



tinbergen
institute

Anoek Castelein

Models for Individual Responses

Explaining and predicting individual behavior

Erasmus Universiteit Rotterdam

MODELS FOR INDIVIDUAL RESPONSES:
EXPLAINING AND PREDICTING
INDIVIDUAL BEHAVIOR

ISBN: 978-90-361-0642-9

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul.

©2021, Aniek Castelein

All rights reserved. Save exceptions stated by the law, no part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, included a complete or partial transcription, without the prior written permission of the author, application for which should be addressed to the author.

This book is no. 775 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Models for Individual Responses: Explaining and predicting individual behavior

Modellen voor individuele reacties:
Het verklaren en voorspellen van individueel gedrag

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
Rector Magnificus

Prof.dr. F.A. van der Duijn Schouten

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on

Thursday March 18, 2021, at 13:00 hours

by

ANOEK CASTELEIN
born in Castricum, the Netherlands.

Doctorate committee

Promotors: Prof.dr. D. Fok
Prof.dr. R. Paap

Other members: Prof.dr. R.L. Lumsdaine
Prof.dr. M. Vandebroek
Prof.dr. M.G. de Jong

Acknowledgements

After five years of work, my thesis is completed. I'm thankful to have been given the opportunity to do a PhD. I've learned a lot, and have been able to dive into my research with much freedom. I've also been surrounded by incredible people, who have supported me during my research or with whom I could relax and enjoy my time.

The process of writing this thesis has been challenging. Countless times, things did not work out the way I thought they would. The main challenge was to find out why. Did the code have a bug? Were there unforeseen issues with the estimation approach? Or did everything work, except for my idea? I spent most of my time testing for bugs. I found out that excellent programming skills are crucial when doing research on developing models. Only when I learned a low-level language and started to program more like a programmer in my fourth year, I could quickly test code and estimate models. Had I been a better programmer, I may have been able to complete my PhD much earlier.

Numerous of people have supported me during the writing of this thesis. I'd like to thank them for their support. First and foremost, my promotors Dennis and Richard. Their support and availability kept me motivated, feel valued, and provided swift help when I needed it. Their domain knowledge helped to address any problem I had. The most valuable lessons I've learned from them are to be able to come up with own solutions to research problems, instead of solely relying on available solutions of others, and to be able to generalize methods.

I'd also like to thank the committee members for accepting a position in my committee and for their helpful comments and suggestions. I'm looking forward to the defense.

My time at the Erasmus University Rotterdam has been interesting and fun thanks to my friendly colleagues, whom I'd like to thank. My roommates Xiao (first year) and Rowan (the years after), with whom I could work in harmony and could share my everyday matters. My roommates from my fourth year onwards in the large PhD office: Daan, Jens, Jiawei, Kevin, Mathijs, Terri, Thomas and Thomas. My fellow PhD candidates, with whom I could blow off steam and enjoy conversations: Albert Jan, Didier, Esmée, Ilka, Indy, Karel, Malin, Matthijs, Max, Myrthe, Nienke, and many others. My colleagues with whom I cooperated in teaching. And my colleagues working at the secretariats of the Econometric and Tinbergen Institute, who always provided swift help.

Finally, I'd like to thank my family. My brothers Tim and Jeroen, who stand beside me as my paranympths. Our bond continues from our childhood in which we did so much together, to the present in which we share the important things in our lives. My family in law, who have supported me, have always been interested, and with whom I've shared many enjoyable moments: Marleen, Hans, Alfons, Nienke, Frank and Valentijn.

My parents Alice and Evert, who have supported me from my childhood onwards, and have provided a stimulating and safe environment. I'm thankful that they've always let me make my own choices. And when things didn't turn out the way I'd hoped, they were always there.

My beautiful children Mette and Lucas. With whom every day is a joy.

My love and my best friend, Luite. Without whom I would have never even thought about doing a PhD. Who was there all the times I was struggling with my research. With whom I can laugh every day, and can share all things in my life. I hope we will have many more adventures together.

Anoek Castelein
Castricum, January 2021

Table of contents

- 1 Introduction** **1**
- 1.1 Inferring individual responses 2
- 1.2 Illustrative example: car preferences 3
- 1.3 Contributions of thesis 6
- 1.4 Overview of thesis 7
- 1.5 Outlook 9

- 2 Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach** **11**
- 2.1 Introduction 11
- 2.2 Related literature 15
- 2.3 Methodology 18
 - 2.3.1 Inference 21
- 2.4 Monte Carlo study 23
 - 2.4.1 Results 25
- 2.5 Case study: multinomial logit model 31
 - 2.5.1 Results 32
 - 2.5.2 Out-of-sample performance 38
- 2.6 Conclusion 40
- 2.A MCMC sampler 41
 - 2.A.1 Draw c_i 43
 - 2.A.2 Draw Σ_q and μ_q 44
 - 2.A.3 Draw λ_{ik} 45
 - 2.A.4 Draw τ_{ik} 46
 - 2.A.5 Draw θ_k 47
 - 2.A.6 Draw γ 47

2.B	Histograms of priors	48
2.C	Hit rates Monte Carlo study	49
3	A multinomial and rank-ordered logit model with inter- and intra- individual heteroscedasticity	51
3.1	Introduction	51
3.2	Background	55
3.3	Methodology	57
3.3.1	Hidden Markov multinomial logit model	58
3.3.2	Hidden Markov rank-ordered logit model	60
3.3.3	Parameter estimation	62
3.4	Monte Carlo study	64
3.4.1	Results	66
3.5	Case study I: learning and fatigue during discrete choice experiments .	68
3.5.1	Results	71
3.6	Case study II: differential capabilities in ranking	76
3.6.1	Results	77
3.7	Conclusion	81
3.A	Maximum simulated likelihood estimation	83
3.A.1	Hidden Markov multinomial logit model	84
3.A.2	Hidden Markov rank-ordered logit model	85
3.A.3	Miscellaneous details	86
3.B	Conditional distribution of S_{it}	86
3.C	Monte Carlo study: results DGPs 4-6	88
4	A dynamic model of clickthrough and conversion probabilities of paid search advertisements	89
4.1	Introduction	89
4.2	Background	93
4.2.1	The mechanism underlying search engine advertising	93
4.2.2	Modeling clickthrough and conversion probabilities of keywords .	94
4.3	General structure of data	96
4.4	Methods	96
4.4.1	Model specification	97
4.4.1.1	Time-varying parameters: the dynamic impact of shocks .	99
4.4.1.2	Unobserved heterogeneity across keywords	100
4.4.1.3	Instrumental variables	100

4.4.2	Parameter identification	101
4.4.3	Bayesian inference	102
4.5	Empirical application	103
4.5.1	Data	103
4.5.2	Baseline results	104
4.5.3	Model comparison	110
4.6	Managerial implications	112
4.7	Summary and conclusions	113
4.A	Gibbs sampler	114
4.A.1	Overview Gibbs sampler	117
4.A.2	Priors	118
4.A.3	Initialization	118
4.A.4	Steps Gibbs sampler	119
4.A.4.1	Sampling Polya-Gamma variables ω	119
4.A.4.2	Sampling α_i , λ_i , and δ_i	119
4.A.4.3	Sampling β_t	120
4.A.4.4	Sampling γ	121
4.A.4.5	Sampling η	121
4.A.4.6	Sampling $\tilde{\alpha}$, $\tilde{\lambda}$, and $\tilde{\delta}$	122
4.A.4.7	Sampling Σ_α , Σ_λ , and Σ_δ	123
4.A.4.8	Sampling Φ	123
4.A.4.9	Sampling Σ_β	124
4.A.4.10	Sampling Σ_η	124
5	Conclusions	125
	References	127
	Abstract in Dutch	137
	About the author	139
	Portfolio	141

Chapter 1

Introduction

We make numerous choices in our lives. From buying a house or choosing an occupation, to choosing which groceries to buy in the supermarket. Public and private organizations increasingly collect and store information on these choices. They use the collected data to help design effective policies. For example, health care providers use information on individual health outcomes to increase the effectiveness of their health treatments. Supermarkets use information on individual purchases to optimize their pricing and promotion strategies.

Going from data to the design of an effective policy is not straightforward. A first insightful step is to examine the average response in a population to a certain policy or policy change. How does a change in price affect the quantity sold? What percentage of patients is cured using a certain health treatment? Knowledge of the average response can lead to accurate predictions of aggregate outcomes of interest that aid the design of effective policies.

To further increase the effectiveness of policies, it is often useful to acknowledge and account for differences across individuals (*heterogeneity*). For example, individuals may respond differently to price changes, or their health may respond differently to certain health treatments. Using an individual-level approach instead of a population-level approach has two key advantages: (i) it helps to gain insight into how different individuals respond and thus gives insight into the complete distribution of policy effects, and (ii) predictions of individual responses can be used for policies that allow for personalization, such as personalized health treatments,

education, or marketing.

In this thesis, I develop approaches to accurately infer individual responses from data. A response refers to the effect of a change in a certain factor (e.g. price) on an individual's choice (e.g. a purchase decision). The developed approaches improve upon existing approaches by allowing for more realistic individual behavior. These improvements can lead to better predictions of individual responses and can therefore be used to gain a better understanding of individual behavior and to design more effective policies. The approaches in this thesis are aimed to be generally applicable to many real-life problems, including problems in health, education, labor, operations research, and consumer choice-making.

1.1 Inferring individual responses

To infer the responses of individuals from data, researchers often make use of a model. A model describes a relationship between the outcome of interest (e.g. heals or not, buys a product or not) and the observed explanatory variables/factors (e.g. type of health treatment, price of a product). This relationship depends on unknown parameters, which include the individual responses, that are to be estimated using the observed data. The parameters represent numerical values indicating the signs and strengths with which an individual's outcome responds to changes in the explanatory variables. When the parameters have been accurately estimated, one can examine what the impact is of a change in an explanatory variable on the outcome of interest.

To allow for individual differences in the responses, the unknown parameters in the model should be made individual-specific. One approach to do so is by considering a separate model for each individual, and estimating the (individual-specific) model parameters using data from that individual alone. In practice, this approach works poorly, especially in settings with a relatively small number of observations per individual and/or with many explanatory variables that may affect the outcome of interest. In these cases, using a separate model for each individual can lead to inaccurate and highly uncertain estimates of an individual's responses.

Instead of using a separate model for each individual, it is often more useful to use a model that shares information across individuals. That is, the model still contains individual-specific parameters, but for inference on those parameters, one uses information on the *underlying population distribution* of the parameters (or the individual responses). This is an approach often used by researchers, and it is also

the approach that I focus on in this thesis.

The underlying population distribution of individual responses describes how the responses across individuals differ. For example, a specific medicine may have a positive effect on a certain health measurement for 50% of individuals, a negative effect for 20% of individuals, and no effect for 30% of individuals. Moreover, the effect may be more positive (or negative) for some individuals than for others.

Hence, the researcher tries to most accurately estimate the underlying distribution of responses. For this purpose, the researcher uses the data from all individuals in the dataset. Using the information on the response distribution, a researcher can more accurately infer per individual where s/he most likely is in the distribution based on the observed data of that specific individual.

The underlying population distribution is usually of a high dimensionality, as in many settings there are quite a number of explanatory variables that may affect the outcome of interest. The response distribution has to jointly consider the responses to all (combinations of) variables. Estimating the shape of the resulting multivariate distribution is therefore not straightforward.

Thus, in many settings, the main challenge when inferring individual responses is to accurately estimate the underlying population distribution of responses. This is the challenge I propose solutions to in this thesis.

1.2 Illustrative example: car preferences

To illustrate how a model can be used to infer individual responses, consider the following example on data from a specific kind of questionnaire: a discrete choice experiment. During a discrete choice experiment, individuals are repeatedly asked to make a (hypothetical) choice amongst a set of alternatives. Each alternative is described by a number of attributes. Data from a discrete choice experiment are also used in this thesis to illustrate several approaches, although the developed approaches are more generally applicable to other types of data.

Suppose we are interested in the preferences of individuals for different types of cars, in particular in the tradeoffs an individual makes when choosing between a gasoline-powered and an electric private lease car. These preferences can be used by governmental institutions to design policies that promote the private lease of an electric car, by car manufacturers to design a desirable electric car and predict the

car's demand, and by car sellers to gain insight into which (types of) individuals would be interested in a specific car as to enable personalized marketing.

A discrete choice experiment can be conducted to elicit the preferences of individuals in the private lease market. During this experiment, individuals can be asked to complete 10 to 15 choice tasks where at each task an individual is asked to choose between two cars, one electric car and one gasoline-powered car: “*If you were in the market to private lease a car, and these were the only alternatives, which would you choose?*”. The cars are described by attributes such as price, average range, and size/luxury. The levels of the attributes vary over the tasks. An example of a choice task is given in Table 1.1.

Table 1.1: Example of a choice task during the discrete choice experiment.

If you were in the market to private lease a car, and these were the only two alternatives, which would you choose?

Attributes	Car 1 Gasoline	Car 2 Electric
Monthly price (in Euros)	250	300
Average fuel price per 100 km (in Euros)	10	5
Average range full battery	-	300 km
CO2 emissions	119 gr/km	-
Segment*	D	B
Option I: cruise control	✓	✓
Option II: leather seats	✓	-

*Each car belongs to one of fourteen segments. A segment indicates the size and class of a car.

Note that in this thesis, I focus on inferring the preferences of individuals *given* the answers to the discrete choice experiment. I do not focus on the design of an (optimal) experiment.

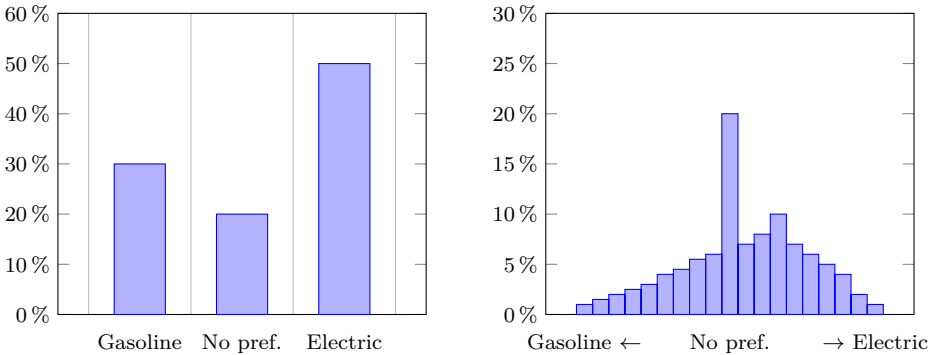
To infer the preferences of the individuals based on the choices made during the experiment, a model is used. This model describes how the combination of attributes of the two cars affects an individual's choice. Given the functional form of the model (that is, the manner in which the attributes may affect the choice), the only challenge remains to estimate the unknown individual-specific model parameters. These parameters correspond to the preferences of individuals for the attributes.

Because of the small number of choice tasks completed by each individual (10 to 15) and the relatively large number of car attributes, using a model that shares

information over individuals is useful. Hence, the interest becomes in estimating the underlying (multivariate) distribution of preferences for the different car attributes.

