our work has suggested that capturing unobserved heterogeneity using generative modelling may solve some of the misspecification issues in discrete choice behaviour modelling, but further investigations into other generative modelling techniques, for instance, GANs, VAE, and Autoencoders can be explored. However, these novel deep learning methods rely on an algorithmic process to efficiently train model and may not have “appropriate” behavioural interpretation. In contrast, the RBM learning algorithm is inspired by physical systems and thermodynamic processes that can be interpreted intuitively [59].

While this research can be useful in practice for travel behaviour modelling, it will require substantial effort to reconfigure to different setting when analysts seek to use this methodology on other types of data, e.g. traffic network flow, which have not been explored in this dissertation.

7.3 Future Work

The research presented in this dissertation focuses on generative machine learning for travel behaviour analysis. It can be used as a tool to complement discrete choice modelling. However, there are no specific applications that necessitate the use of generative modelling if conventional MNL or even regression is sufficient enough. Generative modelling also can work on a range of data-driven transport services such as driving sensors or traffic network signals. More research has to be done on these applications to determine if it will benefit from generative modelling.

The concepts and ideas that have been introduced in this dissertation can be transferred to discrete choice modelling to allow for analyzing a greater range of unobserved heterogeneity, specifically from noisy data. Beyond the comparisons between discrete choice models and machine learning, the analysis of information heterogeneity in behaviour modelling indicates that there is a significant common structure between the contrasting approaches and combining them is a possibility.

To allow more practical application of these new algorithms, for instance, the use in an activity-based travel demand modelling system, several issues need to be resolved. There needs to be a standard or structured way of handling data so that the generative models can be reused on different datasets, in order to simplify benchmarking and comparative testing. The multiple discrete-continuous processes need further investigation into model identification as the model parameters are confounded in the higher-order representation. Another possible approach could be to look into economic and welfare analysis, for example, value of time factor, willingness-to-pay considerations, and elasticites of these measures using the proposed analytical form in this dissertation. It can also be a good idea to see how stacking multiple layers will affect the stability of the model parameters as it is often shown that it may lead to overfitting. This dissertation does not cover the use or comparison with more recent advanced supervised learning methods such as Residual Network models that enable very deep model learning and representation which could be used as a way to increase the depth of the choice model as described in [171].

Based on the findings, the following suggestions are proposed that would improve or integrate deep learning models and algorithms into discrete choice analysis. First, an investigation into machine learning optimization methods, namely, the stochastic gradient descent algorithm and its purpose in, and relation to behavioural choice theory. Second, another interesting perspective would be to look at how the estimation of mixed logit models can be improved with variational Bayesian inference and generative learning techniques that have been recently explored and could be expanded upon [172]. While variational Bayesian inference addresses the shortcomings of MCMC methods, generative modelling methods can also be used to extend Mixed Logit through estimating a flexible underlying latent behaviour model without relying on pre-specified distributions.

The final area for future work would be to adapt these machine learning algorithms and generative models for real-world deployment, for example, in autonomous transit, CAVs and MaaS systems. These can be implemented offline in data analytics or online through a cloud service and integrating networked sensors from traffic and vehicles. Such applications to emerging and innovative transport systems would have significant economic and social benefits, such as the ability to rapidly adapt to changes in travel behaviour to reduce rush-hour delay and increase the throughput capacity of public transit in urban cities.

Appendices

This chapter is organized as follows:

Appendix A briefly summarizes the several fundamentals for the generative modelling and its formulation in discrete choice analysis. This would provide some background on advanced machine learning from the perspective of a choice modeller and is divided into 3 sections. In A.1, the main differences between a Bayesian approach to estimation and conventional log-likelihood approach in discrete choice analysis is compared. In A.2, the basic concepts of stochastic gradient descent used in machine learning is described and an explanation why it enables "data-driven" modelling. In A.3, the concept of undirected graphical model and the role of interactions between observed and unobserved heterogeneity in travel behaviour analysis is described to represent choice behaviour process.

Appendix B provides additional details on the mathematical models and equations used in the application of multiple discrete-continuous models in Chapter 6. Explanation is given as to how the formulation and joint distribution is established. An example of calculating model elasticity is shown in B.2.

Appendix C provides a short snippet of the machine learning code (in Python) used to train the generative model. A complete listing can be found in the Github repository (`https://github.com/mwong009/genome`). The model files in this dissertation uses Theano deep learning libraries for constructing the computational graph essential for calculating the gradient functions.

Appendix D provides a list of variables and a brief description for each of the dataset used in this dissertation.

Appendix A

Advanced machine learning principles for discrete choice modelling

A.1 Bayesian statistical approach

The discrete modelling perspective is based on estimating a set of β values where the true β values are *fixed* but *unobserved*, while the estimates $\hat{\beta}$ is a random variable as a function dependent on the input data set. The variance of the estimated $\hat{\beta}$ is addressed using standard error and t-test statistics. The t-test is an assessment of how much confidence the estimation value reflects the observed data with random sampling. In a Bayesian approach, a probability $p(\beta)$ is used to define the degree of uncertainty of the estimate in different states of the system. E.g., at different values of β, what are the probability that it will occur given the data? Since the data set is observed and assuming the data set represents every possible variation of the population, the true β value is *random*, and we measure the degree of uncertainty between our estimates and the random value of β.

The random variable β can be represented using the prior distribution $p(\beta)$ using a general but broad distribution (e.g. uniform) with high entropy. Often it is also easier to use some simpler distribution, i.e. normal or log-normal before observing the data. After that, samples are drawn from our data sequentially or randomly to update the belief of our initial distribution by using the Bayes rule:

$$\text{posterior} = \frac{\text{likelihood} \times p(\beta)}{\text{marginal}}$$

Relative to the conventional log-likelihood estimation of discrete choice models, the Bayesian statistical approach uses the full distribution over β rather than a point estimate $\hat{\beta}$. The distribution is "learned" by decreasing the density over β where the data does not generate possible values and increasing the density where the data generates possible values of β from the prior. The Bayesian approach is also behaviourally similar to how perception and attitudes (latent constructs) affect the choice outcome. One can associate the Bayesian prior to these latent constructs as the source of human subjective behaviour influencing travel behaviour. In Chapter 5, the uncertainty of the estimates is discussed by analyzing the entropy of the estimates. E.g. have the parameter estimates moved far enough away from a high entropy distribution? This approach is simple to justify, but as with the t-test approach, it is still somewhat ad hoc. The Bayesian approach has its drawbacks and can be computationally intensive when the marginal distribution is complex. What is relatively difficult is expressing the initial Bayesian prior as close as possible to our naïve beliefs of the underlying behaviour. The prior should reflect the perceptions and attitudes, but in practice, a uniform distribution is used with high entropy and perturb the density slowly until it reaches convergence. For real-valued β parameters, it is common to use the **Kullback-Leibler divergence** (KL divergence) as an objective function to minimize the uncertainty between the prior and posterior distribution. In Chapters 3 and 4, the estimation process of an RBM model is shown using the KL divergence function.

Representation learning is one of the key themes of generative modelling [166]. The Bayesian statistical approach operationalizes representation learning in a behaviourally plausible way that reflects choice behaviour. The distinction between generative modelling and choice modelling is that it does not require any formal definition of a dependent variable in the objective function. Indeed, generative modelling is often referred to as an attempt to extract information from a distribution, density estimation, denoising data, clustering or simulation-based prediction. A classic example is Principal Component Analysis (PCA) that finds a representation that preserves much of the original data as possible. However, there are also multiple ways where the distribution is defined: low-dimension transformation (e.g. PCA), sparse distribution (RBM, VAE, Autoencoders) or independent representations [45]. PCA is a low-dimension transformation where a large number of parameters are reduced to remove redundancies for model estimation. Sparse representation is more commonly used in machine learning for its properties which allow the information to be distributed over the latent variable space. Independent representation is much more difficult as it tries to transform the original data into the latent variable space where each latent vector is statistically independent, which might be implausible for typical travel behaviour with unobserved heterogeneity.

A.2 Optimization algorithm

The next aspect of machine learning for discrete choice analysis is the type of optimization algorithm used. Almost all modern machine learning methods and deep learning models use stochastic gradient descent (SGD) or some variant of it. SGD is an extension of the gradient descent algorithm where the gradient of the objective function is used to update the model parameters [45]. Fundamentally, the gradient of the objective function w.r.t. to the dataset is the average gradient, and by the law of large numbers, this can be approximated by a small subsample with a similar value. Rather than iterating through the entire dataset to find the average gradient, the gradient can be approximated by a *mini-batch* sample randomly drawn from the dataset. Thus, the efficiency of SGD increases by the number of observations (large datasets) – as the dataset grows, the number of iterations remains small relative to the total data size. Increasing the size of the dataset allows a much broader range of variability to be modelled without a similar increase in estimation time, hence the justification for "data-driven" modelling. The gradient in SGD is defined as such:

$$\Delta_{\beta}^{(t)} \leftarrow \Delta_{\beta}^{(t-1)} - \eta \frac{1}{m} \frac{\partial L}{\partial \beta} \sum_{i=1}^{m} L(\beta)$$

where m is the batch size, t is the update step and L is the objective function and η is the adjustable parameter updating rate. The SGD algorithm does not guarantee optimal model estimates, but due to the properties of the batch update steps, a value close to optimal can be reached very fast for the estimation of large datasets with a large number of parameters.

A.3 Undirected graph model for behavioural representation

In generative modelling, a way of describing the interaction between observed and unobserved behaviour is by using an undirected graphical model or also known as a Markov Network. It encodes the idea that two variables with strong correlation have a higher magnitude undirected link between them. As opposed to directed graphical models where the causality is one direction, e.g. price to profit relation, undirected graphical models are much more applicable to behavioural analysis and travel based applications. For example, the relationship between travel time and mode choice. Travel time directly influences travel mode choice, but the reverse can also be possible: travel mode choice can increase the overall usage of a particular mode, leading to congestion and increase travel time. Hence, undirected graphical models do not differentiate between the direction of the causality. The relationship is modelled between observed and unobserved behaviour and attitudes as an event in an undirected graphical model. In Chapter 6, an undirected model (RBM) is used to model the relationship between the observed data and the underlying latent behaviour. In Chapters 3 to 5, an undirected graphical model is combined with directional interaction between the observed explanatory variable and the dependent choice variable. Additionally, in Chapter 3, this to model is further extended to the measurement indicators from an SP/RP survey experiment. Since the indicators can be either influencing the behaviour or vice versa, using an undirected graphical model is a plausible way of handling these interactions.