To illustrate the preference distribution for a single attribute, consider the preferences for individuals for choosing between an electric and gasoline-powered car, for *given* levels of the price, range, CO2 emissions, segment, and options. For any specific set-up, some individuals may prefer an electric car, some may prefer a gasoline-powered car, and some may have no clear preference of one type of car over the other. These differences could be for a number of reasons, e.g. due to environmental reasons or the availability of charging stations close to home. Suppose that, for certain given levels of the other attributes, 30% of individuals prefers gasoline-powered cars, 50% prefers electric cars, and 20% has no preference of one type of car over the other. Then, the corresponding preference distribution is given on the left in Figure 1.1a.

Figure 1.1: Examples of underlying population distributions for preferences for gasoline-powered versus electric cars.



(a) Distribution about sign of preferences (b) Distribution about strength of preferences

Of course, the distribution on the left in Figure 1.1a is not useful for inference as it only says something about the “sign” of the preferences for fixed levels of the other attributes: gasoline, neutral or electric. For inference, one also needs information on how strong this preference is: there may be individuals who really prefer an electric car, and those who only prefer it a bit as compared to a gasoline-powered car. These so-called ‘weights’ assigned to attributes are important when examining the relative importance of the different attributes. For example, individuals that assign just a small positive weight to an electric car can easily be persuaded to opt for a gasoline-powered car when the price becomes a bit lower or more options are added. Individuals with a strong preference for an electric car will not as quickly opt

for a gasoline-powered car.

When considering the weights assigned to an attribute, the true distribution of preferences for given levels of the other attributes may be more dispersed and more similar to the distribution on the right in Figure 1.1b. In this distribution, individuals that are in the right tail really prefer electric cars over gasoline-powered cars. Individuals that are closer to the zero weight (no preference), are more indifferent between the two cars.

Next to the preference for the fuel type, a researcher has to simultaneously infer the preferences for the other car attributes. Hence, instead of a distribution as in Figure 1.1b, one obtains a multivariate distribution of much higher dimensionality.

In this thesis, I develop approaches that can accurately estimate the underlying (multivariate) preference distribution. More specifically, I develop an approach that allows for distributions as on the right in Figure 1.1b: individuals may have widely ranging preferences *and* subsets of individuals may be indifferent between certain attribute levels. Moreover, I develop an approach that can accurately infer responses of individuals when (some) individuals become fatigued during the experiment and answer more randomly as the experiment proceeds.

1.3 Contributions of thesis

In this thesis, I develop approaches to accurately estimate the underlying population distribution of individual responses. In the existing literature, a number of approaches have already been developed. These approaches can be quite restrictive in the shape they allow for the underlying distribution. In this thesis, I aim to alleviate a number of important restrictions and provide approaches that allow for more realistic individual behavior. The proposed approaches can lead to improved estimates of individual responses which can be used to gain insight into individual behavior and to design more effective policies.

This thesis reports the developed approaches in three different chapters. The chapters can be read separately from each other. In Chapter 2, an approach is developed that allows for many forms of the underlying (multivariate) distribution of individual responses. In particular, the approach allows for groups of individuals to be unaffected by certain variables *and* for the individuals that are affected by the variables to have widely ranging responses. The proposed approach is generally applicable to problems in a wide range of research fields.

In Chapter 3, an approach is developed to accurately infer individuals' preferences by correcting for possible biases that may arise due to dynamics in the randomness in the choice-making of individuals. In the context of the earlier example on car preferences, this approach can correct for learning and fatigue behavior: at the beginning of the questionnaire some individuals may answer more randomly as they still need to learn about the choice task at hand or about their preferences (*learning*) or as the questionnaire proceeds some individuals may start answering more randomly as they become bored, tired, or irritated (*fatigue*).

In Chapter 4, an approach is developed that allows for individual responses to change over time, for example due to changing preferences or changing environments. This approach is tailored to one specific application: that of accurately estimating and predicting clickthrough and conversion probabilities of paid search advertisements at search engines.

1.4 Overview of thesis

A more detailed summary of the three chapters in this thesis is provided below. The work in Chapters 2 and 3 has been done mostly independently, under close supervision of mentioned co-authors. The original ideas were my own, and further developed in discussion with the co-authors. The implementation and reporting of the research was mostly done independently, a number of improvements were made through feedback on earlier versions of the chapters and discussions with the co-authors. The work in Chapter 4 has been done in close collaboration with the mentioned co-authors.

Chapter 2: A. Castelein, D. Fok and R. Paap: Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach.

In Chapter 2, we develop a general method for heterogeneous variable selection in Bayesian nonlinear panel data models. Heterogeneous variable selection refers to the possibility that subsets of units are unaffected by certain variables. It may be present in applications as diverse as health treatments, consumer choice-making, macroeconomics, and operations research. Our method additionally allows for other forms of cross-sectional heterogeneity. We consider a two-group approach for the model's unit-specific parameters: each unit-specific parameter is either equal to zero (heterogeneous variable selection) or comes from a Dirichlet process (DP) mixture of multivariate normals (other cross-sectional heterogeneity). We develop our approach for general nonlinear panel data models, encompassing multinomial logit and probit

models, poisson and negative binomial count models, exponential models, among many others. For inference, we develop an efficient Bayesian MCMC sampler.

In a Monte Carlo study, we find that our approach is able to capture heterogeneous variable selection whereas a “standard” DP mixture is not. In an empirical application, we find that accounting for heterogeneous variable selection and non-normality of the continuous heterogeneity leads to an improved in-sample and out-of-sample performance and interesting insights. These findings illustrate the usefulness of our approach.

Chapter 3: A. Castelein, D. Fok and R. Paap: A multinomial and rank-ordered logit model with inter- and intra-individual heteroscedasticity.

The heteroscedastic logit model is useful to describe choices of individuals when the randomness in the choice-making varies over time. For example, during surveys individuals may become fatigued and start responding more randomly to questions as the survey proceeds. Or when completing a ranking amongst multiple alternatives, individuals may be unable to accurately assign middle and bottom ranks. The standard heteroscedastic logit model accommodates such behavior by allowing for changes in the signal-to-noise ratio via a time-varying scale parameter. In the current literature, this time-variation is assumed equal across individuals. Hence, each individual is assumed to become fatigued at the same time, or assumed to be able to accurately assign exactly the same ranks. In most cases, this assumption is too stringent.

In Chapter 3, we generalize the heteroscedastic logit model by allowing for differences across individuals. We develop a multinomial and a rank-ordered logit model in which the time-variation in an individual-specific scale parameter follows a Markov process. In case individual differences exist, our models alleviate biases and make more efficient use of data. We validate the models using a Monte Carlo study and illustrate them using data on discrete choice experiments and political preferences. These examples document that inter- and intra-individual heteroscedasticity both exist.

Chapter 4: A. Castelein, D. Fok and R. Paap: A dynamic model of clickthrough and conversion probabilities of paid search advertisements.

In Chapter 4, we develop a dynamic Bayesian model for clickthrough and conversion probabilities of paid search advertisements. These probabilities are subject to changes over time, due to e.g. changing consumer tastes or new product launches. Yet, there is little empirical research on these dynamics. Gaining insight into the dy-

namics is crucial for advertisers to develop effective search engine advertising (SEA) strategies. Our model deals with dynamic SEA environments for a large number of keywords: it allows for time-varying parameters, seasonality, data sparsity and position endogeneity. The model also discriminates between transitory and permanent dynamics. Especially for the latter case, dynamic SEA strategies are required for long-term profitability.

We illustrate our model using a 2 year dataset of a Dutch laptop selling retailer. We find persistent time variation in clickthrough and conversion probabilities. The implications of our approach are threefold. First, advertisers can use it to obtain accurate daily estimates of clickthrough and conversion probabilities of individual ads to set bids and adjust text ads and landing pages. Second, advertisers can examine the extent of dynamics in their SEA environment, to determine how often their SEA strategy should be revised. Finally, advertisers can track ad performances to timely identify when keywords' performances change.

1.5 Outlook

The approaches developed in this thesis can prove useful to practitioners in a wide range of research fields. In particular, they can be used to gain insight into the differences in policy effects across individuals, and to obtain accurate individual-level predictions that enable personalizing certain policies. For future methodological research, it would be interesting to examine approaches that allow for more flexible forms of changing behavior of individuals over time, in particular in settings with relatively little information per individual.

Chapter 2

Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach

2.1 Introduction

Many panel datasets contain information on a large number of cross-sectional units with relatively little information per unit. Such datasets contain too little information to accurately estimate a separate model per unit: estimation inefficiency and overfitting would become problematic. Performing variable selection at the unit-level is therefore not straightforward. Instead, models are used that share information across units. To this end, unit-specific parameters in the model are often shrunk using an underlying population distribution shared across units. Many such distributions have been proposed: continuous distributions such as the multivariate normal or log-normal, finite mixtures of discrete or continuous distributions, and ‘infinite’ mixtures using a Dirichlet process.

In practice, these distributions cannot sufficiently accommodate heterogeneous variable selection on top of other cross-sectional heterogeneity. Heterogeneous variable selection refers to the possibility that subsets of units may be unaffected by certain variables. This is relevant for many applications. For example, in choice situations, groups of individuals may have no preference for or may ignore a certain product attribute when making their decisions. In macroeconomics, unemployment rates in different countries may be differentially affected or unaffected by certain macroeconomic variables. In operations research, the interarrival times of buses or the amount of garbage in bins could differentially depend or not depend on variables as temperature, holidays, or traffic conditions.

We use the term *variable selection* to denote that some units assign no weight to certain variables. Hence, variable selection is part of the data generating process. This is different from the context where variable selection refers to a researcher determining which variables should be selected in a model, also known as *model selection*. Instead of using *variable selection*, other appropriate terms are *variable importance* or *variable relevance* to indicate that for some units, certain variables may be unimportant or irrelevant.

While the literature on modeling heterogeneous responses is extensive, very few approaches have been proposed that accommodate heterogeneous variable selection. That is, the underlying population distribution to which the unit-specific parameters are shrunk, generally does not allow for groups of units to assign no weight to certain variables. Theoretically, heterogeneous variable selection can be captured when the underlying distribution is discrete, such as with a latent class approach. A discrete distribution allows the unit-specific parameters to be equal to one of multiple multivariate discrete outcomes, of which some outcomes may have certain parameters equal to zero. Practically, such a model is infeasible as the discrete distribution would need 2^K possible outcomes to capture all combinations of variable selection, where K is the number of explanatory variables. If, additionally, richer forms of heterogeneity should be allowed for, a multitude of these 2^K outcomes is needed.¹ In models with continuous heterogeneity it is even more problematic to accommodate heterogeneous variable selection, as the continuous heterogeneity distribution cannot have substantial mass at zero unless the variance of the distribution is very close to zero.

¹Alternatively, one could allow for the responses to the different variables to be independent, to avoid needing at least 2^K outcomes. However, this assumption of independence can be too strict.

A number of papers have proposed approaches to accommodate heterogeneous variable selection. They have done so for multivariate linear models (S. Kim et al., 2009, Tang et al., 2020), multivariate binary probit models (S. Kim et al., 2018), and multinomial logit models (Gilbride et al., 2006, Scarpa et al., 2009, Hensher and Greene, 2010, Hole, 2011, Campbell et al., 2011, Hess et al., 2013, Hole et al., 2013, Collins et al., 2013, Hensher et al., 2013). Few of these papers use a Bayesian approach (Gilbride et al., 2006, S. Kim et al., 2009, S. Kim et al., 2018). The papers that use a frequentist approach have strong limitations: when allowing for flexible forms of cross-sectional heterogeneity next to heterogeneous variable selection, the developed models are susceptible to overfitting as the number of parameters quickly grows large relative to the number of observations. Furthermore, the computation time for estimation grows rapidly when the number of variables gets larger, due to the likelihood function containing 2^K terms and, in case of a continuous heterogeneity distribution, the needed use of simulated maximum likelihood due to intractable integrals. Already when there are more than four variables, these approaches can run into problems.² To avoid overfitting, Tang et al. (2020) use a penalization framework. They propose a linear model where each unit-specific parameter comes from a univariate discrete distribution with multiple possible outcomes of which one outcome is set to zero. The parameters of the discrete distributions are estimated by optimizing a penalized objective function. The idea of their approach can also be used for nonlinear models, but, in practice, the use of multiple univariate discrete distributions is too limited to capture the possible rich forms of heterogeneous responses, for example correlations across the responses to different variables.

The few papers that use a Bayesian approach also have their limitations. They are limited in terms of the underlying parametric model: only techniques for heterogeneous variable selection in the context of a multivariate linear, a multinomial logit, and a binary probit model have been proposed. Furthermore, the form of cross-sectional heterogeneity including heterogeneous variable selection is limited in these papers. S. Kim et al. (2018) let the unit-specific parameters come from a categorical distribution that simultaneously incorporates variable selection and other heterogeneity. S. Kim et al. (2009) follow a similar approach but instead consider a categorical distribution with an ‘infinite’ number of outcomes using a Dirichlet process prior. As with standard heterogeneous response models with discrete heterogeneity, the main

²In Hensher et al. (2013) the problem of estimation time is explicitly stated in footnote 5: it took over 100 hours to estimate the parameters based on a dataset with 588 units, 16 observations per unit and 4 variables that were allowed to be ignored.

drawback of the approaches of S. Kim et al. (2018) and S. Kim et al. (2009) is that the number of outcomes of the categorical distribution that is necessary to capture all combinations of variable selection is exponential in the number of explanatory variables. In practice, it is hard to find that many components.

A more parsimonious approach is developed in Gilbride et al. (2006), who let each unit-specific parameter be either equal to zero *or* come from an underlying multivariate normal distribution. However, this single multivariate normal distribution can be insufficient to describe the complex forms of unit-specific responses. Furthermore, the Markov chain Monte Carlo (MCMC) sampler that Gilbride et al. (2006) propose for posterior results can be computationally heavy when there are many variables, as in each MCMC iteration a likelihood function with 2^K terms has to be computed. Moreover, the MCMC sampler uses the prior distribution as candidate for drawing the unit-specific parameters. In case the data is quite informative, this candidate will have low acceptance rates and the sampler will have poor mixing.

In this paper, we generalize and improve the approach of Gilbride et al. (2006), thereby contributing to the literature in three important ways: by (i) generalizing to nonlinear models, (ii) substantially increasing the flexibility in the cross-sectional heterogeneity, and (iii) developing an efficient Bayesian MCMC sampler that also works well for up to 50 or 100 explanatory variables. The increased flexibility is obtained by augmenting the heterogeneous variable selection with an infinite mixture of multivariate normals using a Dirichlet process (DP) prior.

To be more precise, we develop a general method for heterogeneous variable selection in Bayesian nonlinear panel data models. For the model's unit-specific parameters we take a two-group approach: each unit-specific parameter is either zero or comes from a DP mixture of multivariate normals. In case of a single unit-specific parameter, such a two-group approach is referred to as a spike-and-slab prior (Mitchell & Beauchamp, 1988) or as stochastic search variable selection (SSVS) (George and McCulloch, 1993, George and McCulloch, 1997). We develop our approach for general nonlinear panel data models, encompassing multinomial logit and probit models, poisson and negative binomial count models, exponential models, among many others. The model is particularly useful in large N , small T settings, but can also be incorporated in large T settings because of the flexibility of the DP mixture.

We illustrate our approach with a Monte Carlo study and an empirical application. For illustration, we consider a multinomial logit model (MNL) as this model is the focus of most of the literature on heterogeneous variable selection. In the Monte

Carlo study, we find that with our approach we can capture both complex forms of continuous cross-sectional heterogeneity — such as skewness and multimodality — as well as heterogeneous variable selection. When using only a ‘standard’ DP mixture for the unit-specific parameters, we find that heterogeneous variable selection cannot be accommodated. Instead of a spike at zero, this approach generally allocates substantial probability mass to parameter values in a relatively large interval around zero, depending on the shape of the true continuous heterogeneity distribution.