A way of defining an undirected graphical model is to use an energy function E, where:

$$p(\beta) = e^{-E(\beta)}$$

This would ensure that the probability density remains positive for any state of β. An example of this function is the Boltzmann distribution. While Boltzmann machine is historically used to define models with and without latent variables, in deep learning, a Boltzmann machine is exclusively used for models with latent variables [59]. The Boltzmann distribution as a product of models is expressed as follows:

$$p(\beta : \{\beta_1, \beta_2, ..., \beta_K\}) = \prod_{k=1}^{K} e^{-E(\beta_k)}$$

The most straightforward approach to sampling and training a Boltzmann machine is by using Gibbs sampling, alternating between random variables. Due to the separation properties of the graphical model, the draw is conditioned only on the remaining variables that are connected to the targeted random variable (see Section 4.3 for details).

Energy function

The energy function E derives from a quantitative property in thermodynamics that describes the potential energy in a body. The same energy metric can also be used to define behaviour – the formalization of the concept of utility suggests the two domains are similar mathematically and the differences lie primarily in their assignment of importance. It has been shown that behaviour theory can be represented by additive utility gains and an entropic cost involved in processing information [41]. In statistical physics, the Boltzmann distribution satisfies the free energy model $F = U - TS$, which represents the tradeoff between *utility* U and the *entropy* cost S and the temperature term T is the scale term in the discrete choice model. These two terms have been related to utility and information processing cost in rational inattention decision making [30, 58]. Sections 5.2 and 5.3 outlines the key properties of the energy function and how it applies to travel behaviour analysis. This principle can be seen in two ways: First, a minimum relative entropy when the expected utility is fixed. This provides a principle for modelling under uncertainty were utility deviations do not influence the choice decisions. Second, a maximum utility principle when the entropy is fixed. This interpretation leads to a conventional rational decision-making process where the information processing cost is assumed to be homogeneous.

Contrastive divergence (CD) provides a way to estimate the gradient of the energy function. CD gives an approximate idea of the gradient field, and by taking small steps in the direction of the steepest gradient, the optimization approaches a local minimum [102]. Even though it may not be possible to evaluate the energy function (in some cases, the partition function is intractable), CD estimates the *gradient function* given a set of model vectors. The algorithm shown in Appendix A.4 establishes a primary training step on a generative model by computing the gradient term of the variational free energy function and using a stochastic gradient descent step on the model parameters.

A.4 Model training algorithm

Algorithm 2: RBM learning algorithm for generative modelling using N-step Gibbs sample chain (CD_N)

```
Input  : RBM data sample $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, batch sample
         $A_i \subset \mathcal{D}, i = 1, ..., d$, learning rate $\eta$, iteration steps $T$
Output: gradient approximation $\theta = (\mathbf{W}, \mathbf{c}, \mathbf{b})$.

init: $\theta = 0$, $\tau = 1$;
forall $A_\tau \in \mathcal{D}, \tau = 1, ..., T$ do
    forall $(\mathbf{x}_n) \in A_\tau$ do
        for $t = 1$ to $N$ do
            $CD_t$: iterate over Gibbs chain

            positive phase
            $\mathbf{x}^0 \leftarrow \mathbf{x}_n$
            $\mathbf{s}^0 \sim \prod_{j=1}^{H} p(s_j|\mathbf{x}^0)$

            negative phase
            $\mathbf{x}^t \sim \prod_{i=1}^{I} p(x_i|\mathbf{s}^0)$
            $\mathbf{s}^t \sim \prod_{j=1}^{H} p(s_j|\mathbf{x}^t)$
        end
    end
    % Variational free energy term
    $\nabla_{q(\mathbf{s};\theta)}(-\mathcal{F})_{A_\tau} \approx (\langle \mathbf{x}^t \mathbf{s}^t \rangle - \langle \mathbf{x}^0 \mathbf{s}^0 \rangle)$
    % parameter update step
    for $\theta \in \theta$ do
        $\theta_{\tau+1} \leftarrow \theta_\tau - \eta \nabla_{q(\mathbf{s};\theta)}(-\mathcal{F})_{A_\tau}$;
    end
end
```

Appendix B

Mathematical Models and Equations

The contents of this appendix chapter are associated with Chapter 6.

B.1 Multiple discrete and continuous conditional probability generation

Additional details is provided here on how conditional probability generations are formulated. To apply a generative model to travel behaviour choice problems, first, specify the distribution of our required output variable set conditioned on the other variables. Then, this can be further extended to other distributions, not just multinomial and Gaussians, e.g. unimodal distribution for ordinal data [173].

Example 1. *Given an assumption that the simplest possible example consisting of two observed variables* $[x_1, y_1]$ *connected by a single hidden unit* s_j *(Fig. B.1 (a)). The generative model captures the joint distribution of* x, y *and* s *expressed as* $P(x, y, s) = \frac{1}{Z} e^{-E(x,y,s)}$ *as derived from Eq. (6.11). The functional form that represents the variables under an RBM energy model is* $E(x, y, s) = -\sum_{s_j} x_1 W_{1,j} s_j - \sum_{s_j} y_1 W_{1,j} s_j - b_1 x_1 - \sum_{s_j} c_j s_j - d_2 y_1$ *and the conditional probability of* y *given* x *assuming* y *is a multinomial output:*

$$P(y_1|x_1) = \frac{e^{-F(x_1,y_1)}}{\sum_{y_1'} e^{-F(x_1,y_1')}}$$

where its variational free energy $F(x_1, y_1)$ is calculated as:

$$\begin{aligned} F(x_1,y_1) = -\ln \sum_{s_j\in\{0,1\}} e^{-E(x_1,y_1,s_j)} &= -b_1x_1 - d_2y_1 - \ln \sum_{s_j\in\{0,1\}} e^{-s(x_1W_{1,j}+y_1W_{1,j}+c_j)} \\ &= -b_1x_1 - d_2y_1 - \ln(1+e^{-x_1W_{1,j}-y_1W_{1,j}-c_j}) \end{aligned}$$

The first term b_1x_1 is the 'error-corrected' utility component in the model. However, unlike in conventional DCM, b_1 is the beta of variable x_1 contribution to the full *joint* probability $P(x, y, s)$. The second term can be interpreted as the 'alternative specific constant' (ASC) of y_1. For instance, if y_1 is a 3-alternative discrete variable $y_1 : \{y_1^1, y_1^2, y_1^3\}$, then d_2 is a 3-dimension vector representing the ASCs. In the conditional probability $P(y_1|x_1)$, if $y_1^1 = 1$ and 0 otherwise, then the error-corrected utility of alternative y_1^1 is:

$$\begin{aligned} F(x_1,y_1^1) = -\Big(& b_1x_1 + d_2^1\cdot(y_1^1=1) + d_2^2\cdot(y_1^2=0) + d_2^3\cdot(y_1^3=0) \\ &+ \ln(1+e^{-x_1W_{1,j}-y_1W_{1,j}-c_j})\Big) \\ = -\Big(& b_1x_1 + d_2^1 + \underbrace{\ln(1+e^{-x_1W_{1,j}-y_1W_{1,j}-c_j})}_{\text{single correction term}}\Big) \end{aligned}$$

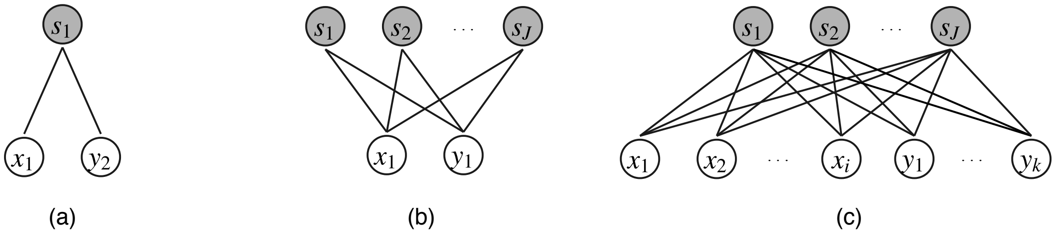


Figure B.1: Generating various multiple discrete-continuous outputs using generative models. (a) Example 1, (b) Example 2, (c) Example 3.

If the weights connections to the hidden units are reduced to zero, i.e. $W_1 = 0, W_2 = 0$ and $c_j = 0$, then the model collapses into a standard MNL. For such a configuration:

$$F(x_1, y_1^1) = -\Big(b_1 x_1 + d_2^1 + \ln(1 + e^0)\Big) = -\underbrace{(b_1 x_1 + d_2^1)}_{\text{MNL utility}}$$

Example 2. *Consider the same example above, but expanding to j hidden units $s_1, ..., s_j$. With j hidden units, additive terms are added to the error-corrected utility (Fig. B.1 (b)):*

$$F(x_1, y_1^1) = -\Big(b_1 x_1 + d_2^1 + \underbrace{\sum_j \ln(1 + e^{-x_1 W_{1,j} - y_1 W_{1,j} - c_j})}_{\textit{multiple correction terms}}\Big)$$

Example 3. *Lastly, consider multiple inputs and multiple discrete-continuous outputs: The joint probability expands to i input variables $x_1, ...x_i$ (Fig. B.1 (c)). Likewise, the error-corrected utility can be derived as:*

$$F(x_1, ...x_i, y_1, ..., y_k) = -\Big(\sum_i b_i x_i + d_2^1 + \sum_j \ln(1 + e^{-\sum_i x_i W_{i,j} - \sum_k y_k W_{k,j} - c_j})\Big)$$

where y_k can be any discrete or continuous variable. These examples above can be extended to multiple discrete-continuous joint distributions, where each y_k component is a Product of Experts model:

$$P(y_1, ...y_k | x_1, ...x_i) = \prod_k P(y_k | x_1, ...x_i)$$

also note here that the correction terms are *marginal decreasing* functions for $x_i \to \infty$ and $W_{i,j} > 0$,

$$\lim_{x_i, ...x_i \to \infty} F(x_1, ...x_i, y_1, ..., y_k) = -(\sum_i b_i x_i + d_2^1) \quad \implies W_{i,j} > 0$$

For continuous variable output with positive only values, the stepped sigmoidal function is applied to $F(x_1, ...x_i, y_{\text{cont}})$:

$$f(y_{\text{cont}}|x_1,...x_i) = \ln(1 + e^{-F(x_1,...x_i,y_1,...,y_k)})$$

If the output is linear with range $-\infty < y_1 < \infty$, then the output would be the variational free energy:

$$f(y_{\text{linear}}|x_1,...x_i) = F(x_1,...x_i,y_1,...,y_k) = \sum_i \frac{(b_i - x_i)^2}{2} - d_2 - \sum_j \ln(1 + e^{-\sum_i x_i W_{i,j} - \sum_k y_k W_{k,j} - c_j})$$

For discrete choice outputs, a similar method described in Example 1 and 2 is used:

$$P(y_{\text{discrete}}|x_1,...x_i) = \frac{e^{-F(x_1,...x,y_{\text{discrete}})}}{\sum_{y'_{\text{discrete}}} e^{-F(x_1,...x,y'_{\text{discrete}})}}$$

B.2 Model elasticity

Analyzing model elasticity is a way to test functional dependency among a set of observations n on the conditional probability distribution. For these tests, we exploit the computational graph used to calculate the backpropagation algorithm in stochastic gradient descent by substituting the final partial derivative $\partial\hat{h}/\partial\mathbf{W}$ with $\partial\hat{h}/\partial x_n$. The advantage of using a Jacobian is that it allows discrimination of linear and non-linear dependence in the model. A Jacobian matrix is generated for each example of the conditional output on the set of inputs and estimates the density of elasticities across the data points.