In the empirical application, we consider responses to a discrete choice experiment on food choices. We find substantial evidence of variable non-attendance and non-normality of the continuous heterogeneity. In particular, the continuous heterogeneity distribution seems skewed. Hence, there seem to be quite some individuals that have strong preferences for certain attributes, and quite some individuals that ignore certain attributes. These findings indicate the usefulness of our approach in practice.

The setup of this paper is as follows. In Section 2.2, we discuss the related literature. In Section 2.3, we develop our approach for general nonlinear panel data models. We also provide the Bayesian MCMC sampler. In Sections 2.4 and 2.5, we discuss the results of our model for a small Monte Carlo study and an empirical application, respectively. In Section 2.6, we conclude.

2.2 Related literature

An overview of papers that develop approaches to accommodate heterogeneous variable selection in panel data models is given in Table 2.1. These papers mainly differ in (i) the type of model they develop (logit, probit, linear, et cetera), (ii) how they incorporate heterogeneous variable selection, (iii) how they deal with cross-sectional heterogeneity other than heterogeneous variable selection, and (iv) if and how they incorporate correlated variable selection.

Heterogeneous variable selection is mostly incorporated using a two-group approach (SSVS, spike-and-slab, latent class). The frequentist approaches rely on latent class techniques (or a categorical distribution) for the unit-specific parameters. That is, these approaches specify 2^K classes where in each class a different combination of variables is selected, i.e. a different combination of parameters are set to zero. Each unit belongs to one of the 2^K classes. For the unit-specific parameters that are not zero, the approaches either restrict them to be equal over units (*constant*), allow them to differ depending on the class the unit is in (*categorical*), or let them be

Table 2.1: Overview of papers that develop approaches to accommodate heterogeneous variable selection.

Paper	Model	Het. variable selection	Additional cross-sectional heterogeneity	Correlated selection*
<i>Frequentist</i>				
Scarpa et al. (2009)	MNL	Latent class	Constant	Partly correlated
Hensher & Green (2010)	MNL	Latent class	Categorical	Partly correlated
Hole (2011)	MNL	Latent class	Constant	Partly correlated
Campbell et al. (2011)	MNL	Latent class	Categorical	Partly correlated
Hess et al. (2013)	MNL	Latent class	Multivariate normal	Uncorrelated
Hole et al. (2013)	MNL	Latent class	Multivariate normal	Partly correlated
Collins et al. (2013)	MNL	Latent class	Multivariate normal	Partly correlated
Hensher et al. (2013)	MNL	Latent class	Multivariate normal per latent class	Fully correlated
Tang et al. (2020)	Linear	Penalty	Categorical per variable	Uncorrelated
<i>Bayesian</i>				
Gilbride et al. (2006)	MNL	SSVS	Multivariate normal	Uncorrelated
Kim et al. (2009)	Linear	Spike-and-slab**	Categorical (infinite # of outcomes)	Uncorrelated
Kim et al. (2018)	Probit	Spike-and-slab**	Categorical	Uncorrelated
This paper	General	SSVS	Infinite mixture of multivariate normals	Uncorrelated

* The partly correlated methods are based on either considering only a subset of variables to be ignored together or letting the membership probabilities being a function of unit-specific variables.

** In S. Kim et al. (2009) and S. Kim et al. (2018), the underlying distribution for the unit-specific parameters incorporates heterogeneous variable selection within the categorical distribution that governs other cross-sectional heterogeneity.

independent of the class a unit is in and let them come from an underlying multivariate normal distribution. Exceptions are Campbell et al. (2011) who use a single multivariate normal and additionally allow for a different scale parameter per class, and Hensher et al. (2013) who allow for a different multivariate normal per class. The Bayesian approaches rely on a spike-and-slab prior or stochastic search variable selection (SSVS). That is, when a variable is ignored/unselected, the corresponding unit-specific parameter is either zero (spike-and-slab prior) or comes from a distribution closely centered around zero (SSVS). Within the Bayesian approaches, S. Kim et al. (2009) and S. Kim et al. (2018) incorporate heterogeneous variable selection within the categorical distribution that describes other cross-sectional heterogeneity. In contrast, Gilbride et al. (2006) let these two types of heterogeneous responses be independent: a unit-specific parameter is either zero or comes from a separate multivariate normal distribution. Our approach is most similar to Gilbride et al. (2006). We extend upon their approach by generalizing to nonlinear models and using a Dirichlet process mixture of multivariate normals for the other heterogeneity to realistically capture differences across units. Moreover, we improve upon their MCMC sampler to allow the approach to be used for up to 50 or 100 explanatory variables.

Alternatively to the two-group approach, Tang et al. (2020) use a penalization framework to shrink the unit-specific parameters towards zero or towards a specific value out of a set of outcomes to be estimated. Similar penalization frameworks for heterogeneous variable selection are employed in image and video classification problems, see e.g. Wu et al. (2012) and Zhao et al. (2015), where the used term is often *heterogeneous feature selection* or *sparsification*. In contrast to the approach developed in Tang et al. (2020), these latter approaches shrink the corresponding unit-specific parameter to zero in case a variable is selected, and not to some underlying population distribution shared across units.

Another main difference between the available approaches for heterogeneous variable selection is if and how they deal with correlated variable selection. Correlated variable selection refers to the phenomenon that some variables may be more likely to be selected/ignored together. This correlation can be divided into explained correlation (using observed unit-specific variables) and unexplained correlation. Most of the papers on heterogeneous variable selection do not allow for correlated variable selection. The ones that do can be divided into three groups: (i) letting each class/component have its own membership probability causing the number of membership probability

parameters to be exponential in the number of explanatory variables (Hensher et al., 2013), (ii) allowing for variable selection and correlation only across predefined subsets of variables (Scarpa et al., 2009, Hensher and Greene, 2010, Campbell et al., 2011 and Collins et al., 2013), or (iii) letting the class membership probabilities be a function of unit-specific variables (Hole, 2011, Hole et al., 2013). In this paper, we do not explicitly allow for correlated variable selection. However, our approach can be extended to allow for both explained and unexplained correlated variable selection.

Approaches have also been developed that use a DP mixture for cross-sectional heterogeneity, and *aggregate* variable selection to analyze which variables should not be in the model for all units (see e.g. Cai and Dunson, 2005 and M. Yang, 2012). Furthermore, related approaches have been developed for models that do not include unit-specific parameters: the combination of a DP mixture and variable selection are used for a set of pooled parameters. These approaches are often used in settings with many explanatory variables to shrink coefficients towards zero (variable selection) or each other (DP mixture), both in supervised problems (see e.g. Dunson et al., 2008, MacLehose et al., 2007, and Korobilis, 2013) and unsupervised clustering problems (see e.g. S. Kim et al., 2006, Wang and Blei, 2009, Yu et al., 2010, Fan and Bouguila, 2013).

2.3 Methodology

In this section, we develop our approach to simultaneously allow for heterogeneous variable selection and other flexible forms of cross-sectional heterogeneity in nonlinear panel data models. We provide the model specification and the details of the MCMC sampler to obtain posterior samples.

We consider a dataset with N cross-sectional units and T_i observations for unit $i = 1, \dots, N$. The interest is in modeling a scalar dependent random variable Y_{it} in terms of observed explanatory variables in x_{it} and z_{it} for unit i at time t . The responses to the variables in the $(K_x \times 1)$ vector x_{it} are assumed unit-specific and captured in the $(K_x \times 1)$ parameter vector β_i . For identification, x_{it} may contain time-varying variables only, other than an intercept.³ The responses to the variables in the $(K_z \times 1)$ vector z_{it} are assumed equal across units and captured in the $(K_z \times 1)$ parameter vector γ . The variables in x_{it} and z_{it} cannot overlap.

³We recommend to mean center any continuous variable in x_{it} . Furthermore, for multinomial models, instead of a single intercept, x_{it} may contain an intercept per possible outcome for Y_{it} , minus one, or other time-invariant alternative-specific variables.

We consider a nonlinear model for Y_{it} as given by

$$Y_{it}|\beta_i, \gamma \sim f(g(x_{it}, \beta_i, z_{it}, \gamma)), \quad (2.1)$$

where f is a known continuous or discrete probability distribution, g is a known (possibly multivariate) deterministic link function that maps x_{it} , β_i , z_{it} and γ to the parameters of the probability distribution, and we assume the observations Y_{it} to be conditionally independent over units and time periods.

For example, for multinomial data such as discrete choices, f could represent a multinomial distribution with size 1 and probability vector $p_{it} = g(x_{it}, \beta_i, z_{it}, \gamma)$ based on e.g. the softmax link function to obtain a multinomial logit model. For count data, f could represent a Poisson or negative binomial distribution with parameters $g(x_{it}, \beta_i, z_{it}, \gamma)$. Continuous distributions may also be used, such as the normal or the exponential distribution. We take the distribution $f()$ and the link function $g()$ as given.

The parameters in β_i capture the responses of unit i to the variables in x_{it} . To allow for flexible forms of cross-sectional heterogeneity, we take

$$\beta_{ik} = \tau_{ik}\lambda_{ik}, \quad (2.2)$$

for $k = 1, \dots, K_x$. Heterogeneous variable selection is captured in the latent indicator τ_{ik} which indicates whether variable k is selected by unit i and, if selected, lets β_{ik} be equal to λ_{ik} which follows an infinite mixture of multivariate normals distribution using a Dirichlet process prior. We take $\tau_{ik} \in \{\kappa, 1\}$, where κ is zero or close to zero and is set by the researcher. In case $\kappa = 0$, we obtain a spike-and-slab prior, in case $\kappa \neq 0$ but close to zero our approach becomes an example of stochastic search variable selection. For estimation efficiency, it is not necessary to set $\kappa \neq 0$. Hence, for interpretation it may be most suitable to set $\kappa = 0$.

We assume the variable selection indicator (τ_{ik}) to be independent of λ_{ik} . The probability that unit i selects variable k is denoted by

$$\Pr[\tau_{ik} = 1|\theta_k] = \theta_k, \quad (2.3)$$

with $0 \leq \theta_k \leq 1$, for $k = 1, \dots, K_x$.⁴

⁴One can allow for explained correlated variable selection using unit-specific probabilities θ_{ik} that are a deterministic function of unit-specific variables.

For flexible continuous heterogeneity, we let $\lambda_i = (\lambda_{i1}, \dots, \lambda_{iK_x})'$ come from an infinite mixture of multivariate normals using the DP prior (Ferguson et al., 1974, Antoniak, 1974, Rossi, 2014). The mixture for λ_i is given by

$$\lambda_i | \{\pi_q\}_q, \{\mu_q\}_q, \{\Sigma_q\}_q \sim \sum_{q=1}^{\infty} \pi_q MVN(\mu_q, \Sigma_q), \quad (2.4)$$

where π_q indicates the component membership probability of component q , μ_q denotes component's q mean, and Σ_q denotes component's q covariance matrix. The DP prior puts a prior on the mixture parameters π_q , μ_q and Σ_q . The DP prior has two hyperparameters: a tightness parameter α and a base distribution G_0 that invoke the following priors on π_q , μ_q and Σ_q

$$\pi_q = \eta_q \prod_{r=1}^{q-1} (1 - \eta_r), \quad \eta_q \sim \text{Beta}(1, \alpha), \quad (2.5)$$

$$\mu_q, \Sigma_q \sim G_0 \equiv p(\mu_q, \Sigma_q), \quad (2.6)$$

for $q = 1, 2, \dots$, where the base distribution G_0 of the DP is the prior distribution $p(\mu_q, \Sigma_q)$. This representation of the DP mixture is known as the stick-breaking representation (Rossi, 2014).

The component membership probabilities π_q are completely governed by the tightness parameter α . The specification implies that π_q declines as the component indicator q increases. The larger α , the more mass the Beta distribution has at zero. Hence, the larger α , the smaller we expect the η_q 's for the first components to be, and the more components we expect to have reasonably large membership probabilities. Given that there are N units, at most N unique components can be identified from the data.

For the base distribution, we take the conjugate prior $p(\mu_q, \Sigma_q) = p(\mu_q | \Sigma_q) p(\Sigma_q)$ as given by

$$p(\mu_q | \Sigma_q) = MVN(\mu_0, d^{-1} \Sigma_q), \quad (2.7)$$

$$p(\Sigma_q) = IW(\nu, \nu v I). \quad (2.8)$$

This conjugate prior allows for efficient estimation. The hyperparameter ν affects the variances of the components: a large ν puts substantial prior mass on components with 'large' variance, whereas a small ν puts substantial prior mass on components

with ‘small’ variance (Rossi, 2014).

Finally, for γ and θ_k we take the following priors

$$p(\gamma) = MVN(\gamma_0, \Sigma_\gamma), \quad (2.9)$$

$$p(\theta_k) = Beta(a, b), \quad \text{for } k = 1, \dots, K_x. \quad (2.10)$$

The hyperparameters α , μ_0 , d , v , ν , γ_0 , Σ_γ , a and b should either be set by the researcher or should have a prior itself. The proposed approach for heterogeneous responses is particularly useful in large N , small T settings, but can also be incorporated in large T settings because of the flexibility of the DP mixture.

As a final remark, we note that one may wish to restrict the variable selection to hold for multiple variables simultaneously. For example, in case one includes different levels of the same categorical variable through multiple dummy variables, one may want the variable selection to hold for all levels of that categorical variable. More formally, some of the elements in $\tau_i = (\tau_{i1}, \dots, \tau_{iK_x})'$ should be allowed to be restricted to be equal to one another. Such restrictions can be incorporated by introducing the unknown $(K_x \times 1)$ vector τ_i^* with elements that can all differ from each other, and a known $(K_x \times K_x^*)$ selection matrix D^* to correctly map τ_i^* to τ_i via $\tau_i = D^* \tau_i^*$, where $K_x^* \leq K_x$. The selection matrix D^* should be set by the researcher, its elements are either zero or one, and it can have only a single one per row. In case $D^* = I_{K_x}$ we obtain the original formulation. Details of the prior specification and inference can be easily adapted.

2.3.1 Inference

For inference, we develop an efficient Bayesian MCMC sampler. The details of the MCMC sampler are outlined in Appendix 2.A. Specialized code was written in R and C++ to obtain the posterior samples.⁵ In this section, we present the main ideas.

⁵The code for the MCMC sampler was tested using the identity (Geweke, 2004 and Cook et al., 2006)

$$p(\omega) = \int p(\omega|\tilde{y})p(\tilde{y}|\tilde{\omega})p(\tilde{\omega})d\tilde{y}d\tilde{\omega}$$

where ω are the model parameters, $\tilde{\omega}$ is a draw from the prior density $p(\omega)$, \tilde{y} is a draw from the DGP with likelihood function $p(y|\tilde{\omega})$ given $\tilde{\omega}$, and $p(\omega|\tilde{y})$ is the posterior density of ω given \tilde{y} . During testing, we used many replications to approximate the integral on the right-hand side and checked whether the approximated marginal densities of ω matched the prior marginal densities. That is, for each replication, we drew $\tilde{\omega}$ from its prior and used this draw to generate data \tilde{y} from the DGP. Next, we used the MCMC sampler to obtain posterior draws for ω given the generated data \tilde{y} . Finally, for each parameter in ω , we considered the posterior draws over all replications,

To draw the DP mixture parameters, we use algorithm 2 in Neal (2000). That is, we augment the parameter space with the latent membership indicator c_i that indicates which mixture component unit i belongs to. This procedure is similar to that for a finite mixture, except that for the DP mixture, components may appear or disappear in subsequent MCMC iterations. Due to the conjugacy of the base distribution $p(\mu_q, \Sigma_q)$, we can use a computationally efficient Gibbs step to draw c_i . Moreover, in this Gibbs step we draw c_i unconditional on the component membership probabilities π . Hence, there is no need to draw π .

Per MCMC iteration, we draw (i) the DP mixture parameters $\{\lambda_i\}_{i=1}^N$, $\{c_i\}_{i=1}^N$, $\{\mu_q\}_q$ and $\{\Sigma_q\}_q$, (ii) the variable selection parameters $\{\tau_i\}_{i=1}^N$ and θ , and (iii) γ . Conditional on $\{c_i\}_{i=1}^N$, drawing $\{\lambda_i\}_{i=1}^N$, $\{\mu_q\}_q$ and $\{\Sigma_q\}_q$ becomes straightforward: λ_i can be drawn using a random walk Metropolis-Hastings (M-H) step (Metropolis et al., 1953, Hastings, 1970), μ_q can be drawn from a multivariate normal using only the λ_i from the units for which $c_i = q$, and similarly Σ_q can be drawn from an inverse Wishart distribution. Furthermore, we draw γ using a random walk M-H step, τ_{ik} using a Bernoulli distribution, and θ_k from a Beta distribution.

For some models, including the linear model, the M-H steps to draw λ_i and γ can be directly replaced by Gibbs steps. For models in which this is not the case, we do not recommend to perform any further data augmentation to enable a Gibbs step for λ_i and γ . For example, we would not recommend to augment the latent utilities in the multinomial logit model (using e.g. the augmentation schemes in Polson et al., 2013 or Frühwirth-Schnatter and Frühwirth, 2010). Such types of data augmentation can lead to poor mixing in the MCMC sampler. The main reason for poor mixing is that, for the example of the multinomial logit model, the latent utilities are drawn conditional on the variable selection indicators τ_i . In case in a MCMC iteration, one obtains a draw $\tau_{ik} = 0$, the draw for the latent utility will assign no weight to the k^{th} variable. In the next MCMC iteration, this may cause a high probability to again draw $\tau_{ik} = 0$ conditional on the latent utility. That is, the correlation between posterior draws of τ_i and the latent utilities can be quite high.

To improve mixing of the sampler, we jointly draw λ_{ik} and τ_{ik} for each variable k , and we randomize the order over k across the MCMC iterations. Alternatively, one may jointly draw λ_i and τ_i over all variables. In that case, the computation of the likelihood function requires the evaluation of 2^{K_x} terms of likelihood contributions of unit i due to all possible combinations of variables selected. These evaluations

and checked whether the posterior marginal densities coincided with the prior marginal densities.

can generally not be simplified. Hence, this should only be done when K_x is small, say smaller than five. By drawing separately per variable, the likelihood function contains only 2 terms to compute (one for $\tau_{ik} = 1$ and one for $\tau_{ik} = \kappa$) and this has to be repeated K_x times.

Our model and Bayesian MCMC sampler can be used for any nonlinear model of the form in Equation (2.1). The sampler does rely on the computation of the likelihood function conditional on λ_i and γ , for performing the M-H steps for λ_{ik} and γ and for drawing τ_{ik} . For many models, this likelihood function can be analytically computed, e.g. for the multinomial logit model, poisson model, and negative binomial model. For other models, the likelihood function has to be approximated, e.g. for the multinomial probit model (MNP) when the number of possible outcomes for Y_{it} exceeds two. For these later cases, our MCMC sampler can become slow due to the computations necessary for approximating the likelihood function, and more efficient approaches could entail further data augmentation, for example the latent utilities for the MNP. Again, care must be taken, because conditioning on the augmented parameters can lead to high correlation in the chains due to the conditioning on the variable selection indicators τ_i .

2.4 Monte Carlo study

In this section, we perform a small Monte Carlo study to examine the performance of our proposed approach for accommodating heterogeneous variable selection. For this purpose, we consider a multinomial logit model (McFadden, 1973, Manski, 1977). At each observation t , a unit i selects one of J alternatives. Each alternative j is described by K_x variables in the vector x_{itj} . The multinomial logit model is given by

$$Y_{it} \sim \text{Multinomial}(1, p_{it}), \quad (2.11)$$

$$p_{itj} \equiv \Pr[Y_{it} = j | \beta_i] = \frac{\exp(x'_{itj} \beta_i)}{\sum_{l=1}^J \exp(x'_{itl} \beta_i)}, \quad j = 1, \dots, J, \quad (2.12)$$

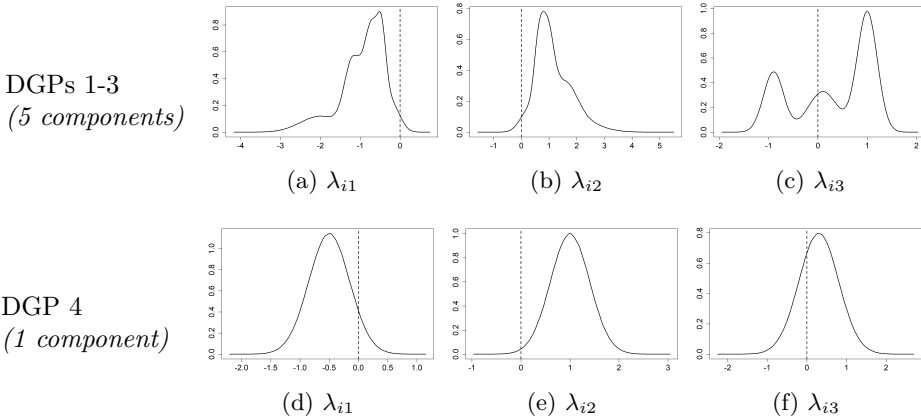
where $p_{it} = (p_{it1}, \dots, p_{itJ})'$.

We consider four data generating processes (DGPs) and perform 100 Monte Carlo replications per DGP. In each DGP, we consider 1,000 units, 20 observations per unit, 3 alternatives per observation, and 3 variables: x_{1itj} from a standard normal distribution and x_{2itj}, x_{3itj} from a Bernoulli distribution with probability of outcome

1 equal to 0.5. For all DGPs, we let $\beta_{ik} = \tau_{ik}\lambda_{ik}$, where $\tau_{ik} \in \{0, 1\}$ is the variable selection indicator, for $k = 1, 2, 3$.

For DGPs 1 to 3, we let λ_i come from a mixture of multivariate normals with five components. The components' means, covariance matrices and weights are equal across the three DGPs, whereas the amount of variable selection differs across the DGPs. In the mixture, the marginal density of λ_{i1} mostly has mass on the negative domain, is skewed and has an extra mode in the tail, that of λ_{i2} is skewed with mass mostly on the positive domain, and that of λ_{i3} is multimodal with a mode at zero and substantial mass on both the positive and negative domain, see Figures 2.1 (a)-(c).⁶ Hence, the first variable could represent price, the second variable a quality indicator, and the third variable a brand indicator. For the heterogeneous variable selection part, we take the following probabilities that a variable is relevant for a unit, i.e., that the unit assigns weight to the variable. In DGP 1, the variables are relevant for the majority of units: $\theta = (0.90, 0.85, 0.95)$. In other words, 90% of units assign weight to the first variable, 85% to the second variable, and 95% to the third variable. In DGP 2, the variables are relevant for all units: $\theta = (1, 1, 1)$. In DGP 3, there are quite some units for which the variables are irrelevant: $\theta = (0.80, 0.70, 0.75)$.

Figure 2.1: True marginal densities of λ_{i1} , λ_{i2} and λ_{i3} for DGPs 1 to 3 (top) and DGP 4 (bottom).



⁶For DGPs 1-3 with five mixture components we use the following setting. We set the membership probabilities to $\pi = (0.25, 0.1, 0.15, 0.1, 0.4)$, the components' means to $\mu_1 = (-1.2, -0.45, -2, -0.2, -0.7)$, $\mu_2 = (1.6, 0.6, 2, 0.25, 0.9)$ and $\mu_3 = (0.1, 1, -0.9, -0.9, 1)$, and the components' covariance matrices with standard deviations, $\sigma_1 = (0.2, 0.1, 0.5, 0.2, 0.2)$, $\sigma_2 = (0.4, 0.15, 0.75, 0.3, 0.25)$, and $\sigma_3 = (0.3, 0.2, 0.2, 0.2, 0.2)$, and correlations (equal across components) $\rho_{12} = 0.2$, $\rho_{13} = 0.1$ and $\rho_{23} = 0.2$.

For DGP 4, we use one mixture component for λ_i , see Figures 2.1 (d)-(f).⁷ We use the same amount of variable selection as in DGP 1, that is, $\theta = (0.90, 0.85, 0.95)$.

We estimate a MNL using three different approaches for the heterogeneous responses: (1) our proposed DP mixture with heterogeneous variable selection (HVS-DPM), (2) a “standard” DP mixture without heterogeneous variable selection (DPM), and (3) a *single* multivariate normal distribution with heterogeneous variable selection (HVS-M). We set the priors’ hyperparameters to $\alpha = 1$, $\mu_0 = 0$, $d = 0.5$, $\nu = K_x + 5$, $v = 0.2$, and $a = b = 1$. Hence, the prior distribution for θ_k is uniform over the unit interval. Appendix 2.B gives the histograms of the prior number of components based on α and N , the marginal prior on μ and the marginal prior on the standard deviations on the diagonal of Σ . Furthermore, we set $\kappa = 0$ in estimation.

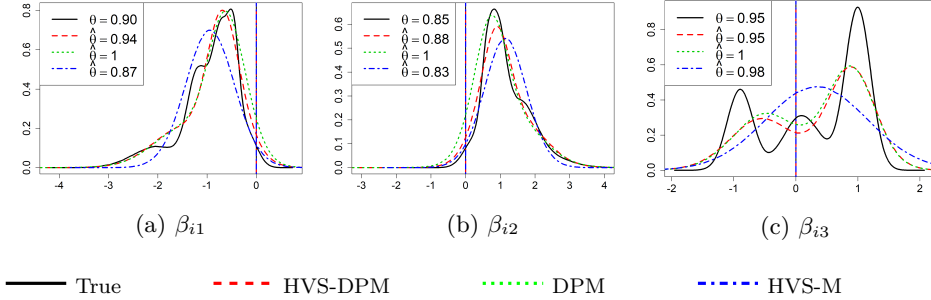
For the posterior results per replication, we use 15,000 simulations after 5,000 burn-in draws and keep every 4th draw. We visualize the results per DGP using the posterior marginal densities of β_{i1} , β_{i2} , and β_{i3} . For this purpose, we first construct the posterior marginal densities for each of the 100 replications. That is, for each replication, we take the equally weighted mixture of the 15,000/4 posterior draws of marginal densities, where each draw of the density directly results from the draws of the parameters of the mixture of multivariate normals (π, μ, Σ) and of the heterogeneous variable selection (θ) . For each DGP, we plot the equally weighted mixture of these 100 marginal densities.

2.4.1 Results

The posterior results for DGP 1, with substantial variable relevance and non-normal continuous heterogeneity, are shown in Figure 2.2.⁸ In this figure, we plot the marginal posterior densities of β_{i1} , β_{i2} , and β_{i3} , by plotting the underlying continuous heterogeneity distribution (the mixture of multivariate normals) as a continuous density. Moreover, we represent the heterogeneous variable selection, i.e. the relative number of units that assign no weight to the variable, by a vertical line through zero. The probability mass at zero is equal to one minus the mean across replications of the posterior mean of θ , displayed in the top left corner.

⁷For DGP 4 with one mixture component we set the mean to $\mu = (-0.5, 1.0, 0.3)$ and the covariance matrix to Σ with standard deviations $\sigma = (0.35, 0.40, 0.50)$ for the three variables, respectively, and correlations $\rho_{12} = 0.2$, $\rho_{13} = 0.1$ and $\rho_{23} = 0.4$.

⁸To obtain 1.000 draws from the posterior for a Monte Carlo replication generated from DGP 1, it takes about 50 seconds for the HVS-DPM, 10 seconds for the DPM and 45 seconds for the HVS-M. Simulations were done using 1 core on an Intel Core i7 processor with 2.6GHz frequency.

Figure 2.2: (Posterior) marginal densities of β_{i1} , β_{i2} and β_{i3} for DGP 1.

Our proposed model is well able to capture the skewness and multimodality in the continuous heterogeneity. The fit is not perfect, mainly because we find components that are less peaked than they are in reality, that is, we find components with larger variances. Due to this smoothing, primarily caused by the prior on the covariance matrices, the mass close to zero of the continuous heterogeneity distribution is slightly overestimated and therefore the probability that a variable is selected is overestimated. In sum, for the skewed distribution for variables one and two, our model is able to capture the modes and the heavy tails. For variable three, the mode at zero of the continuous heterogeneity distribution is missed, and the modes at the positive and negative side are less extreme than in reality.

Compared to the alternative approaches, our approach seems to best capture the underlying distribution of heterogeneous responses. The standard DP mixture without variable selection cannot capture the spike at zero. Instead, more mass is allocated between -0.5 and 0.5 . The single multivariate normal approach with variable selection cannot capture the non-normality in the continuous heterogeneity, and compensates by shifting the mode away from zero for the skewed distributions, and finding much less heavy tails.

To further compare the performance of the three approaches for modeling heterogeneous responses, we consider the predictive performance. We generate five more observations for each unit. For each Monte Carlo replication and each approach, we compute the predictive log-likelihood contribution per unit based on these five out-of-sample observations.⁹ For easy comparison, we take the sum of predictive log-likelihood contributions across units and subtract the value obtained with one of the alternative approaches (DPM or HVS-M) from the value obtained with our

⁹The predictive log-likelihood contribution of unit i can be approximated using the posterior

approach (HVS-DPM). A positive number indicates our approach leads to a better predictive performance, a negative number indicates the alternative approach leads to a better predictive performance.

Table 2.2: Difference between the sum of predictive log-likelihood contributions for our approach (HVS-DPM) against two alternative approaches (DPM and HVS-M) per DGP. Based on 100 replications. Averages and percentages of replications for which the difference is greater than zero.

DGP	HVS-DPM against DPM		HVS-DPM against HVS-M	
	Mean	% > 0	Mean	% > 0
DGP 1	2.3	84%	24.9	100%
DGP 2	-0.5	41%	37.0	100%
DGP 3	5.5	98%	11.4	98%
DGP 4	1.0	75%	-0.2	36%

The results for the predictive log-likelihood contributions are in Table 2.2. We report the means over the Monte Carlo replications and the fraction of Monte Carlo replications for which our approach has a better predictive performance. For DGP 1, we find that the predictions obtained with our approach are substantially better than those obtained with the alternative approaches. This holds in particular in comparison with the single multivariate normal approach (HVS-M): none of the replications of the HVS-M approach has a higher log-likelihood value.