Lemma 1. *Given the conditional probability function $p_n(\mathbf{x})$, its elasticity ε is defined as:*

$$\varepsilon = \frac{Jp_n(\mathbf{x})\mathbf{x}_n}{p_n(\mathbf{x})} = \frac{\partial p_n(\mathbf{x})}{\partial x_n} \cdot \frac{\mathbf{x}_n}{p_n(\mathbf{x})}$$

The Jacobian matrix $Jp_n(\mathbf{x})$ of each observation n, for each fixed input vector $\mathbf{x}$ is defined as the backpropagation derivative w.r.t p_n:

$$Let \quad p_n(\mathbf{x}) = g(\mathbf{W}^{(1)} \cdot h(\mathbf{W}^{(0)} \cdot \mathbf{x}_n)), \quad then$$

$$Jp_n(\mathbf{x}) = \frac{\partial p_n(\mathbf{x})}{\partial x_n} = \underbrace{\frac{\partial p_n(\mathbf{x})}{\partial \hat{h}} \cdot \frac{\partial \hat{h}}{\partial x_n}}_{backpropagation\ terms} = \begin{bmatrix} \frac{\partial p(\mathbf{x})_1}{\partial \hat{h}_1} & \cdots & \frac{\partial p(\mathbf{x})_1}{\partial \hat{h}_s} \\ \vdots & \ddots & \vdots \\ \frac{\partial p(\mathbf{x})_k}{\partial \hat{h}_1} & \cdots & \frac{\partial p(\mathbf{x})_k}{\partial \hat{h}_s} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial \hat{h}_1}{\partial x_1} & \cdots & \frac{\partial \hat{h}_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \hat{h}_s}{\partial x_1} & \cdots & \frac{\hat{h}_s}{\partial x_n} \end{bmatrix}$$

B.3 Example of derivation of the joint probability

For the case study example described in Section 6.4, the derivations of the joint probability is shown, energy function and estimation steps:

Energy function

The model is defined by the following energy function:

$$E(\mathbf{x}, \mathbf{s}, y) = -\Big(b^y y + \sum_{m,j} x_m W_{m,j} s_j + \sum_j y W_j s_j + \sum_m b^x_m x_m + \sum_j c_j s_j\Big) \qquad \text{(B.1)}$$

where $s_j \in \{0,1\}^J$ and $x_m, y \in \mathbb{R}^{\mathcal{D}}$ are referred to as latent and observed variables respectively in the RBM model. m is the number of explanatory variables used, x_m are the explanatory variables (time, speed, distance, location etc.) and y is the mode choice dependent variable vector. For 5 latent variables, set $j = 5$. The weight parameters are $\theta = (W, b^x, b^y, c)$.

Joint probability

The joint probability distribution of the observed and hidden variables follows the Boltzmann distribution $p(\mathbf{x}, \mathbf{s}, y) = e^{-E(\mathbf{x},\mathbf{s},y)}/Z$, where Z is a normalization factor such that $0 < p(\mathbf{x}, \mathbf{s}, y) \leq 1$.

Model estimation process

Given a sufficient number of latent variables, the RBM model parameters can be tuned such that the negative free energy is minimized:

$$(-F) \quad = \ln \sum_{s_j \in \{0,1\}} \left(e^{b^y y + \sum_{m,j} x_m W_{m,j} s_j + \sum_j y W_j s_j + \sum_m b^x_m x_m + \sum_j c_j s_j} \right) - \ln Z \tag{B.2}$$

The training task is then to minimize the negative free energy term by taking the derivative w.r.t. the model parameters and updating the parameters using an SGD training process. The gradient update step is as follows:

1. Draw Gibbs samples $\mathbf{x}^0, \mathbf{y}^0, \mathbf{s}^0, ..., \mathbf{x}^t, \mathbf{y}^t, \mathbf{s}^t$ for t steps (Appendix A.4).
2. Compute $\frac{\partial}{\partial\theta}\Big(- F(\mathbf{x}^0, \mathbf{y}^0, \mathbf{s}^0, ..., \mathbf{x}^t, \mathbf{y}^t, \mathbf{s}^t) \Big)$and update model parameters θ.

The model can be used to predict new observations by "clamping" the explanatory variables, generate latent variable samples from it, and using the generated samples to compute the choice probability $P(y|\mathbf{x}, \mathbf{s})$.

Appendix C

Code listing for generative model estimation

Code listing provided in this dissertation is released as open-source code under the MIT License.

Readers are recommended to be familiar with Python (3.5+) and the Theano deep learning library.

A complete software package (under development) for writing discrete choice models with machine learning elements, specifically generative models, can be downloaded from `https://github.com/litrans/genome`. The software supports both conventional discrete choice models (logit, etc.) and neural network-based models (RBM) using stochastic gradient descent as the primary model optimizer.