For further evaluation, we consider the hit rates: how well are the MNLs based on the three approaches able to accurately assign, at the unit-level, posterior mass to β_{ik} . The results are in Table 2.9. In this table, we show the percentage of units for which the posterior draw of β_{ik} lies in the interval $[-\epsilon, \epsilon]$ for different values of ϵ , averaged over draws, variables and replications. We do this for four groups: (1) all units, (2) units for which the true β_{ik} lies within the interval, (3) units for which the true β_{ik} does not lie within the interval, and (4) units for which the true $\beta_{ik} = 0$. For DGP 1, we find that our approach slightly underestimates the mass between

samples:

$$\log p(y_i^* | y) \approx \log \left[\frac{1}{S} \sum_{s=1}^S \left(\prod_{t \in \mathcal{T}_i^*} \Pr[Y_{it}^* = y_{it}^* | \beta_i^{(s)}] \right) \right],$$

where y_i^* denotes the out-of-sample observations for individual i that predictions are made for, y denotes the in-sample observations which were used to obtain the posterior draws, S is the number of draws of the MCMC sampler after burn-in, \mathcal{T}_i^* is the set of observations for unit i that was left out of the training sample (i.e. observations 21 until 25), and $\beta_i^{(s)}$ is the s^{th} posterior draw of β_i which can be computed directly using the s^{th} posterior draws for δ_i and τ_i .

$[-0.3, 0.3]$, but not as much as the standard DP mixture approach. In contrast, the single multivariate normal approach leads to an overestimation of the mass close to zero, and underestimation of the mass in the tails. Because of this, the HVS-M approach is better able to assign posterior mass to units that assign weights close to zero but does worse for units with weights further away from zero.

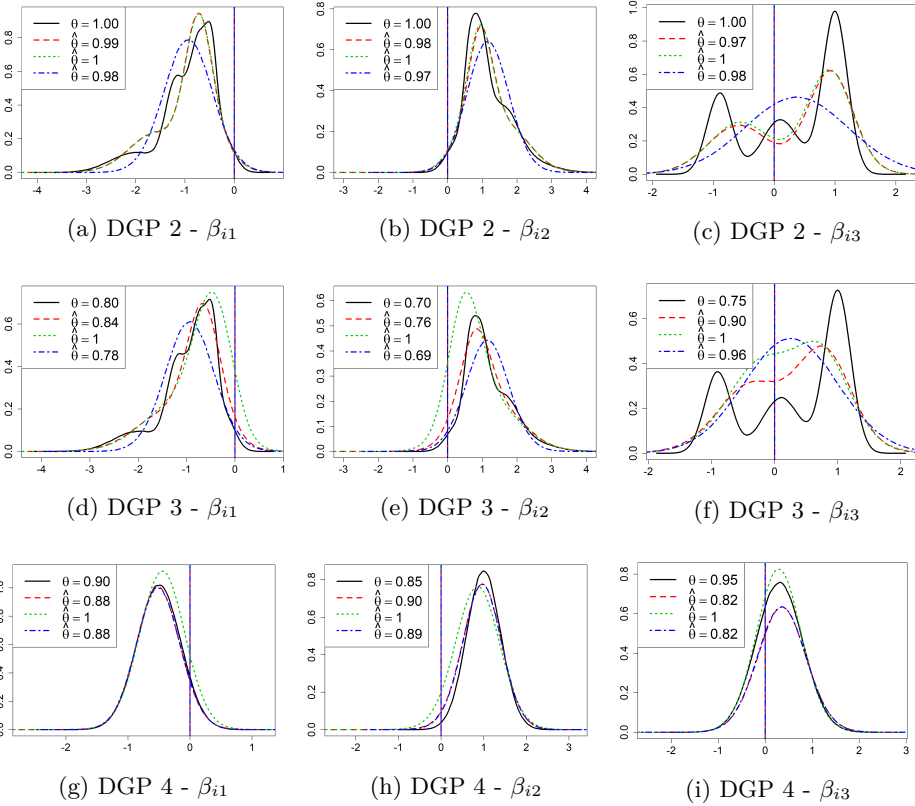
For DGP 2 where the variables are relevant for all units, we find that, as expected, the results of our approach closely match those of the standard DP mixture approach, see Figures 2.3 (a)-(c). With our approach, we do find evidence of a small amount of units which do not assign weight to certain variables (1%-3%). The predictive log-likelihoods indicate that our approach leads to a similar predictive performance as the standard DP mixture, with the standard DPM being slightly better, see Table 2.2.

The results for DGP 3, where for quite some units the variables are irrelevant, are in Figures 2.3 (d)-(f). Again, we find that our approach seems to be most accurate in capturing the true density. Furthermore, the improvement in predictive performance of our approach as compared to the standard DP mixture approach is greater than for DGP 1, see Table 2.2. Hence, the more units that assign no weight to certain variables, the more important it becomes to account for heterogeneous variable selection.

When the continuous heterogeneity follows a normal distribution as in DGP 4, our approach and the HVS-M approach with a single multivariate normal for the continuous heterogeneity find a similar shape for the underlying distribution of heterogeneous responses, see Figures 2.3 (g)-(i). For the third variable, the amount of variable selection is underestimated by both approaches: an estimated 82% of units assign weight to the third variable, whereas in reality it is 95%. This affects the shape of continuous heterogeneity found, which underestimates the mass between -0.5 and 0.5. The predictive log-likelihoods in Table 2.2 indicate that our approach leads to a similar predictive performance as the HVS-M approach.

As a final note. In this Monte Carlo study, we use $K_x = 3$ variables. Already with this small number of variables, we see that our approach with heterogeneous variable selection performs better than the standard DP mixture approach. In case there are more variables, we expect this difference in performance to be even greater, as the standard DP mixture would need at least 2^{K_x} components to capture all combinations of variable selection.

Figure 2.3: (Posterior) marginal densities of β_{i1} , β_{i2} and β_{i3} for DGPs 2-4.



— True
- - - HVS-DPM
... DPM
- . - . HVS-M

Table 2.3: Percentage of units for which the posterior draw of β_{ik} falls within $-\epsilon \leq \beta_{ik} \leq \epsilon$ for multiple values of ϵ (averaged over Monte Carlo replications, draws and variables) for DGP 1. The results for DGPs 2 to 4 are in Appendix 2.C.

ϵ	(1) All			(2) True $-\epsilon \leq \beta_{ik} \leq \epsilon$			(3) True $\beta_{ik} < -\epsilon$ or $\beta_{ik} > \epsilon$			(4) True $\beta_{ik} = 0$			
	True	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M
0.00	10	7	0	10	24	0	36	6	0	8	24	0	36
0.10	13	11	5	15	29	13	38	8	4	11	31	12	42
0.20	17	14	10	19	36	25	44	10	7	14	39	25	48
0.30	20	19	16	23	43	36	49	13	11	17	46	36	54
0.40	23	24	22	28	49	45	55	16	15	20	54	47	60
0.50	28	29	29	33	55	53	59	19	19	24	61	57	66
0.75	43	46	47	48	67	68	68	30	32	32	77	76	78
1.00	64	64	65	62	78	79	76	40	41	38	88	88	87
1.50	88	86	86	85	92	92	90	42	44	52	97	97	97
2.00	95	94	94	96	96	96	97	59	60	81	99	99	99
2.50	98	98	98	99	98	98	100	78	77	95	100	100	100

Results for four different groups:

- (1) all units,
- (2) units for which true β_{ik} falls within interval,
- (3) units for which true β_{ik} does not fall within interval,
- (4) units for which true $\beta_{ik} = 0$ ($\tau_{ik} = 0$).

2.5 Case study: multinomial logit model

In this section, we illustrate our approach with an empirical application. We again consider the multinomial logit model in Equations (2.11) and (2.12). We consider responses obtained from a discrete choice experiment on food choices (Koç & van Kippersluis, 2017).¹⁰ During the choice experiment, respondents had to complete 18 choice tasks. In each task, a respondent was asked which out of two meals s/he would eat most regularly. The meals were described by attributes as price and taste, and by attributes describing how healthy the meal is.

The respondents were divided into three groups. Each respondent group obtained different types of choice tasks in terms of the attributes describing how healthy the meal is and the amount of health information provided in the text. For group 1 (1,206 respondents), the meals were described by four attributes: price, cooking time, taste, and health consequences. All health information was provided in the final attribute health consequences. For groups 2 (1,154 respondents) and 3 (1,185 respondents), the meals were described by six attributes: price, cooking time, taste, number of calories, grams of saturated fat, and grams of sodium. Group 2 obtained health information in the text regarding what amount of calories, saturated fat, and sodium constitutes a healthy meal, whereas group 3 did not obtain health information. The ordering of the tasks within each respondent group was randomized over the respondents.

Table 2.4: Attributes and attribute levels in the choice tasks for the discrete choice experiment on healthy food choices. The final column indicates which respondents groups (1,2 or 3) saw which attributes in the choice experiment.

Attribute	Attribute levels			Respondent groups
Price	2 Euro	6 Euro	10 Euro	1, 2, 3
Cooking time	10 min	30 min	50 min	1, 2, 3
Taste	OK	Good	Very good	1, 2, 3
Health consequences	Unhealthy	Health neutral	Healthy	1
Number of kilocalories	800	1,100	1,400	2, 3
Grams of saturated fat	10	20	30	2, 3
Milligrams of sodium	900	1,200	1,500	2, 3

Each of the attributes took on one of three values. The attribute levels had a clear ordering, see Table 3.2. For example, the price of the meal could either be 2 Euros, 6 Euros, or 10 Euros. In the model, we include a separate dummy variable per attribute level, with the exception of a baseline level per attribute (the middle level). Furthermore, we restrict the variable selection to hold for all levels of the same

¹⁰We thank the LISS panel and the experiment designers for providing this dataset.

attribute. That is, we consider whether an individual finds an attribute relevant (such as price), and not just one of the attribute levels (such as price 2 Euros). Heterogeneous variable selection in such an application is also known as attribute non-attendance (Scarpa et al., 2009).

As in the Monte Carlo study, we use three approaches for modeling heterogeneous responses in the MNL: (1) our proposed DP mixture with heterogeneous variable selection (HVS-DPM), (2) a “standard” DP mixture without heterogeneous variable selection (DPM), and (3) a single multivariate normal distribution with heterogeneous variable selection (HVS-M). For posterior results, we use 60,000 simulations after 40,000 burn-in draws and we keep every 10^{th} draw. We use the same priors as in the Monte Carlo study and set $\kappa = 0$.

The MCMC sampler converges rather quickly and mixes well in general. For extreme quantiles of the heterogeneity distribution, the mixing is less good. This is not surprising as only very few observations are informative for such quantiles. Trace plots are given in the Supplementary Materials, available upon request.

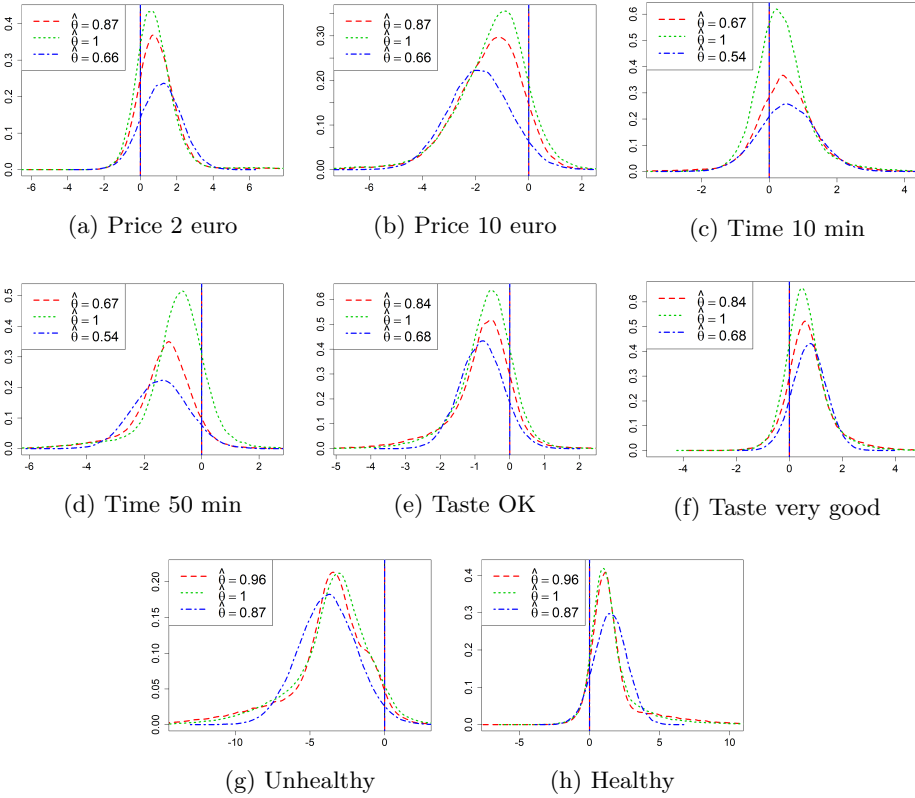
2.5.1 Results

The posterior marginal densities of β_i for the first respondent group are displayed in Figure 2.4.¹¹ For this group, the meals were described by four attributes. Using our approach with a DP mixture and heterogeneous variable selection, we find evidence of the existence of groups of respondents that ‘ignore’ attributes, for all four attributes. Ignorance of attributes, or attribute non-attendance, can mean that either a respondent did not consider the attribute or is indifferent between the attribute levels. The health attribute is least ignored (4%), followed by price (13%), taste (16%), and cooking time (33%). The marginal distributions seem skewed, most mass is usually at either the positive or the negative side, and there is a heavy tail away from zero. For the health attribute levels, the tail is especially thick, indicating that there are groups of respondents that highly value this attribute.

The HVS-M approach with a single multivariate normal clearly cannot capture the skewness in the marginal distributions. Instead, to somewhat capture the heavy tail and that most mass is on one side of the distribution, the mode of the distribution is shifted further away from zero, leading to selection probabilities that are substantially lower than we find with our approach. Finally, the standard DP mixture without

¹¹The posterior marginal densities are constructed from the posterior draws in the same way as in the Monte Carlo study.

Figure 2.4: Posterior marginal density of β_{ik} for respondent group 1.



Baseline levels are price 6 euro, cooking time 30 minutes, taste good, and health neutral.

- HVS-DPM
- ... DPM
- .- HVS-M

variable selection finds roughly the same forms of the density as our approach with variable selection, but as it cannot capture the peak at zero, it distributes more mass between -1.0 and 1.0. This can be seen most clearly in Table 2.6, which shows the percentage of draws for β_{ik} in the interval $[-\epsilon, \epsilon]$. In this table, we also see that the DP mixture approaches assign more mass in the tails than the HVS-M approach.