RBM model class definition

The model class definition initializes the framework of the model and constructs parameter variables, function to add latent and observed variables and the Gibbs sampling phases. The following code is used in Chapters 4 to 6.

```
class RBM(Network):
    ''' define the RBM toplevel '''
    def __init__(self, name, hyperparameters=OrderedDict()):
        Network.__init__(self, name, hyperparameters)
        # tensors
        self.label = []                             # label tensors
        self.input, self.output = [], []            # list of tensors
        self.in_dtype, self.out_dtype = [], []      # list of str dtypes

        # parameters
        self.W_params, self.B_params = [], []               # xWh, hWy, xBy params
        self.V_params, self.U_params = [], []               # xWh, hWy params
        self.hbias, self.vbias, self.cbias = [], [], []     # bias

        # flattened version
        self.W_params_f, self.B_params_f = [], []   # xWh, hWy, xBy params
        self.V_params_f, self.U_params_f = [], []   # xWh, hWy params
        self.vbias_f, self.cbias_f = [], []

        # sigmas
        self.vsigmas, self.csigmas = [], []
        self.vsigmas_f = []

        # parameter masks
        self.B_params_m, self.U_params_m = [], []   # list of the Uh mask
        self.cbias_m = []

    def add_latent(self, name='hbias'):
        """
        add_latent func

        Parameters
        ----------
        name : ‘str‘, optional
            Name of hidden node e.g. ‘’hbias’‘
        shp_hidden : ‘tuple‘, optional
            Size of the hidden units

        Updates
        -------
        self.hbias[] : sequence of ‘theano.shared()‘
        self.params[name] : OrderedDict of ‘theano.shared()‘
        """
        try:
            shp_hidden = self.hyperparameters['n_hidden']
        except KeyError as e:
            print("hidden unit shape not defined!")

        if name in self.model_values.keys():
            value = self.model_values[name]
        else:
            value = np.random.normal(0., 0.01, shp_hidden)
        hbias = theano.shared(value, name=name)

        self.hbias.append(hbias)
        self.params[name] = hbias
        self.params_shp[name] = shp_hidden

    def add_node(self, var_dtype, name, shp_visible):
        """
        add_node func

        Parameters
        ----------
        var_dtype : `str`
            Type of variables e.g. 'binary', 'category',
            see hyperparameters for more information
        name : `str`
            Name of visible node e.g. 'age'
        shp_visible : `tuple`, optional
            Size of the visible units

        Updates
        -------
        self.input[] : sequence of `T.tensor3()`\n
        self.in_dtype[] : sequence of `str`\n
        self.W_params[] : sequence of `theano.shared()`\n
        self.vbias[] : sequence of `theano.shared()`\n
        self.params['x_'+name] : OrderedDict of `theano.shared()`\n
        """
        self.hyperparameters['shapes'][name] = shp_visible

        shp_hidden = self.hyperparameters['n_hidden']
        size = shp_visible + shp_hidden

        # create the tensor symbolic variables
        tsr_variable = init_tensor(shp_visible, name)

        # create the tensor shared variables
        if 'W_' + name in self.model_values.keys():
            value = self.model_values['W_'+name]
        else:
            value = np.random.normal(0., 0.01, np.prod(size))

        W_f = theano.shared(value.astype(DTYPE_FLOATX), 'W_'+name)
        W = T.reshape(W_f, size)
        print('W', name, size)

        if 'vbias_' + name in self.model_values.keys():
            value = self.model_values['vbias_'+name]
        else:
            value = np.random.normal(0, 0.01, np.prod(shp_visible))

        vbias_f = theano.shared(value.astype(DTYPE_FLOATX), 'vbias_'+name)
        vbias = T.reshape(vbias_f, shp_visible)

        self.input.append(tsr_variable)
        self.in_dtype.append(var_dtype)
        self.W_params.append(W)
        self.W_params_f.append(W_f)
        self.V_params.append(W)
        self.V_params_f.append(W_f)
        self.vbias.append(vbias)
        self.vbias_f.append(vbias_f)

        self.params['W_' + name] = W_f
        self.params['vbias_' + name] = vbias_f
        self.params_shp['W_' + name] = shp_visible + shp_hidden
        self.params_shp['vbias_' + name] = shp_visible

    def add_connection_to(self, var_dtype, name, shp_output):
        """
        add_connection_to func

        Parameters
        ----------
        var_dtype : `str`
            Type of variables e.g. `'binary'`, `'category'`, see
            hyperparameters for more information
        name : `str`
            Name of visible node e.g. `'mode_prime'`
        shp_output : `tuple`, optional
            Size of the visible units

        Updates
        -------
        self.output[] : sequence of `T.matrix()`
        self.W_params[] : sequence of `theano.shared()`
        self.cbias[] : sequence of `theano.shared()`
        self.B_params[] : sequence of `theano.shared()`
        self.params[] : sequence of `theano.shared()`
        """
        self.hyperparameters['shapes'][name] = shp_output
        shp_hidden = self.hyperparameters['n_hidden']

        # create the tensor symbolic variables
        tsr_variable = init_tensor(shp_output, name)
        tsr_label = T.ivector(name + '_label')

        # create logit mask for W
        size = shp_hidden + shp_output
        mask = np.ones(size, DTYPE_FLOATX)
        mask[..., -1] = 0.
        mask = (mask.T).flatten()

        # create the tensor shared variables W
        w_name = 'W_' + name
        size = shp_output + shp_hidden
        print('W', w_name, size)
        if w_name in self.model_values.keys():
            value = self.model_values[w_name]
        else:
            value = np.random.normal(0., 0.01, np.prod(size))  # * mask

        W_f = theano.shared(value.astype(DTYPE_FLOATX), w_name)
        W_m = theano.shared(mask, w_name+'_mask')
        W = T.reshape(W_f, size)

        # create logit mask for H->cbias
        mask = np.ones(shp_output, DTYPE_FLOATX)
        mask[..., -1] = 0.
        mask = mask.flatten()

        # create the tensor shared variables cbias
        c_name = 'cbias_' + name
        print('cbias', name, shp_output)
        if c_name in self.model_values.keys():
            value = self.model_values[c_name]
        else:
            value = np.random.normal(0, 0.01, np.prod(shp_output))

        cbias_f = theano.shared(value, c_name)
        cbias_m = theano.shared(mask, c_name+'_mask')
        cbias = T.reshape(cbias_f, shp_output)

        self.output.append(tsr_variable)
        self.out_dtype.append(var_dtype)
        self.label.append(tsr_label)
        self.W_params.append(W)
        self.U_params.append(W)
        self.cbias.append(cbias)
        self.W_params_f.append(W_f)
        self.U_params_f.append(W_f)
        self.U_params_m.append(W_m)
        self.cbias_f.append(cbias_f)
        self.cbias_m.append(cbias_m)
        self.csigmas.append(None)

        self.params['W_' + name] = W_f
        self.params[c_name] = cbias_f
        self.params_shp['W_' + name] = shp_output + shp_hidden
        self.params_shp[c_name] = shp_output

        # condtional RBM connection (B weights)
        for node in self.input:
            var_name = node.name
            shp_visible = self.hyperparameters['shapes'][var_name]

            # create logit mask for B
            size = shp_visible + shp_output
            mask = np.ones(size, DTYPE_FLOATX)
            mask[..., -1] = 0.
            mask = mask.flatten()

            # create the tensor shared variables B
            b_name = 'B_' + var_name + '_' + name
            if b_name in self.model_values.keys():
                value = self.model_values[b_name]
            else:
                value = np.zeros(np.prod(size), DTYPE_FLOATX)  # * mask

            B_f = theano.shared(value, b_name)
            B_m = theano.shared(mask, b_name+'_mask')
            B = T.reshape(B_f, size)

            self.B_params.append(B)
            self.B_params_f.append(B_f)
            self.B_params_m.append(B_m)

            self.params[b_name] = B_f
            self.params_shp[b_name] = shp_visible + shp_output

    def free_energy(self, input, utility=0):
        """
        Free energy function

        Parameters
        ----------
        self : RBM class object

        input : `[T.tensors]`, optional
            Used when calculating free energy of gibbs chain sampling

        Returns
        -------
        F(y, x) :
            Scalar value of the generative model free energy

        :math:
        `F(y, x, h) = -(xWh + yWh + vbias*x + hbias*h + cbias*y)`\n
        `    wx_b = xW + yW + hbias`\n
        `  F(y, x) = -{vbias*x + cbias*y + sum_k[ln(1+exp(wx_b))]}`\n

        """
        # collect parameters
        visibles = input
        vbiases = self.vbias
        W_params = self.W_params

        dtypes = self.in_dtype
        hbias = self.hbias[0]

        # input shapes as (rows, items, cats) or (rows, outs)
        # weight shapes as (items, cats, hiddens) or (outs, hiddens)
        # bias shapes as (items, cats) or (outs,)
        wx_hbias = hbias
        for dtype, v, W, vbias in zip(dtypes, visibles, W_params, vbiases):
            # vbias_x: (rows,)
            if dtype == VARIABLE_DTYPE_CATEGORY:
                vbias_x = T.tensordot(v, vbias, axes=[
                    list(range(v.ndim)[1:]), list(range(vbias.ndim)[-2:])])
                utility -= vbias_x

                wx = T.tensordot(v, W, axes=[
                    list(range(v.ndim)[1:]), list(range(W.ndim)[:-1])])

            else:
                vbias_x = T.sum(0.5 * T.sqr(v - vbias[None, ...]),
                                axis=list(range(v.ndim)[1:]))
                utility += vbias_x
                wx = T.tensordot(v, W, axes=[
                    list(range(v.ndim)[1:]), list(range(W.ndim)[:-1])])

            # wx_hbias: (rows, hiddens)
            wx_hbias += wx

        # sums over hidden axis --> (rows,)
        return utility - T.sum(T.log(1. + T.exp(wx_hbias)), axis=1)

    def sample_h_given_v(self, v0_samples, vtype='xy'):
        """
        sample_h_given_v func
            Binomial hidden units

        Parameters
        ----------
        v0_samples : `[T.tensors]`
            theano Tensor variable

        Returns
        -------
        h1_preactivation : `scalar` (-inf, inf)
            preactivation function e.g. logit utility func
        h1_means : `scalar` (0, 1)
            sigmoid activation
        h1_samples : `integer` 0 or 1
            binary samples
        """
        # prop up
        W_params = self.W_params
        dtypes = self.in_dtype

        hbias = self.hbias
        h1_preactivation = self.propup(v0_samples, W_params, hbias[0], dtypes)

        # h ~ p(h|v0_sample)
        h1_means = T.nnet.sigmoid(h1_preactivation)
        h1_samples = self.theano_rng.binomial(
            size=h1_means.shape, p=h1_means, dtype=DTYPE_FLOATX)

        return h1_preactivation, h1_means, h1_samples

    def propup(self, samples, weights, bias, dtypes):

        preactivation = bias
        # (rows, items, cats), (items, cats, hiddens)
        # (rows, outs), (outs, hiddens)
        for v, W, dtype in zip(samples, weights, dtypes):
            preactivation += T.tensordot(v, W, axes=[
                    list(range(v.ndim)[1:]), list(range(W.ndim)[:-1])])

        return preactivation

    def sample_v_given_h(self, h0_samples, vtype='xy'):
        """
        sample_v_given_h func
            Binomial hidden units

        Parameters
        ----------
        h0_samples : `[T.tensors]`
            theano Tensor variable

        Returns
        -------
        v1_preactivation : `[scalar]` (-inf, inf)
            sequence of preactivation function e.g. logit utility func
        v1_means : `[scalar]` (0, 1)
            sequence of sigmoid activation
        v1_samples : `[binary]` or `[integer]` or `[float32]` or `[array[j]]`
            visible unit samples
        """
        # prop down
        V_params = self.W_params
        bias = self.vbias
        dtypes = self.in_dtype

        v1_preactivations = self.propdown(h0_samples, V_params, bias)
        # v ~ p(v|h0_sample)
        v1_means = []
        v1_samples = []
        for v1, dtype in zip(v1_preactivations, dtypes):
            if dtype == VARIABLE_DTYPE_BINARY:
                v1_mean = T.nnet.sigmoid(v1)
                v1_sample = self.theano_rng.binomial(
                    size=v1.shape, p=v1_mean, dtype=DTYPE_FLOATX)

            elif dtype == VARIABLE_DTYPE_CATEGORY:
                uniform = self.theano_rng.uniform(
                    size=v1.shape, low=1e-10, high=1.0, dtype=DTYPE_FLOATX)

                # reshape softmax tensors to 2D matrix
                if v1.ndim == 3:
                    (d1, d2, d3) = v1.shape
                    v1 = v1.reshape((d1 * d2, d3))
                    reshp_flag = 1

                v1_mean = T.nnet.softmax(v1)
                v1_sample = self.theano_rng.multinomial(
                    pvals=v1_mean, dtype=DTYPE_FLOATX)

                if reshp_flag == 1:
                    # reshape back into original dimensions
                    v1_mean = v1_mean.reshape((d1, d2, d3))
                    v1_sample = v1_sample.reshape((d1, d2, d3))

            elif dtype == VARIABLE_DTYPE_REAL:
                v1_std = T.nnet.sigmoid(T.abs_(v1))
                normal_sample = self.theano_rng.normal(
                    size=v1.shape,  avg=v1, std=v1_std, dtype=DTYPE_FLOATX)
                v1_mean = v1
                v1_sample = normal_sample

            elif dtype == VARIABLE_DTYPE_INTEGER:
                v1_std = T.nnet.sigmoid(v1)
                normal_sample = self.theano_rng.normal(
                    size=v1.shape, avg=v1, std=v1_std, dtype=DTYPE_FLOATX)
                v1_mean = T.nnet.softplus(v1)
                v1_sample = T.nnet.softplus(normal_sample)

            else:
                raise NotImplementedError

            v1_means.append(v1_mean)
            v1_samples.append(v1_sample)

        return v1_preactivations, v1_means, v1_samples

    def propdown(self, samples, weights, bias):

        preactivation = []
        for W, b in zip(weights, bias):
            if W.ndim == 2:
                W = W.dimshuffle(1, 0)
            else:
                W = W.dimshuffle(0, 2, 1)
            # add visible bias
            preactivation.append(T.dot(samples, W) + b)
        return preactivation

    def gibbs_hvh(self, h0_samples):
        v1_pre, v1_means, v1_samples = self.sample_v_given_h(h0_samples)
        h1_pre, h1_means, h1_samples = self.sample_h_given_v(v1_samples)

        return v1_pre + v1_means + v1_samples + \
            [h1_pre] + [h1_means] + [h1_samples]

    def gibbs_vhv(self, *v0_samples):
        h1_pre, h1_means, h1_samples = self.sample_h_given_v(v0_samples)
        v1_pre, v1_means, v1_samples = self.sample_v_given_h(h1_samples)

        return [h1_pre] + [h1_means] + [h1_samples] + \
            v1_pre + v1_means + v1_samples

    def get_generative_cost_updates(self, k=1):
        """
        get_generative_cost_updates func
            updates weights for W^(1), W^(2), a, c and d
        """
        # prepare visible samples from x input and y outputs
        v0_samples = self.input
        v0_size = len(v0_samples)

        # perform positive Gibbs sampling phase
        # one step Gibbs sampling p(h|v1,v2,...) = p(h|v1)+p(h|v2)+...
        h0_pre, h0_means, h0_samples = self.sample_h_given_v(v0_samples)

        # start of Gibbs sampling chain
        # we only want the samples generated from the Gibbs sampling phase
        chain_start = h0_samples
        scan_out = 3 * v0_size * [None] + [None, None, chain_start]

        # theano scan function to loop over all Gibbs steps k
        # [v1_pre[], v1_means[], v1_samples[], h1_pre, h1_means, h1_samples]
        # outputs are given by outputs_info
        # [[t,t+1,t+2,...], [t,t+1,t+2,...], ], gibbs_updates
        # NOTE: scan returns a dictionary of updates
        outputs, gibbs_updates = theano.scan(
            fn=self.gibbs_hvh, outputs_info=scan_out, n_steps=k,
            name='gibbs_hvh'
        )

        # note that we only need the visible samples at the end of the chain
        chain_end = []
        for output in outputs:
            chain_end.append(output[-1])
        vn_pre = chain_end[:v0_size]
        vn_means = chain_end[v0_size: 2 * v0_size]
        vn_samples = chain_end[2 * v0_size: 3 * v0_size]

        # calculate the model cost
        params = self.V_params_f + self.vbias_f + self.hbias
        positive_phase = T.mean(self.free_energy(v0_samples))
        negative_phase = T.mean(self.free_energy(vn_means))
        cost = positive_phase - negative_phase

        # calculate the gradients
        grads = T.grad(cost=cost, wrt=params, consider_constant=vn_means)
        jacobians = T.grad(cost=cost,
                           wrt=self.input,
                           consider_constant=vn_means)

        # update Gibbs chain with update expressions from updates list[]
        updates = self.update_opt(params, grads, self.decay*self.learning_rate)
        for parameter, expression in updates:
            gibbs_updates[parameter] = expression

        monitoring_cost = self.pseudo_loglikelihood(
            inputs=v0_samples, preactivation=vn_pre)

        return (monitoring_cost, gibbs_updates, positive_phase,
                negative_phase, jacobians, self.input, vn_means)

    def get_v_samples(self, k):
        # prepare visible samples from input
        v0_samples = self.input
        print(v0_samples)
        v0_size = len(v0_samples)
        h0_pre, h0_means, h0_samples = self.sample_h_given_v(v0_samples)
        scan_out = 3 * v0_size * [None] + [None, None, h0_samples]

        # theano scan function to loop over all Gibbs steps k
        # [v1_pre[], v1_means[], v1_samples[], h1_pre, h1_means, h1_samples]
        # outputs are given by outputs_info
        # [[t,t+1,t+2,...], [t,t+1,t+2,...], ], gibbs_updates
        # NOTE: scan returns a dictionary of updates
        gibbs_output, gibbs_updates = theano.scan(
            fn=self.gibbs_hvh,
            outputs_info=scan_out,
            n_steps=k,
            name='gibbs_sampling'
        )

        ## note that we only need the visible samples at the end of the chain
        chain_end = []
        for output in gibbs_output:
            chain_end.append(output[-1])
        vn_pre = chain_end[:v0_size]
        vn_means = chain_end[v0_size: 2 * v0_size]
        vn_samples = chain_end[2 * v0_size: 3 * v0_size]

        return vn_means, gibbs_updates

    def pseudo_loglikelihood(self, inputs, preactivation):
        """
        pseudo_loglikelihood func
            Function to calculate the (pseudo) neg loglikelihood

        Parameters
        ----------
        inputs : `[T.tensors]`
            list of input tensors
        preactivation : `[T.shared]`
            list of precomputed "logits"

        Returns
        -------
        pll : `scalar`
            value of the pseudo log likelihood
        """
        dtypes = self.in_dtype
        epsilon = 1e-10  # small value to prevent log(0.)
        cross_entropy = 0
        loglikelihood = 0
        mse_r = 0
        mse_i = 0
        for input, v1, dtype in zip(inputs, preactivation, dtypes):
            if dtype == VARIABLE_DTYPE_BINARY:
                cross_entropy -= T.mean(T.sum(
                    input * T.log(T.nnet.sigmoid(v1))), axis=1
                )

            elif dtype == VARIABLE_DTYPE_CATEGORY:
                (d1, d2, d3) = v1.shape
                v1_mean = T.nnet.softmax(v1.reshape((d1 * d2, d3)))
                # reshape back into original dimensions
                v1_mean = v1_mean.reshape((d1, d2, d3))
                loglikelihood -= T.mean(input * T.log(v1_mean))

            elif dtype == VARIABLE_DTYPE_REAL:
                v = v1
                mse_r += T.sqrt(T.mean(T.sqr(input - v)))

            elif dtype == VARIABLE_DTYPE_INTEGER:
                v = T.nnet.softplus(v1)
                mse_i += T.sqrt(T.mean(T.sqr(input - v)))

            else:
                raise NotImplementedError

        return [loglikelihood, mse_r, mse_i]

    def generator(self, h5pydataset, var_list):
        shared_inputs_valid = []
        for var in var_list:
            shared_inputs_valid.append(
                theano.shared(h5pydataset[var]['data'][:].astype(DTYPE_FLOATX),
                              borrow=True))

        gibbs_sampling_steps = T.iscalar('steps')
        vsamples, vsamples_updates = self.get_v_samples(gibbs_sampling_steps)

        tensor_inputs = self.input
        self.sample = theano.function(
            inputs=[gibbs_sampling_steps],
            outputs=vsamples,
            updates=vsamples_updates,
            givens={
                key: val[:]
                for key, val in zip(tensor_inputs, shared_inputs_valid)},
            name='sample',
            allow_input_downcast=True,
            on_unused_input='ignore'
        )

    def initialize(self, x):
        self.add_latent()

        for item in x:
            print('x', item.name.strip('/'), item['data'].shape[1:])
    self.add_node(
        var_dtype=item.attrs['dtype'],
        name=item.name.strip('/'),
        shp_visible=item['data'].shape[1:]
    )

k = self.hyperparameters['gibbs_steps']
batch_size = self.hyperparameters['batch_size']
n_samples = self.hyperparameters['n_samples']

(
    monitoring_cost, gibbs_updates, positive_phase, negative_phase,
    jacobians, batch_inputs, batch_outputs
) = self.get_generative_cost_updates(k)

tensor_inputs = self.input
elasticity = [
    jacobians[0] *
    batch_inputs[4].dimshuffle((0, 1, 'x')) /
    batch_outputs[0]
]

tensor_outputs = monitoring_cost + [positive_phase, negative_phase] +\
    elasticity
tensor_updates = gibbs_updates

shared_inputs = [
    theano.shared(
        item['data'][:].astype(DTYPE_FLOATX),
        borrow=True) for item in x]

ind = T.iscalar('index')
decay_rate = T.scalar('decay_rate')
start_idx = ind * batch_size
end_idx = (ind + 1) * batch_size

print('constructing Theano computational graph...')
self.train = theano.function(
    inputs=[ind],
    outputs=tensor_outputs,
    updates=tensor_updates,
    givens={
        key: val[start_idx: end_idx]
        for key, val in zip(tensor_inputs, shared_inputs)},
    name='train',
    allow_input_downcast=True
)

self.decay_learning_rate = theano.function(
    inputs=[decay_rate],
    outputs=self.decay * decay_rate,
    updates=((self.decay, self.decay * decay_rate), ),
    name='decay_learning_rate'
)
```

Appendix D

Dataset description

Table D.1: Table of descriptive statistics for Train Hôtel dataset

variable	description	type	mean
DrvLicens	Driving License	binary	0.865
PblcTrst	Public transit pass	binary	0.639
Ag1825	Age between 18 to 25	binary	0.072
Ag2545	Age between 25 to 45	binary	0.414
Ag4565	Age between 45 to 65	binary	0.376
Ag65M	Age above 65	binary	0.137
Male	1: male 0: female	binary	0.446
Fulltime	Full time employed	binary	0.569
Edu_Highschl	Highest education high school	binary	0.193
Edu_BSc	Highest education bachelors	binary	0.634
Edu_MscPhD	Highest education graduate	binary	0.172
HH_Veh0	0 household vehicles	binary	0.219
HH_Veh1	1 household vehicle	binary	0.552
HH_Veh2M	2+ household vehicles	binary	0.228
HH_Chld0	0 children in household	binary	0.748
HH_Chld1	1 children in household	binary	0.139
HH_Chld2M	2+ children in household	binary	0.112
HH_Inc020K	Income less than 20K	binary	0.204
HH_Inc2060K	Income between 20K and 60K	binary	0.361
HH_Inc60KM	Income more than 60K	binary	0.32
Choice	1: Car, 2: Car Rental, 3: Bus, 4: Plane, 5: Train, 6: Train Hotel	categorical	

Dataset can be obtained from the Train Hôtel study [76].
A. Sobhani and B. Farooq. “Innovative Intercity Transport Mode: Application of Choice Preference Integrated with Attributes Nonattendance and Value Learning”. In: 21st International Federation of Operational Research Societies, Québéc City. 2017.