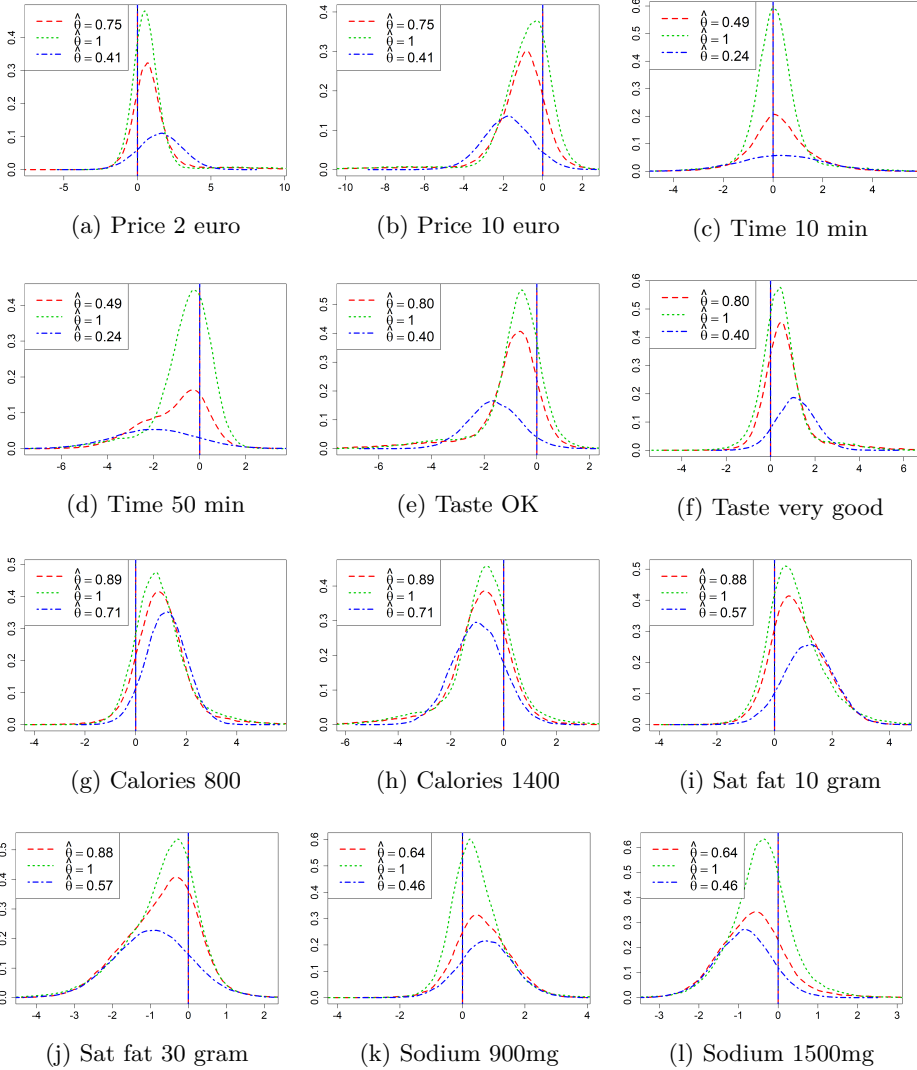
The results for the second (health info) and third (no health info) respondent groups are in Figures 2.5 and 2.6, respectively. For these groups, the meals were described by six attributes. For both respondent groups we again find evidence of variable ignorance and non-normality of the heterogeneity when using our approach.

To better show the difference in variable selection across the respondent groups, Table 2.5 concisely displays the posterior means and 95% highest posterior density intervals (HPDIs) of θ_k — the probability that attribute k is selected — per respondent group and attribute for the DP mixture approach with heterogeneous variable selection. For the meals described by six attributes (groups 2 and 3), including the health information seems to have the respondents made more aware of calories and saturated fat, but the opposite seems to hold for sodium. Furthermore, compared to the first group, the individuals in the second and third group seem to more often ignore the standard attributes price, cooking time, and taste. The 95% HPDIs are quite wide, indicating that there is quite some uncertainty in these values.

Table 2.5: Posterior means and 95% HPDIs of attribute selection probabilities θ per respondent group and attribute (results of HVS-DPM).

Attribute	Group 1		Group 2		Group 3	
	Mean	95% HPDI	Mean	95% HPDI	Mean	95% HPDI
Price	0.87	(0.79,0.96)	0.75	(0.65,0.84)	0.63	(0.52,0.76)
Cooking time	0.67	(0.57,0.77)	0.49	(0.39,0.58)	0.39	(0.31,0.48)
Taste	0.84	(0.72,0.97)	0.80	(0.67,0.97)	0.75	(0.65,0.85)
Health	0.96	(0.93,0.99)	-	-	-	-
Number of kilocalories	-	-	0.89	(0.80,0.98)	0.81	(0.72,0.89)
Grams of saturated fat	-	-	0.88	(0.78,1.00)	0.86	(0.73,0.97)
Milligrams of sodium	-	-	0.64	(0.51,0.79)	0.74	(0.56,0.91)

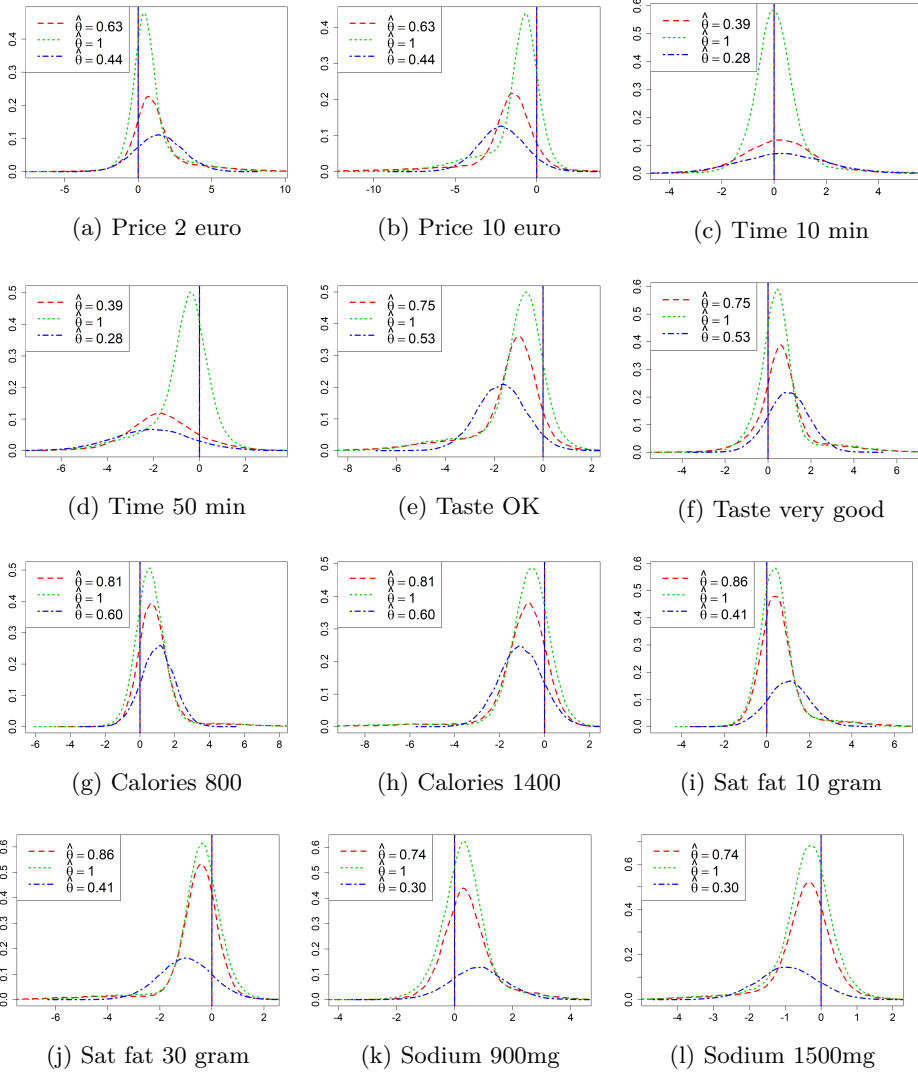
Figure 2.5: Posterior marginal density of β_{ik} for respondent group 2.



Baseline levels are price 6 euro, cooking time 30 minutes, taste good, calories 1100, saturated fat 20 gram, and sodium 1200 mg.

- HVS-DPM
- ... DPM
- .- HVS-M

Figure 2.6: Posterior marginal density of β_{ik} for respondent group 3.



Baseline levels are price 6 euro, cooking time 30 minutes, taste good, calories 1100, saturated fat 20 gram, and sodium 1200 mg.

- HVS-DPM
- ... DPM
- .- HVS-M

Table 2.6: Percentage of individuals for which the posterior draw of β_{ik} falls within $-\epsilon \leq \beta_{ik} \leq \epsilon$ (averaged over draws and variables) per respondent group.

ϵ	Group 1			Group 2			Group 3		
	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M
0.00	17	0	31	26	0	54	33	0	58
0.10	20	6	34	31	9	55	37	9	59
0.20	24	12	36	36	17	57	42	18	61
0.30	28	18	39	41	25	59	46	27	63
0.40	32	24	42	46	33	61	50	35	64
0.50	36	30	45	50	41	63	55	43	66
0.75	46	43	52	60	57	68	64	59	71
1.00	55	54	58	70	69	73	72	71	75
1.50	70	70	70	83	84	83	84	85	83
2.00	79	79	79	90	91	90	90	91	89
2.50	84	85	85	94	95	95	93	93	94
3.00	88	89	90	96	96	97	95	95	97
4.00	93	93	94	98	98	99	97	97	99
5.00	96	96	97	99	99	100	99	99	100

2.5.2 Out-of-sample performance

We next evaluate the forecasting performance of our approach for modeling heterogeneous responses for the empirical dataset on food choice. We compare the performance across the three approaches (HVS-DPM, DPM, and HVS-M) using two measures: (i) hit rates and (ii) predictive log-likelihood contributions.

To construct forecasts, we split the sample into two parts: (i) a forecasting sample consisting of two randomly chosen observations per individual, and (ii) a training sample with the remaining observations. We obtain posterior samples using again 60,000 simulations after 40,000 burn-in draws and we keep every 10^{th} draw. We repeat this exercise ten times per respondent group, using different randomly chosen forecasting samples to increase the robustness of the results to the selection of the forecasting sample.

First, we consider the hit rates. The hit rate is equal to the percentage of choices that are correctly predicted, that is, the percentage of choices for which the predicted choice probability is higher than 0.5.¹² The results are in Table 2.7, where we show the average hit rate over the ten forecasting samples for each approach and respondent group.

Table 2.7: Hit rates for each approach to modeling heterogeneous responses (HVS-DPM, DPM, HVS-M) and respondent group. Averaged across ten different forecasting samples.

	HVS-DPM	DPM	HVS-M
Group 1	79.9%	79.8%	79.4%
Group 2	77.9%	78.1%	77.5%
Group 3	78.0%	77.8%	77.7%

The approach with a single multivariate normal and heterogeneous variable selection (HVS-M) gives the lowest hit rate for all three respondent groups. Hence, this approach seems to give the weakest forecasting performance. For respondent groups

¹²The choice probabilities are approximated using the posterior samples:

$$\Pr[Y_{it}^* = y_{it}^* | y] \approx \frac{1}{S} \sum_{s=1}^S \Pr[Y_{it}^* = y_{it}^* | \beta_i^{(s)}],$$

where y_{it}^* denotes the out-of-sample observation for individual i that a prediction is made for, y denotes the in-sample observations which were used to obtain the posterior draws, S is the number of draws of the MCMC sampler after burn-in and $\beta_i^{(s)}$ is the s^{th} posterior draw of β_i .

1 and 3, our HVS-DPM approach provides the best forecasts, and for group 2 the DPM approach provides the best forecasts. The differences are, however, small.

Finally, we consider the predictive log-likelihood contribution per individual. We subtract the sum of predictive log-likelihood contributions obtained with the two alternative approaches from that obtained with our approach. These measures are computed in the same manner as in the Monte Carlo study. The results for the predictive log-likelihood contributions are given in Table 2.8. For each respondent group, we report the averages across the ten replications, and the percentages of replications for which the difference is positive.

Table 2.8: Difference between the sum of predictive log-likelihood contributions for our approach (HVS-DPM) against two alternative approaches (DPM and HVS-M) per DGP. Based on 10 forecasting samples. Averages and percentages of replications for which the difference is greater than zero.

DGP	HVS-DPM against DPM		HVS-DPM against HVS-M	
	Mean	% > 0	Mean	% > 0
Group 1	1.13	70%	12.26	90%
Group 2	0.85	60%	8.16	70%
Group 3	6.39	80%	10.80	100%

The averages are all positive, indicating that our approach leads to a better forecasting performance than the alternative approaches. Our approach clearly stands out as compared to the HVS-M approach with a single multivariate normal that was proposed by Gilbride et al. (2006).¹³ The differences in predictive log-likelihood contributions are much larger than zero for the majority of forecasting samples for all three respondent groups. Hence, for this dataset on food choices, there is sufficient evidence of non-normality in the distribution of preferences. These findings indicate that allowing for flexible cross-sectional heterogeneity via a mixture of multivariate normals is important for understanding and predicting choice behavior.

Our approach also compares favorably to the DPM approach without heterogeneous variable selection, although the results are less overwhelming for respondent groups 1 and 2. A possible reason for the relative small difference in predictive performance could be that quite some individuals have strong preferences (> 2) for one attribute or another, as indicated by the heavy tails. In the multinomial logit model, such attributes will dominate the choice predictions, making it less important to accurately

¹³For the HVS-M approach, we use the model specification provided in Gilbride et al. (2006) and our computationally efficient Bayesian MCMC sampler.

estimate the preferences close to zero. The final respondent group, for which the predictive performance of our approach clearly stands out, was the group which had to make decisions on the largest number of attributes (six) without obtaining objective information on the health attributes. For this group, allowing for heterogeneous variable selection substantially improves the predictive performance.

2.6 Conclusion

In this paper, we develop a general method for heterogeneous variable selection in Bayesian nonlinear panel data models. We allow for flexible cross-sectional heterogeneity by letting the model's unit-specific parameters follow a Dirichlet process mixture of multivariate normals. Our main contribution is that we augment the DP mixture with heterogeneous variable selection. This allows modeling the possibility that subsets of units are unaffected by certain variables, as may be present in applications as diverse as health treatments, choice situations, macroeconomics, and operations research. We develop our approach for nonlinear panel data models including multinomial logit and probit models, count models, exponential models, among many others. Finally, we develop an efficient Bayesian MCMC sampler to allow for inference for datasets with up to 50 or 100 explanatory variables.

We illustrate the model with a Monte Carlo study and an empirical application. For illustration, we consider a multinomial logit model as this model is the focus of most literature on heterogeneous variable selection. In the Monte Carlo study we find that our approach is able to capture both complex forms of continuous cross-sectional heterogeneity — such as skewness and multimodality — as well as heterogeneous variable selection. A 'standard' DP mixture cannot capture heterogeneous variable selection. Instead of a spike at zero, this approach generally allocates probability mass to a relatively large region around zero, depending on the shape of the continuous heterogeneity. In the empirical application, we consider responses to a discrete choice experiment on food choices. We find substantial evidence of attribute non-attendance and non-normality of the continuous heterogeneity. In particular, the continuous heterogeneity seems skewed. These findings indicate the usefulness of our approach in practice.

A limitation of the proposed approach is the use of a conjugate prior for the components' means and covariance matrices. Although this prior is advantageous for estimation, it may be unrealistic as the prior on the component's mean directly de-

depends on the component's covariance matrix. This implies that the marginal prior on the mean is tighter when the corresponding variance is small. If the conjugacy of the prior would be relaxed, it is required to draw the component membership indicators with a Metropolis-Hastings step instead of a Gibbs step. This could dramatically increase computation time due to worse mixing properties of the resulting MCMC sampler.