Table D.2: Table of descriptive statistics for Santander dataset

variable	description	type	mean	std dev
age	Age of customer	continuous	42.9	0.02593
loyalty	Customer loyalty (years)	continuous	8.032	0.01191
income	Income (‘000s)	continuous	0.142	0.000522
sex	1: male 0: female	binary	0.388	0.000967
employee	Is employee	binary	0.001	4.86E-05
active	Is active customer	binary	0.959	0.000395
new_cust	1: loyalty < 6mos 0: > 6mos	binary	0.045	0.000412
resident	Is resident	binary	1	1.42E-05
foreigner	Is foreigner	binary	0.045	0.000413
european	Is European	binary	1	1.11E-05
vip	Is VIP customer	binary	0.117	0.000637
savings	Savings account	binary	0.00015	2.43E-05
current	Current account	binary	0.572	0.000982
derivada	Derivada account	binary	0.001	5.96E-05
payroll_acc	Payroll account	binary	0.416	0.000978
junior	Junior account	binary	9.46E-05	1.93E-05
masparti	Mas Particular account	binary	0.017	0.000254
particular	Particular account	binary	0.168	0.000742
partiplus	Particular Plus account	binary	0.113	0.000628
e_acc	E-Account	binary	0.255	0.000866
choice	Product codes 1: aval 2: deco 3: fond 4: hip 5: plan 6: pres 7: reca 8: tjcr 9: valo 10: nomina	categorical		

Dataset obtained from

https://www.kaggle.com/c/santander-product-recommendation/data

Table D.3: Table of descriptive statistics for Mtltrajet dataset

variable	description	type	mean	std dev
avg_speed	Average trip speed (km/h)	continuous	25.870	0.062
duration	Trip duration (mins)	continuous	21.072	0.287
n_coord	Number of links	continuous	91.501	0.232
trip_km	Trip distance (km)	continuous	6.893	0.017
mode choice	1: cycling 2: driving 3: driving+transit 4: transit 5: walk 6: other	categorical		
activity choice	1: education 2: health 3: leisure 4: meal 5: errand 6: shopping 7: home 8: work 9: meetings 10: others	categorical		
district	District ID: 34 neighbourhoods	categorical		
interval_15	24hr in 15 min intervals	categorical		

Dataset obtained from
http://donnees.ville.montreal.qc.ca/dataset/mtl-trajet

References

[1] M. Kamargianni et al. "A critical review of new mobility services for urban transport". In: *Transportation Research Procedia* 14 (2016), pp. 3294–3303.

[2] Y. Wang et al. "Enhancing transportation systems via deep learning: A survey". In: *Transportation research part C: emerging technologies* (2018).

[3] J. Van Brummelen et al. "Autonomous vehicle perception: The technology of today and tomorrow". In: *Transportation research part C: emerging technologies* 89 (2018), pp. 384–406.

[4] L. Zha et al. "Economic analysis of ride-sourcing markets". In: *Transportation Research Part C: Emerging Technologies* 71 (2016), pp. 249–266.

[5] R. R. Clewlow and G. S. Mishra. *Disruptive transportation: The adoption, utilization, and impacts of ride-hailing in the United States*. Tech. rep. UCD-ITS-RR-17-07. University of California, Davis, Institute of Transportation Studies, Davis, CA, 2017.

[6] M. Kamargianni and M. Matyas. "The business ecosystem of mobility-as-a-service". In: *transportation research board*. Vol. 96. Transportation Research Board. 2017.

[7] R. Krueger et al. "Preferences for shared autonomous vehicles". In: *Transportation research part C: emerging technologies* 69 (2016), pp. 343–355.

[8] European Commission. *On the road to automated mobility: An EU strategy for mobility of the future*. Tech. rep. COM(2018) 283. 2018.

[9] C. Simpson et al. *Mobility 2030: Transforming the Mobility Landscape*. Tech. rep. United Kingdom: KPMG Global Strategy Group, 2019.

[10] C. Chen et al. “The promises of big data and small data for travel behavior (aka human mobility) analysis”. In: *Transportation research part C: emerging technologies* 68 (2016), pp. 285–299.

[11] M. Ben-Akiva and M. Bierlaire. “Discrete choice methods and their applications to short term travel decisions”. In: *Handbook of transportation science*. Springer, 1999, pp. 5–33.

[12] K. He et al. “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[13] X. Li and J. She. “Collaborative variational autoencoder for recommender systems”. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 305–314.

[14] T. Brathwaite et al. “Machine learning meets microeconomics: The case of decision trees and discrete choice”. In: *arXiv preprint arXiv:1711.04826* (2017).

[15] L. Breiman et al. “Statistical modeling: The two cultures”. In: *Statistical science* 16.3 (2001), pp. 199–231.

[16] K. Friston et al. “The anatomy of choice: active inference and agency”. In: *Frontiers in human neuroscience* 7 (2013), p. 598.

[17] F. J. Bremner et al. “Hinton diagrams: Viewing connection strengths in neural networks”. In: *Behavior Research Methods, Instruments, & Computers* 26.2 (1994), pp. 215–218.

[18] C. A. Sims. “Implications of rational inattention”. In: *Journal of monetary Economics* 50.3 (2003), pp. 665–690.

[19] G. M. Becker et al. “Stochastic models of choice behavior”. In: *Behavioral science* 8.1 (1963), pp. 41–55.

[20] D. Kahneman and A. Tversky. “Prospect Theory: An Analysis of Decision under Risk”. In: *Econometrica* 47.2 (1979), pp. 263–292.

[21] C. R. Bhat et al. “Flexible Model Structures for Discrete Choice Analysis”. In: vol. 2. Emerald, Inc., 2008. Chap. 5, pp. 70–90.

[22] A. De Palma et al. “Risk, uncertainty and discrete choice models”. In: *Marketing Letters* 19.3-4 (2008), pp. 269–285.

[23] A. Tversky and D. Kahneman. “Advances in Prospect Theory: Cumulative Representation of Uncertainty”. In: *A practical Guide to Sentiment Analysis*. Ed. by E. Cambria et al. Vol. 5. Socio-Affective Computing. Springer International Publishing, 2017. Chap. 24, pp. 493–519.

[24] L. Masiero and D. A. Hensher. “Analyzing loss aversion and diminishing sensitivity in a freight transport stated choice experiment”. In: *Transportation Research Part A: Policy and Practice* 44.5 (2010), pp. 349–358.

[25] M. Fosgerau and G. Jiang. “Travel time variability and rational inattention”. In: *Transportation Research Part B: Methodological* 120 (2019), pp. 1–14.

[26] D. A. Gopinath. “Modeling Heterogeneity in Discrete Choice Processes: Application to Travel Demand”. PhD thesis. MIT, 1995.

[27] K. E. Train. *Discrete choice methods with simulation*. 2nd ed. Cambridge University Press, 2009.

[28] D. A. Hensher and W. H. Greene. “The mixed logit model: the state of practice”. In: *Transportation* 30.2 (2003), pp. 133–176.

[29] C. R. Bhat. “A new flexible multiple discrete–continuous extreme value (MDCEV) choice model”. In: *Transportation Research Part B: Methodological* 110 (2018), pp. 261–279.

[30] F. Matějka and A. McKay. “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model”. In: *American Economic Review* 105.1 (2015), pp. 272–298.

[31] M. Fosgerau et al. “A link based network route choice model with unrestricted choice set”. In: *Transportation Research Part B: Methodological* 56 (2013), pp. 70–80.

[32] C. G. Chorus et al. “A random regret-minimization model of travel choice”. In: *Transportation Research Part B: Methodological* 42.1 (2008), pp. 1–18.

[33] M. Ben-Akiva et al. “Process and context in choice models”. In: *Marketing Letters* 23.2 (2012), pp. 439–456.

[34] M. Paulssen et al. “Values, attitudes and travel behavior: a hierarchical latent variable mixed logit model of travel mode choice”. In: *Transportation* 41.4 (2014), pp. 873–888.

[35] A. Vij and J. L. Walker. “How, when and why integrated choice and latent variable models are latently useful”. In: *Transportation Research Part B: Methodological* 90 (2016), pp. 192–217.

[36] J. Shen. “Latent class model or mixed logit model? A comparison by transport mode choice data”. In: *Applied Economics* 41.22 (2009), pp. 2915–2924.

[37] D. Li et al. “Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets”. In: *Transportation Research Part C: Emerging Technologies* 67 (), pp. 31–46.

[38] M. Ben-Akiva et al. “Integration of choice and latent variable models”. In: *Perpetual motion: Travel behaviour research opportunities and application challenges* (2002), pp. 431–470.

[39] P. Schwartenbeck et al. “Evidence for surprise minimization over value maximization in choice behavior”. In: 5.16575 (2015), pp. 1–14.

[40] P. Dayan et al. “The helmholtz machine”. In: *Neural computation* 7.5 (1995), pp. 889–904.

[41] P. A. Ortega and D. A. Braun. “Thermodynamics as a theory of decision-making with information-processing costs”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469.2153 (2013), p. 20120683.

[42] C. L. Buckley et al. “The free energy principle for action and perception: A mathematical review”. In: *Journal of Mathematical Psychology* 81 (2017), pp. 55–79.

[43] D. McFadden. “Conditional logit analysis of qualitative choice behaviour”. In: *Frontiers in Economics*. Ed. by P. Zarembka. New York: Academic Press. Chap. 4, pp. 105–142.

[44] J. L. Walker et al. “Correcting for endogeneity in behavioral choice models with social influence variables”. In: *Transportation Research Part A: Policy and Practice* 45.4 (2011), pp. 362–374.

[45] I. Goodfellow et al. *Deep Learning.* MIT Press, 2016.

[46] C. G. Chorus and M. Kroesen. "On the (im-) possibility of deriving transport policy implications from hybrid choice models". In: *Transport Policy* 36 (2014), pp. 217–222.

[47] I. H. Witten et al. *Data mining: Practical machine learning tools and techniques.* 3rd ed. Morgan Kaufmann series in data management systems. Morgan Kaufman, 2011.

[48] H. Abdi and L. J. Williams. "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.

[49] R. Salakhutdinov et al. "Restricted Boltzmann machines for collaborative filtering". In: *Proceedings of the 24th international conference on Machine learning.* ACM. 2007, pp. 791–798.

[50] R. Salakhutdinov. "Learning deep generative models". In: *Annual Review of Statistics and Its Application* 2 (2015), pp. 361–385.

[51] Y. Lecun. "A theoretical framework for back-propagation". In: *Proceedings of the 1988 Connectionist Models Summer School, CMU, Pittsburg, PA.* Morgan Kaufmann, 1988, pp. 21–28.

[52] X. Glorot et al. "Deep Sparse Rectifier Neural Networks". In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics.* Ed. by G. Gordon et al. Vol. 15. JMLR Proceedings. JMLR.org, 2011, pp. 315–323.

[53] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the 13th international conference on artificial intelligence and statistics.* 2010, pp. 249–256.

[54] Y. Bengio et al. "Better mixing via deep representations". In: *International conference on machine learning.* 2013, pp. 552–560.

[55] D. J. MacKay. "Bayesian neural networks and density networks". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 354.1 (1995), pp. 73–80.

[56] G. E. Hinton. "Training products of experts by minimizing contrastive divergence". In: *Neural Computation* 14.8 (2002), pp. 1771–1800.

[57] H. Larochelle and Y. Bengio. "Classification using discriminative restricted Boltzmann machines". In: (2008), pp. 536–543.

[58] C. A. Sims. "Rational inattention and monetary economics". In: *Handbook of monetary economics*. Ed. by B. M. Friedman and M. Woodford. Vol. 3. Elsevier, 2010. Chap. 4, pp. 155–181.

[59] G. E. Hinton et al. *Boltzmann machines: Constraint satisfaction networks that learn*. Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, 1984.

[60] K. J. Friston and K. E. Stephan. "Free-energy and the brain". In: *Synthese* 159.3 (2007), pp. 417–458.

[61] D. E. Bell. "Regret in decision making under uncertainty". In: *Operations research* 30.5 (1982), pp. 961–981.

[62] A. Anas. "Discrete choice theory, information theory and the multinomial logit and gravity models". In: *Transportation Research Part B: Methodological* 17.1 (1983), pp. 13–23.

[63] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer Series in Statistics. Springer New York, 1985.

[64] M. Biehl et al. "Expanding the Active Inference Landscape: More Intrinsic Motivations in the Perception-Action Loop". In: *Frontiers in Neurorobotics* 12 (2018), p. 45.

[65] M. Wong et al. "Discriminative conditional restricted Boltzmann machine for discrete choice and latent variable modelling". In: *Journal of Choice Modelling* 29 (2018), pp. 152–168.

[66] Y. Bengio and S. Bengio. "Modeling high-dimensional discrete data with multi-layer neural networks". In: *Advances in Neural Information Processing Systems*. 2000, pp. 400–406.

[67] D. M. Blei et al. "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[68] D. McFadden and K. Train. “Mixed MNL models for discrete response”. In: *Journal of applied Econometrics* 15.5 (2000), pp. 447–470.

[69] C. R. Bhat et al. “Introducing non-normality of latent psychological constructs in choice modeling with an application to bicyclist route choice”. In: *Transportation Research Part B: Methodological* 78 (2015), pp. 341–363.

[70] M. Wong et al. “Discriminative conditional restricted Boltzmann machine for discrete choice and latent variable modelling”. In: *Journal of Choice Modelling* (2017).

[71] A. Daly et al. “Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour”. In: *Transportation* 39.2 (2012), pp. 267–297.

[72] K. Ashok et al. “Extending discrete choice models to incorporate attitudinal and other latent variables”. In: *Journal of marketing research* 39.1 (2002), pp. 31–46.

[73] T. F. Golob. “Structural equation modeling for travel behavior research”. In: *Transportation Research Part B: Methodological* 37.1 (2003), pp. 1–25.

[74] T. Morikawa et al. “Discrete choice models incorporating revealed preferences and psychometric data”. In: *Advances in Econometrics*. Emerald Group Publishing Limited, 2002, pp. 29–55.

[75] T. J. Klette and Z. Griliches. “The inconsistency of common scale estimators when output prices are unobserved and endogenous”. In: *Journal of Applied Econometrics* (1996), pp. 343–361.

[76] A. Sobhani and B. Farooq. “Innovative Intercity Transport Mode: Application of Choice Preference Integrated with Attributes Nonattendance and Value Learning”. In: *21st International Federation of Operational Research Societies, Québéc City*. 2017.

[77] C. M. Rungie et al. “Latent variables in discrete choice experiments”. In: *Journal of Choice Modelling* 5.3 (2012), pp. 145–156.

[78] N. Le Roux and Y. Bengio. “Representational Power of Restricted Boltzmann Machines and Deep Belief Networks”. In: *Neural Computation* 20.6 (2008), pp. 1631–1649.

[79] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach.* Springer Science & Business Media, 2003.

[80] B. Atasoy et al. “Attitudes towards mode choice in Switzerland”. In: *disP-The Planning Review* 49.2 (2013), pp. 101–117.

[81] C. R. Bhat and S. K. Dubey. “A new estimation approach to integrate latent psychological constructs in choice modeling”. In: *Transportation Research Part B: Methodological* 67 (2014), pp. 68–85.

[82] B. Eric et al. “Active preference learning with discrete choice data”. In: *Advances in neural information processing systems.* Vol. 20. 2008, pp. 409–416.

[83] E. Morey et al. “Using angler characteristics and attitudinal data to identify environmental preference classes: a latent-class model”. In: *Environmental and Resource Economics* 34.1 (2006), pp. 91–115.

[84] A. Hackbarth and R. Madlener. “Consumer preferences for alternative fuel vehicles: A discrete choice analysis”. In: *Transportation Research Part D: Transport and Environment* 25 (2013), pp. 5–17.

[85] A. Yazdizadeh et al. “A Generic Form for Capturing Unobserved Heterogeneity in Discrete Choice Modeling: Application to Neighborhood Location Choice”. In: *Transportation Research Board 96th Annual Meeting.* 17-05144. 2017.

[86] S. Hess and A. Daly. *Handbook of choice modelling.* Edward Elgar Publishing, 2014.

[87] A. Glerum et al. “Forecasting the Demand for Electric Vehicles: Accounting for Attitudes and Perceptions”. In: *Transportation Science* 48.4 (2014).

[88] S. Hess et al. “Accommodating underlying pro-environmental attitudes in a rail travel context: application of a latent variable latent class specification”. In: *Transportation Research Part D: Transport and Environment* 25 (2013), pp. 42–48.

[89] R. Maldonado-Hinarejos et al. “Exploring the role of individual attitudes and perceptions in predicting the demand for cycling: a hybrid choice modelling approach”. In: *Transportation* 41.6 (2014), pp. 1287–1304.

[90] J. Kim et al. “Expanding scope of hybrid choice models allowing for mixture of social influences and latent attitudes: Application to intended purchase of electric cars”. In: *Transportation research part A: policy and practice* 69 (2014), pp. 71–85.

[91] G. Poucin et al. “Pedestrian Activity Pattern Mining in WiFi-Network Connection Data”. In: *Transportation Research Board 95th Annual Meeting*. 16-5846. 2016.

[92] N. K. Ahmed et al. “An empirical comparison of machine learning models for time series forecasting”. In: *Econometric Reviews* 29.5-6 (2010), pp. 594–621.

[93] A. Rosenfeld et al. “Combining psychological models with machine learning to better predict people’s decisions”. In: *Synthese* 189.1 (2012), pp. 81–93.

[94] T. Osogami and M. Otsuka. “Restricted Boltzmann machines modeling human choice”. In: *Advances in Neural Information Processing Systems* 26 (2014), pp. 73–81.

[95] M. Aggarwal. “On Learning of Choice Models with Interactive Attributes”. In: *IEEE Transactions on Knowledge and Data Engineering* 28.10 (2016), pp. 2697–2708.

[96] M. Ben-Akiva et al. “Hybrid choice models: progress and challenges”. In: *Marketing Letters* 13.3 (2002), pp. 163–175.

[97] G. E. Hinton and R. R. Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* 313.5786 (2006), pp. 504–507.

[98] M. Wong et al. “Next Direction Route Choice Model for Cyclist using Panel Data”. In: *51st Annual Conference of Canadian Transportation Research Forum*. 2016.

[99] J. K. Vermunt and J. Magidson. “Latent class cluster analysis”. In: *Applied latent class analysis* 11 (2002), pp. 89–106.

[100] G. E. Hinton. “A practical guide to training restricted Boltzmann machines”. In: *Neural networks: Tricks of the trade*. Ed. by G. Montavon et al. Springer Berlin Heidelberg, 2012. Chap. 24, pp. 599–619.

[101] A. Y. Ng and M. I. Jordan. “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes”. In: *Advances in neural information processing systems*. Vol. 14. 2002, pp. 841–848.

[102] M. A. Carreira-Perpinan and G. E. Hinton. “On contrastive divergence learning”. In: *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*. Ed. by R. G. Cowell and Z. Ghahramani. Society for Artificial Intelligence and Statistics, 2005, pp. 33–40.

[103] V. Mnih et al. “Conditional restricted boltzmann machines for structured output prediction”. In: *arXiv preprint arXiv:1202.3748* (2012).

[104] G. E. Hinton et al. “A fast learning algorithm for deep belief nets”. In: *Neural computation* 18.7 (2006), pp. 1527–1554.

[105] J. C. Helton et al. *Sensitivity analysis techniques and results for performance assessment at the waste isolation pilot plant*. Tech. rep. Sandia National Labs., 1991.

[106] A. Saltelli et al. “Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index”. In: *Computer Physics Communications* 181.2 (2010), pp. 259–270.

[107] G. W. Taylor et al. “Modeling human motion using binary latent variables”. In: *Advances in neural information processing systems*. Vol. 19. 2007, pp. 1345–1352.

[108] A. Vij and R. Krueger. “Random taste heterogeneity in discrete choice models: Flexible nonparametric finite mixture distributions”. In: *Transportation Research Part B: Methodological* 106 (2017), pp. 76–101.

[109] M. Nikolić and M. Bierlaire. “Data-driven spatio-temporal discretization for pedestrian flow characterization”. In: *Transportation research procedia* 23 (2017), pp. 188–207.

[110] D. Bolduc and R. Alvarez-Daziano. “On Estimation of Hybrid Choice Models”. In: *Choice Modelling: The State-of-the-art and The State-of-practice*. Ed. by S. Hess and A. Daly. Edward Elgar, 2010, pp. 259–287. ISBN: 9781849507721.

[111] E. Cherchi and J. W. Polak. “Assessing user benefits with discrete choice models: Implications of Specification errors under random taste heterogeneity”. In: *Transportation Research Record* 1926.1 (2005), pp. 61–69.

[112] H. Alizadeh et al. “An online survey to enhance the understanding of car drivers route choices”. In: *Transportation Research Procedia* 32 (2018), pp. 482–494.