We note three interesting venues for future research. First, one can allow for correlated variable selection. This could be incorporated, for example, by allowing for a different membership probability per combination of variables selected and putting a (Dirichlet) prior on these membership probabilities. Second, the model can be generalized to allow for time-varying parameters, including time-varying variable selection. In choice situations, this could model changing preferences of individuals, or learning and fatigue effects. Finally, the nonlinear (univariate) panel data model can be extended to multivariate outcomes. This would require inference on the correlations across outcomes.

Appendix

2.A MCMC sampler

In this section, we develop the MCMC sampler for our nonlinear panel data model with heterogeneous variable selection in Equations (2.1)-(2.10). In summary, the model is given by

$$Y_{it} | \beta_i, \gamma \sim f(g(x_{it}, \beta_i, z_{it}, \gamma)),$$

$$\beta_{ik} | \tau_{ik}, \lambda_{ik} = \tau_{ik} \lambda_{ik},$$

with variable selection priors

$$\tau_{ik} \in \{\kappa, 1\},$$

$$\Pr[\tau_{ik} = 1 | \theta_k] = \theta_k,$$

$$\theta_k \sim \text{Beta}(a, b),$$

DP mixture priors

$$\begin{aligned} \lambda_i | \{\pi_q\}_q, \{\mu_q\}_q, \{\Sigma_q\}_q &\sim \sum_{q=1}^{\infty} \pi_q MVN(\mu_q, \Sigma_q) \\ \pi_q &= \eta_q \prod_{r=1}^{q-1} (1 - \eta_r), \quad \eta_q \sim \text{Beta}(1, \alpha), \\ \mu_q | \Sigma_q &\sim MVN(\mu_0, d^{-1} \Sigma_q), \\ \Sigma_q &\sim IW(\nu, \nu v I), \end{aligned}$$

and finally

$$\gamma \sim MVN(\gamma_0, \Sigma_\gamma).$$

The hyperparameters α , μ_0 , d , ν , v , γ_0 , Σ_γ , a and b , are assumed fixed. The sampler can be easily extended to allow for priors on these hyperparameters.

The MCMC sampler is given by

- (1) $c_i | c_{-i}, \lambda_i, \{\mu_q\}_q, \{\Sigma_q\}_q$ for $i = 1, \dots, N$, (Gibbs, multinomial),
- (2) $\mu_q, \Sigma_q | \{\lambda_i\}_{i=1}^N, \{c_i\}_{i=1}^N$ for every unique q in $\{c_1, \dots, c_N\}$:
 - (2a) $\Sigma_q | \{\lambda_i\}_{i=1}^N, \{c_i\}_{i=1}^N$ (Gibbs, inverse Wishart),
 - (2b) $\mu_q | \{\lambda_i\}_{i=1}^N, \{c_i\}_{i=1}^N, \Sigma_q$ (Gibbs, multivariate normal),
- (3) $\lambda_{ik}, \tau_{ik} | y_i, \lambda_{i,-k}, \tau_{i,-k}, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta_k, \gamma$ for $i = 1, \dots, N$, and $k = 1, \dots, K_x$:
 - (3a) $\lambda_{ik} | y_i, \lambda_{i,-k}, \tau_{i,-k}, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta_k, \gamma$ (M-H, random walk),
 - (3b) $\tau_{ik} | y_i, \lambda_i, \tau_{i,-k}, \theta_k, \gamma$ (Gibbs, Bernoulli),
- (4) $\theta_k | \{\tau_{ik}\}_{i=1}^N$ for $k = 1, \dots, K_x$, (Gibbs, Beta),
- (5) $\gamma | \{y_i\}_{i=1}^N, \{c_i\}_{i=1}^N, \{\lambda_i\}_{i=1}^N, \{\tau_i\}_{i=1}^N$ (M-H, random walk),

In this sampler, we jointly draw μ_q and Σ_q , and we jointly draw λ_{ik} and τ_{ik} .

Two remarks on this sampler. First, in case some of the variables should be simultaneously selected, and thus $K_x^* < K_x$ (see Section 2.3, right before Section 2.3.1), step 3 should be slightly altered to loop over all $k = 1, \dots, K_x^*$ and, per k , to jointly draw $\{\lambda_{il}, \tau_{il}\}$ over all l for which $D_{l,k}^* = 1$. Second, in case K_x is really small, say $K_x < 5$, the MCMC sampler could be more efficient when λ_i and τ_i are jointly drawn over all variables instead of per variable k . In this case, step 3 can be replaced by

step 3* below

$$\begin{aligned}
 (3^*) \quad & \lambda_i, \tau_i \mid y_i, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta, \gamma && \text{for } i = 1, \dots, N: \\
 (3a^*) \quad & \lambda_i \mid y_i, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta, \gamma && \text{(M-H, random walk),} \\
 (3b^*) \quad & \tau_i \mid y_i, \lambda_i, \theta, \gamma && \text{(Gibbs, Multinomial).}
 \end{aligned}$$

The starting values are generated as follows. First, we set the number of components to 10, and generate the component membership indicators c_i from a multinomial distribution with 10 outcomes, each with equal probability. Second, the components means μ_q are set to a vector of zeroes, and the component covariance matrices Σ_q to an identity matrix. Third, to draw λ_i and γ , we first compute the maximum likelihood estimates of the parameters of the corresponding model with homogeneous responses and no variable selection. Then, for each individual, we take the λ_i that optimizes a weighted log-likelihood function.¹⁴ Fourth, we set $\theta_k = 0.95$ for $k = 1, \dots, K_x$. Finally, we draw τ_{ik} by first drawing a r_{ik} from a Bernoulli distribution with parameter θ_k and then setting τ_{ik} to 1 when $r_{ik} = 1$ and to κ otherwise.

2.A.1 Draw c_i

We use algorithm 2 from Neal (2000) to sample c_i from the full conditional posterior. Let n_c denote the number of units in component c , and $n_{c,-i}$ denote the number of units in component c if we would not count unit i . Let \mathcal{Q}_i be the current set of distinct components if we would not count unit i . That is, \mathcal{Q}_i consists of the distinct components in $\{c_1, \dots, c_N\} \setminus \{c_i\}$. Let Q_i be the size of the set \mathcal{Q}_i . We draw c_i from a multinomial distribution with $Q_i + 1$ outcomes. The first Q_i possible outcomes are the objects in \mathcal{Q}_i , the final component is a new component. The corresponding

¹⁴The weighted log-likelihood function that is optimized over λ_i is similar to the one used in Rossi (2015) for the MNL and is given by

$$0.9 * \log f(y_i \mid \lambda_i, \gamma) + 0.1 * \frac{T_i}{\sum_i T_i} (-0.5 * z'z),$$

where $f(y_i \mid \lambda_i, \gamma)$ is the likelihood function of observing y_i conditional on $\beta_i = \lambda_i$ and γ , and $z = L(\lambda_i - \hat{\lambda})$ where $\hat{\lambda}$ is the pooled maximum likelihood estimate of λ , and L is the Cholesky decomposition of the negative Hessian of the pooled log-likelihood function at the maximum likelihood estimates.

probabilities are given by

$$\Pr[c_i=q|c_{-i},\lambda_i,\{\mu_q\}_q,\{\Sigma_q\}_q]=\begin{cases} \frac{n_{q,-i}f(\lambda_i|\mu_q,\Sigma_q)}{\sum_{r \in \mathcal{Q}_i} n_{r,-i}f(\lambda_i|\mu_r,\Sigma_r)+\alpha \int f(\lambda_i|\mu,\Sigma)f(\mu,\Sigma)d\mu d\Sigma}, & \text{if } q \in \mathcal{Q}_i, \\ \frac{\alpha \int f(\lambda_i|\mu,\Sigma)f(\mu,\Sigma)d\mu d\Sigma}{\sum_{r \in \mathcal{Q}_i} n_{r,-i}f(\lambda_i|\mu_r,\Sigma_r)+\alpha \int f(\lambda_i|\mu,\Sigma)f(\mu,\Sigma)d\mu d\Sigma}, & \text{if } q \notin \mathcal{Q}_i, \end{cases}$$

where $f(\lambda_i|\mu_q,\Sigma_q)$ is the density of a multivariate normal distribution with mean μ_q and covariance matrix Σ_q evaluated at λ_i , $f(\mu,\Sigma)$ is the prior density of a μ and Σ based on the prior distribution in Equation (2.7)-(2.8) and the marginal density of λ_i is given by

$$f(\lambda_i) = \int f(\lambda_i|\mu,\Sigma)f(\mu,\Sigma)d\mu d\Sigma = \left(\frac{d}{\pi(d+1)}\right)^{K_x/2} \frac{\Gamma_{K_x}((\nu+1)/2)}{\Gamma_{K_x}(\nu/2)} \frac{|\nu\nu I|^{\nu/2}}{|\hat{S}_i|^{(\nu+1)/2}},$$

where Γ_K is the multivariate Gamma function, $|\cdot|$ denotes the determinant and \hat{S}_i is the scale matrix of the distribution of Σ conditional on λ_i as given by

$$\hat{S}_i = \nu\nu I + (\lambda_i - \hat{\mu}_i)(\lambda_i - \hat{\mu}_i)' + d(\mu_0 - \hat{\mu}_i)(\mu_0 - \hat{\mu}_i)',$$

where $\hat{\mu}_i$ is the mean of the distribution of μ conditional on λ_i as given by

$$\hat{\mu}_i = \frac{d\mu_0 + \lambda_i}{d+1}.$$

For these derivations, we use the conjugacy of the normal-inverse Wishart prior on μ and Σ .

When in the multinomial distribution we draw a new component $c_i \notin \mathcal{Q}_i$, we also need to draw a new component mean μ_{c_i} and covariance matrix Σ_{c_i} . These are drawn from their posterior. For this purpose, we first draw Σ_{c_i} conditional on λ_i , and then μ_{c_i} conditional on Σ_{c_i} and λ_i . That is, we draw Σ_{c_i} from an inverse Wishart distribution with $\nu+1$ degrees of freedom and scale matrix \hat{S}_i . Next, we draw μ_{c_i} from a multivariate normal distribution with mean $\hat{\mu}_i$ and covariance matrix $(d+1)^{-1}\Sigma_{c_i}$.

2.A.2 Draw Σ_q and μ_q

We can jointly draw Σ_q and μ_q conditional on $\{\lambda_i\}_{i=1}^N$ and $\{c_i\}_{i=1}^N$ by first drawing Σ_q conditional on $\{\lambda_i\}_{i=1}^N$ and $\{c_i\}_{i=1}^N$ and then drawing μ_q conditional on Σ_q , $\{\lambda_i\}_{i=1}^N$ and $\{c_i\}_{i=1}^N$, for $q = 1, \dots, Q$.

We draw Σ_q from an inverse Wishart distribution with degrees of freedom $\nu + N_q$ and scale matrix

$$\hat{S}_q = \nu \nu I + \sum_{i=1}^N I[c_i = q](\lambda_i - \hat{\mu}_q)(\lambda_i - \hat{\mu}_q)' + d(\mu_0 - \hat{\mu}_q)(\mu_0 - \hat{\mu}_q)'$$

where N_q is the number of units in component q , and

$$\hat{\mu}_q = \frac{d\mu_0 + \sum_{i=1}^N I[c_i = q]\lambda_i}{d + N_q}.$$

Next, we draw μ_q from a multivariate normal distribution with mean $\hat{\mu}_q$ and covariance matrix $(d + N_q)^{-1}\Sigma_q$.

2.A.3 Draw λ_{ik}

We use a random walk Metropolis-Hastings step to draw λ_{ik} conditional on y_i , $\lambda_{i,-k}$, $\tau_{i,-k}$, c_i , μ_{c_i} , Σ_{c_i} , θ_k , and γ . Conditional on $\lambda_{i,-k}$ and $\tau_{i,-k}$, we know $\beta_{i,-k}$. Moreover, given $\lambda_{i,-k}$, μ_{c_i} and Σ_{c_i} , the prior for λ_{ik} is a univariate normal distribution with mean $\tilde{\mu}_{\lambda_{ik}}$ and variance $\tilde{\sigma}_{\lambda_{ik}}^2$ given by

$$\begin{aligned}\tilde{\mu}_{\lambda_{ik}} &\equiv E[\lambda_{ik} | \lambda_{i,-k}, \mu_{c_i}, \Sigma_{c_i}] = \mu_{c_i,k} + \Sigma_{c_i,k,-k} \Sigma_{c_i,-k,-k}^{-1} (\lambda_{i,-k} - \mu_{c_i,-k}), \\ \tilde{\sigma}_{\lambda_{ik}}^2 &\equiv \text{Var}(\lambda_{ik} | \lambda_{i,-k}, \Sigma_{c_i}) = \Sigma_{c_i,kk} - \Sigma_{c_i,k,-k} \Sigma_{c_i,-k,-k}^{-1} \Sigma_{c_i,-kk},\end{aligned}$$

where $\Sigma_{c_i,k,-k}$ refers to the k^{th} row of Σ and all columns except for the k^{th} .

The candidate for λ_{ik} is drawn from the normal distribution

$$\lambda_{ik}^* \sim N(\lambda_{ik}, \rho_{\lambda_{ik}}^2 \tilde{\sigma}_{\lambda_{ik}}^2),$$

where $\rho_{\lambda_{ik}}$ is a parameter to be tuned such the acceptance rate is about 0.44 (Roberts et al., 1997, Roberts, Rosenthal et al., 2001). Tuning is performed during the burn-in MCMC iterations. The candidate is accepted with probability

$$\min \left[1, \frac{f(y_i | \lambda_{ik}^*, \beta_{i,-k}, \theta_k, \gamma) f(\lambda_{ik}^* | \mu_{c_i}, \Sigma_{c_i}, \lambda_{i,-k})}{f(y_i | \lambda_{ik}, \beta_{i,-k}, \theta_k, \gamma) f(\lambda_{ik} | \mu_{c_i}, \Sigma_{c_i}, \lambda_{i,-k})} \right],$$

where the likelihood contribution conditional on λ_{ik} and $\beta_{i,-k}$ is given by

$$\begin{aligned}
 f(y_i|\lambda_{ik}, \beta_{i,-k}, \theta_k, \gamma) &= \sum_{\tilde{\tau}_{ik} \in \{\kappa, 1\}} f(y_i, \tilde{\tau}_{ik}|\lambda_{ik}, \beta_{i,-k}, \theta_k, \gamma), \\
 &= \sum_{\tilde{\tau}_{ik} \in \{\kappa, 1\}} \Pr[\tau_{ik} = \tilde{\tau}_{ik}|\theta_k] f(y_i|\lambda_{ik}, \tilde{\tau}_{ik}, \beta_{i,-k}, \gamma), \\
 &= \sum_{\tilde{\tau}_{ik} \in \{\kappa, 1\}} \theta_k^{I[\tilde{\tau}_{ik}=1]} (1 - \theta_k)^{I[\tilde{\tau}_{ik}=\kappa]} f(y_i|\tilde{\beta}_i, \gamma), \\
 &= \sum_{\tilde{\tau}_{ik} \in \{\kappa, 1\}} \theta_k^{I[\tilde{\tau}_{ik}=1]} (1 - \theta_k)^{I[\tilde{\tau}_{ik}=\kappa]} \left(\prod_{t=1}^{T_i} f(y_{it}|\tilde{\beta}_i, \gamma) \right),
 \end{aligned}$$

where $\tilde{\beta}_i$ has k^{th} element $\tilde{\tau}_{ik}\lambda_{ik}$ and $f(y_{it}|\tilde{\beta}_i, \gamma)$ is the likelihood contribution of observation t of unit i conditional on $\tilde{\beta}_i$ and γ given in Equation (2.1). For the prior density of λ_{ik} we have that

$$f(\lambda_{ik}|\mu_{c_i}, \Sigma_{c_i}, \lambda_{i,-k}) \propto \exp \left\{ -\frac{1}{2} \frac{(\lambda_{ik} - \tilde{\mu}_{\lambda_{ik}})^2}{\tilde{\sigma}_{\lambda_{ik}}^2} \right\}.$$

In case λ_i should be drawn jointly over all variables k , the candidate should be a multivariate normal distribution and the tuning parameter ρ_λ should be tuned such to obtain an acceptance rate of about 0.234 (Roberts et al., 1997, Roberts, Rosenthal et al., 2001).