[113] M. Fosgerau et al. “Discrete Choice and Rational Inattention: a General Equivalence Result”. In: *arXiv preprint arXiv:1709.09117* (2017).

[114] D. Ellsberg. “Risk, ambiguity, and the Savage axioms”. In: *The quarterly journal of economics* (1961), pp. 643–669.

[115] J. Steiner et al. “Rational Inattention Dynamics: Inertia and Delay in Decision-Making”. In: *Econometrica* 85.2 (2017), pp. 521–553.

[116] C. Teye et al. “Entropy maximising facility location model for port city intermodal terminals”. In: *Transportation Research Part E: Logistics and Transportation Review* 100 (2017), pp. 1–16.

[117] B. Leard. “Consumer inattention and the demand for vehicle fuel cost savings”. In: *Journal of choice modelling* 29 (2018), pp. 1–16.

[118] A. Ullah. “Entropy, divergence and distance measures with econometric applications”. In: *Journal of Statistical Planning and Inference* 49.1 (1996), pp. 137–162.

[119] M. Ranzato et al. “Efficient learning of sparse representations with an energy-based model”. In: *Advances in neural information processing systems*. 2007, pp. 1137–1144.

[120] D. P. Kingma and M. Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[121] A. Alwosheel et al. “Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis”. In: *Journal of choice modelling* 28 (2018), pp. 167–182.

[122] K. He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[123] Y. W. Teh et al. “Energy-based models for sparse overcomplete representations”. In: *Journal of Machine Learning Research* 4 (2003), pp. 1235–1260.

[124] Ville de Montréal. *Déplacements MTL Trajet*. https://ville.montreal.qc.ca/mtltrajet/. 2016.

[125] M. Ranzato et al. “Sparse feature learning for deep belief networks”. In: *Advances in neural information processing systems*. 2008, pp. 1185–1192.

[126] B. Farooq et al. “Simulation based population synthesis”. In: *Transportation Research Part B: Methodological* 58 (2013), pp. 243–263.

[127] A. Golan et al. “A maximum entropy approach to recovering information from multinomial response data”. In: *Journal of the American Statistical Association* 91.434 (1996), pp. 841–853.

[128] X. Ma et al. “Mining smart card data for transit riders’ travel patterns”. In: *Transportation Research Part C: Emerging Technologies* 36 (2013), pp. 1–12.

[129] X. Zheng et al. “Big data for social transportation”. In: *IEEE Transactions on Intelligent Transportation Systems* 17.3 (2016), pp. 620–630.

[130] H. Akaike. “Information theory and an extension of the maximum likelihood principle”. In: *Breakthroughs in Statistics: Foundations and Basic Theory*. Ed. by S. Kotz and N. L. Johnson. Vol. 1. Springer Series in Statistics. Springer New York, 1992. Chap. 38, pp. 610–624.

[131] M. Keane. “Current issues in discrete choice modeling”. In: *Marketing Letters* 8.3 (1997), pp. 307–322.

[132] G. Menghini et al. “Route choice of cyclists in Zurich”. In: *Transportation Research Part A: Policy and Practice* 44.9 (2010), pp. 754–765.

[133] A. Sobhani et al. “Metropolis-Hasting based Expanded Path Size Logit model for cyclists’ route choice using GPS data”. In: *International Journal of Transportation Science and Technology* (2018). in press.

[134] L. Shen and P. R. Stopher. “Review of GPS travel survey and GPS data-processing methods”. In: *Transport Reviews* 34.3 (2014), pp. 316–334.

[135] C. R. Bhat. “The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models”. In: *Transportation Research Part B: Methodological* 45.7 (2011), pp. 923–939.

[136] A. Vij and K. Shankari. “When is big data big enough? Implications of using GPS-based surveys for travel demand analysis”. In: *Transportation Research Part C: Emerging Technologies* 56 (2015), pp. 446–462.

[137] Z. Ghahramani. “Probabilistic machine learning and artificial intelligence”. In: *Nature* 521.7553 (2015), p. 452.

[138] H. Wang and C. Zhai. “Generative Models for Sentiment Analysis and Opinion Mining”. In: *A practical Guide to Sentiment Analysis*. Ed. by E. Cambria et al. Vol. 5. Socio-Affective Computing. Springer International Publishing, 2017. Chap. 6, pp. 107–134.

[139] W. K. Hastings. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. In: *Biometrika* 57.1 (1970), pp. 97–109.

[140] T.-C. Lee et al. “New approach to modeling mixed traffic containing motorcycles in urban areas”. In: *Transportation Research Record: Journal of the Transportation Research Board* 2140 (2009), pp. 195–205.

[141] G. M. Allenby and P. E. Rossi. “Marketing models of consumer heterogeneity”. In: *Journal of Econometrics* 89.1 (1998), pp. 57–78.

[142] R. A. Daziano and D. Bolduc. “Incorporating pro-environmental preferences towards green automobile technologies through a bayesian hybrid choice model”. In: *Transportmetrica A: Transport Science* 9.1 (2013), pp. 74–106.

[143] L. Sun and A. Erath. “A Bayesian network approach for population synthesis”. In: *Transportation Research Part C: Emerging Technologies* 61 (2015), pp. 49–62.

[144] I. Saadi et al. “Hidden Markov Model-based population synthesis”. In: *Transportation Research Part B: Methodological* 90 (2016), pp. 1–21.

[145] N. Bhatnagar et al. “The computational complexity of estimating MCMC convergence time”. In: *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. Ed. by L. A. Goldberg et al. Vol. 6845. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2011, pp. 424–435.

[146] A. P. Dempster et al. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1977), pp. 1–38.

[147] C. R. Bhat. “An endogenous segmentation mode choice model with an application to intercity travel”. In: *Transportation science* 31.1 (1997), pp. 34–48.

[148] K. E. Train. “EM algorithms for nonparametric estimation of mixing distributions”. In: *Journal of Choice Modelling* 1.1 (2008), pp. 40–69.

[149] C. R. Bhat. “The multiple discrete-continuous extreme value (MDCEV) model: role of utility function parameters, identification considerations, and model extensions”. In: *Transportation Research Part B: Methodological* 42.3 (2008), pp. 274–303.

[150] J. Kim et al. “Modeling Consumer Demand for Variety”. In: *Marketing Science* 21.3 (2002), pp. 229–250.

[151] C. Anda et al. “Transport modelling in the age of big data”. In: *International Journal of Urban Sciences* 21.sup1 (2017), pp. 19–42.

[152] L. Sun and Y. Yin. “Discovering themes and trends in transportation research using topic modeling”. In: *Transportation Research Part C: Emerging Technologies* 77 (2017), pp. 49–66.

[153] D. M. Blei et al. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[154] T. Huynh et al. “Discovery of Activity Patterns using Topic Models”. In: *10th International Conference on Ubiquitous Computing*. ACM. 2008, pp. 10–19.

[155] S. Hasan and S. V. Ukkusuri. “Urban activity pattern classification using topic models from online geo-location data”. In: *Transportation Research Part C: Emerging Technologies* 44 (2014), pp. 363–381.

[156] I. Peled et al. "Model-Based Machine Learning for Transportation". In: *Mobility Patterns, Big Data and Transport Analytics.* Elsevier, 2019, pp. 145–171.

[157] A. Muralidharan et al. "Probabilistic graphical models of fundamental diagram parameters for simulations of freeway traffic". In: *Transportation Research Record* 2249.1 (2011), pp. 78–85.

[158] T. A. Wheeler and M. J. Kochenderfer. "Factor graph scene distributions for automotive safety analysis". In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC).* IEEE. 2016, pp. 1035–1040.

[159] D. P. Kingma et al. "Improved variational inference with inverse autoregressive flow". In: *Proceedings of the 30th Annual Conference on Neural Information Processing Systems* (Barcelona, Spain). Ed. by D. Lee et al. Vol. 29. Advances in Neural Information Processing Systems. Red Hook New York, 2016, pp. 4743–4751.

[160] B. Schwehn. "Using the Natural Gradient for training Restricted Boltzmann Machines". MA thesis. University of Edinburgh, 2010.

[161] S. Bekhor et al. "Evaluation of choice set generation algorithms for route choice models". In: *Annals of Operations Research* 144.1 (2006), pp. 235–247.

[162] Y. W. Teh and G. E. Hinton. "Rate-coded restricted boltzmann machines for face recognition". In: *Advances in neural information processing systems.* Vol. 13. 2001, pp. 908–914.

[163] M.-A. Côté and H. Larochelle. "An Infinite restricted boltzmann machine". In: *Neural Computation* 28.7 (2016), pp. 1265–1288.

[164] R. M. Neal. *Probabilistic Inference Using Markov Chain Monte Carlo Methods.* Tech. rep. CRG-TR-93-1. University of Toronto, 1993.

[165] M. Braun and J. McAuliffe. "Variational Inference for Large-Scale Models of Discrete Choice". In: *Journal of the American Statistical Association* 105.489 (2010), pp. 324–335.

[166] C. Bishop. *Pattern Recognition and Machine Learning.* 1st ed. Information Science and Statistics. Springer New York, 2006.

[167] D. J. C. MacKay, ed. *Information Theory, Inference and Learning Algorithms.* 1st ed. Cambridge University Press, 2003.

[168] G. Collell and J. Fauquet. “Brain activity and cognition: a connection from thermodynamics and information theory”. In: *Frontiers in psychology* 6 (2015), p. 818.

[169] D. J. C. MacKay. “Probable networks and plausible predictions–a review of practical Bayesian methods for supervised neural networks”. In: *Network: Computation in Neural Systems* 6.3 (1995), pp. 469–505.

[170] X.-H. Yu. “Can backpropagation error surface not have local minima”. In: *IEEE Transactions on Neural Networks* 3.6 (1992), pp. 1019–1021.

[171] K. He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 770–778.

[172] R. Krueger et al. “Variational Bayesian Inference for Mixed Logit Models with Unobserved Inter-and Intra-Individual Heterogeneity”. In: *arXiv preprint arXiv:1905.00419* (2019).

[173] J. F. P. da Costa et al. “The unimodal model for the classification of ordinal data”. In: *Neural Networks* 21.1 (2008), pp. 78–91.

[174] E. Cambria et al., eds. *A practical Guide to Sentiment Analysis.* Vol. 5. Socio-Affective Computing. Springer International Publishing, 2017.