2.A.4 Draw τ_{ik}

We draw τ_{ik} conditional on y_i , λ_i , $\tau_{i,-k}$, θ_k and γ using a Bernoulli distribution. The conditional probability that τ_{ik} is equal to 1 is given by

$$\begin{aligned}
 \Pr[\tau_{ik} = 1|y_i, \lambda_i, \tau_{i,-k}, \theta_k, \gamma] &= \frac{\Pr[\tau_{ik} = 1|\theta_k] f(y_i|\lambda_{ik}, \tau_{ik} = 1, \beta_{i,-k}, \gamma)}{\sum_{\tilde{\tau}_{ik} \in \{\kappa, 1\}} \Pr[\tau_{ik} = \tilde{\tau}_{ik}|\theta_k] f(y_i|\lambda_{ik}, \tilde{\tau}_{ik}, \beta_{i,-k}, \gamma)} \\
 &= \frac{\theta_k \left(\prod_{t=1}^{T_i} f(y_{it}|\beta_{ik} = \lambda_{ik}, \beta_{i,-k}, \gamma) \right)}{(1 - \theta_k) \left(\prod_{t=1}^{T_i} f(y_{it}|\beta_{ik} = \kappa\lambda_{ik}, \beta_{i,-k}, \gamma) \right) + \theta_k \left(\prod_{t=1}^{T_i} f(y_{it}|\beta_{ik} = \lambda_{ik}, \beta_{i,-k}, \gamma) \right)},
 \end{aligned}$$

where the likelihood contribution of observation t of unit i conditional on β_i and γ is given in Equation (2.1). Hence, we can draw a r_{ik} from a Bernoulli distribution with the probability above. Then, we obtain a draw of τ_{ik} by setting τ_{ik} equal to 1 when $r_{ik} = 1$ and equal to κ when $r_{ik} = 0$.

2.A.5 Draw θ_k

We can directly draw θ_k conditional on $\{\tau_{ik}\}_{i=1}^N$ from the Beta distribution

$$\theta_k | \{\tau_{ik}\}_{i=1}^N \sim \text{Beta} \left(a + \sum_i I[\tau_{ik} = 1], b + \sum_i I[\tau_{ik} = \kappa] \right),$$

for $k = 1, \dots, K_x$.

2.A.6 Draw γ

We use a random walk Metropolis-Hastings step to draw γ conditional on $\{y_i\}_{i=1}^N$, $\{c_i\}_{i=1}^N$, $\{\lambda_i\}_{i=1}^N$, and $\{\tau_i\}_{i=1}^N$. First notice that given $\{\lambda_i\}_{i=1}^N$, and $\{\tau_i\}_{i=1}^N$, we know $\{\beta_i\}_{i=1}^N$. At the s^{th} draw, the candidate for γ , γ^* , is drawn from

$$\gamma^* \sim \text{MVN}(\gamma^{(s-1)}, \rho_\gamma^2 \Sigma_{c_i}),$$

where $\gamma^{(s-1)}$ is the current draw for γ , and $\rho_{\lambda,i}$ is a parameter to be tuned such the acceptance rate is about 0.234 (Roberts et al., 1997, Roberts, Rosenthal et al., 2001).

The acceptance probability is given by

$$\min \left[1, \frac{f(y_i | \gamma^*, \beta_i) f(\gamma^*)}{f(y_i | \gamma^{(s-1)}, \beta_i) f(\gamma^{(s-1)})} \right],$$

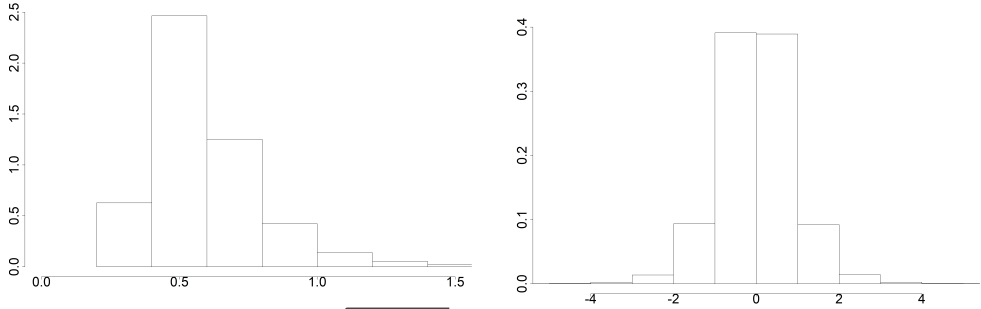
where

$$f(y_i | \gamma, \beta_i) = \prod_{t=1}^{T_i} f(y_{it} | \beta_i, \gamma),$$

and $f(\gamma)$ is the prior density of γ . In case γ^* is not accepted, we set $\gamma^{(s)} = \gamma^{(s-1)}$.

2.B Histograms of priors

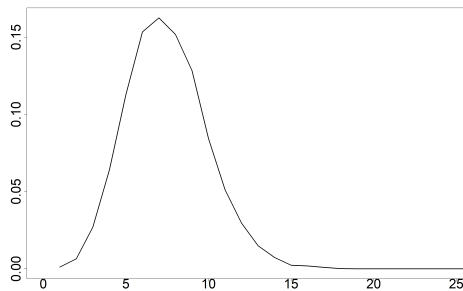
Figure 2.7: Priors μ_q and Σ_q



(a) Prior standard deviation $\sqrt{\text{Diag}(\Sigma_q)}$ in Monte Carlo study and empirical applications. This is the marginal density of the square root of a variance from the diagonal of a covariance matrix based on an $IW(\nu, \nu v I)$ distribution with $K = 3$, $\nu = K + 5$, and $v = 0.2$.

(b) Prior mean for μ_q , marginalized over Σ_q , in Monte Carlo study and empirical applications. This is the marginal density based on a $MVN(0, 0.5^{-1}\Sigma_q)$ prior for μ , $K = 3$, and the prior $\Sigma_q \sim IW(\nu, \nu v I)$ with $\nu = K + 5$, and $v = 0.2$.

Figure 2.8: Implied prior on number of components ($N = 1,000$ and $\alpha = 1$)



2.C Hit rates Monte Carlo study

Table 2.9: Percentage of units for which the posterior draw of β_{ik} falls within $-\epsilon \leq \beta_{ik} \leq \epsilon$ for multiple values of ϵ (averaged over Monte Carlo replications, draws and variables).

ϵ	True	(1) All		(2) True $-\epsilon \leq \beta_{ik} \leq \epsilon$		(3) True $\beta_{ik} < -\epsilon$ or $\beta_{ik} > \epsilon$		(4) True $\beta_{ik} = 0$		
	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-DPM	DPM	HVS-DPM	DPM	
<i>DGP 2</i>										
0.00	0	2	2	~	~	~	2	2	-	-
0.10	3	5	3	16	13	17	4	3	-	-
0.20	7	8	6	26	24	26	5	5	-	-
0.30	10	11	11	35	33	37	8	7	-	-
0.40	14	15	15	42	41	44	11	10	-	-
0.50	19	21	20	47	46	50	14	14	-	-
0.75	37	39	39	61	61	61	26	26	-	-
1.00	59	59	59	74	75	72	36	36	-	-
1.50	86	83	84	90	91	88	39	39	-	-
2.00	94	93	93	96	96	97	58	59	-	-
2.50	98	98	98	98	98	99	79	79	-	-
<i>DGP 3</i>										
0.00	25	16	19	33	0	40	11	0	33	40
0.10	28	21	24	39	15	45	13	5	40	15
0.20	30	25	28	46	29	51	16	10	47	29
0.30	33	30	33	52	42	57	18	15	54	42
0.40	36	35	38	59	53	62	22	21	61	53
0.50	40	41	40	64	61	67	25	26	67	63
0.75	53	56	58	75	76	75	36	38	81	81
1.00	70	71	72	83	84	81	43	45	90	91
1.50	90	88	89	93	93	92	45	47	98	98
2.00	95	95	95	97	97	98	62	63	100	100
2.50	98	98	98	99	99	100	79	79	100	100
<i>DGP 4</i>										
0.00	10	14	14	25	0	26	12	0	25	0
0.10	17	20	9	34	16	35	17	8	34	15
0.20	24	26	19	45	32	45	20	15	43	29
0.30	31	33	28	54	46	54	24	20	51	43
0.40	39	40	38	61	57	61	27	26	60	54
0.50	47	48	47	68	66	66	30	31	67	64
0.75	66	66	68	80	81	80	31	30	82	82
1.00	81	81	82	88	89	88	51	54	91	91
1.50	97	96	96	97	97	97	77	76	99	99
2.00	100	100	99	100	99	100	94	93	100	100
2.50	100	100	100	100	100	100	99	99	100	100

Results for four different groups: (1) all units, (2) units for which true β_{ik} falls within interval, (3) units for which true β_{ik} does not fall within interval, (4) units for which true $\beta_{ik} = 0$ ($\tau_{i,k} = 0$).

Chapter 3

A multinomial and rank-ordered logit model with inter- and intra-individual heteroscedasticity

3.1 Introduction

Understanding and predicting choices of individuals is important for numerous applications. These applications include predicting product demand, designing effective policies, and constructing meaningful product recommendations. In many cases, individuals are observed while making repeated choices over time. For example when responding to survey questions, choosing supermarket products across different visits, or completing a ranking by consecutively choosing best, second best, et cetera. To deduce an individual's preferences based on observed discrete choices, the (multinomial) logit model is often employed.

The logit model is based on a utility framework: an individual obtains utility from choosing a certain alternative/option and chooses the alternative which gives the highest utility (Manski, 1977). The utility is comprised of an explained part (the preferences/signal) and an unexplained part (the noise). The noise captures that the

actual choice can differ from the choice that yields the highest signal. Hence, the noise can capture that (i) the signal fails to capture all preferences of an individual, and/or (ii) an individual can make ‘mistakes’ and choose an alternative that does not accord with her underlying preferences. Logit models assume an extreme value distribution for the noise.

The logit model has been extended in many ways to realistically capture certain aspects of individual behavior. One such aspect of behavior is that the randomness in the choice-making of individuals may vary over time. For example, during surveys, individuals may become fatigued and start responding more randomly to questions as the survey proceeds. Or when completing a ranking amongst multiple alternatives, individuals may be unable to accurately assign middle and bottom ranks, due to the required cognitive effort or lack of information. For supermarket purchases, an individual that is new to a certain product category (e.g. diapers) may at first pick alternatives quite randomly after which the preferences are learned and choices are more and more based on the underlying preferences.

The heteroscedastic logit model is able to estimate individual preferences while accounting for changes in the randomness in choices (Hausman and Ruud, 1987, Bradley and Daly, 1994). For this purpose, the model explicitly allows for changes in the relative importance of the explained and the unexplained part of utility (the signal-to-noise ratio). When choices become more random, this can be captured in the unexplained part becoming more dominant. Mathematically, the heteroscedastic logit model allows for changes in the signal-to-noise ratio via a time-varying scale parameter in the unexplained part of the utility specification.

The main drawback of the standard heteroscedastic logit model is that the scale parameter is specified at the population-level. Hence, the model assumes that the changes in the randomness in decision-making is equal across individuals, thereby only allowing for within-individual (intra-individual) heteroscedasticity. In the context of the earlier examples, this implies that each individual is assumed to become fatigued at the same time, or assumed to be able to accurately assign exactly the same ranks.

In this paper, we generalize the heteroscedastic logit model by allowing for differences across individuals in the changes in the scale parameter. That is, we allow for intra- and inter-individual heteroscedasticity (or *heterogeneous heteroscedasticity*): each individual has her own sequence of scale parameters over time, and the time-variation in the scale parameters can differ across individuals. For example, for some

individuals the scale parameters may stay constant, for others the scale parameters may increase several times, and for again some others the scale parameters may first decrease and then increase.

In case such individual differences exist, using an individual-level instead of a population-level approach is beneficial for several reasons. First, existing population-level approaches generally lead to biased estimators for the preference parameters. That is, there will be a bias towards zero, because at each time period a number of individuals could be answering more randomly.¹ Second, population-level approaches make inefficient use of data. This is because it is assumed that at each time period, each individual provides the same amount of information in her choices. Finally, population-level approaches only give insight into the average time-variation in the scale parameter. In some cases, one might find a constant average scale parameter while in reality there is heterogeneous heteroscedasticity. An individual-level approach also gives more insight into the behavior of different individuals. To allow for individual differences, some structure is needed to model the heteroscedastic process.

We develop a multinomial logit model (MNL) and a rank-ordered logit model (ROL) that allow for heterogeneous heteroscedasticity. For this purpose, we include individual- and time/rank-specific scale parameters. We let the dynamics in the sequence of an individual's scale parameters be governed by a Markov process. We also allow for unobserved preference heterogeneity. For inference, we develop a maximum simulated likelihood estimation approach.

The Markov process assumes that, for subsequent choices to make or for consecutive ranks to assign, an individual can go through a number of phases. Each phase is marked by a different scale parameter. When an individual moves to a phase with higher scale, the choices become more random. When an individual moves to a phase with lower scale, the choices become more predictable and more in line with the underlying preferences. For the example of fatigue during surveys, there may exist phases with a rather high scale. Respondents who become fatigued, enter these phases with high scale and answer more randomly as the survey proceeds. Respondents that do not become fatigued, remain in a phase with low scale.

In the literature, a related individual-level ROL has been proposed in Fok et al. (2012). They propose a latent class ROL where they allow for individuals to have

¹Even when no individual differences exist, the existing estimators for the preference parameters in the heteroscedastic logit model (Bradley & Daly, 1994) are often biased away from zero, because the preference parameters are scaled such that the time period with lowest estimated signal-to-noise ratio has a scale of one. The estimator for our proposed model does not suffer from this shortcoming.

different ranking abilities: different individuals may be able to assign a different number of top ranks accurately. When applied to the ROL, our approach can be seen as a generalization of Fok et al. (2012). First, we allow for unobserved preference heterogeneity and allow the model to be used for panel data. Especially in the context of non-constant scale, allowing for heterogeneity is important to avoid spurious findings. That is, when preference heterogeneity is unaccounted for, an individual who has preferences that deviate from the “average” individual is likely falsely classified as having a large scale parameter. Second, Fok et al. (2012) allow for individuals to assign a specific rank either accurately or completely randomly. Instead, our model allows for the decisions of individuals to be more in between, which is also possible in the standard heteroscedastic ROL. As a consequence, our model is a generalization of the heteroscedastic ROL, whereas the latent class ROL is not. Finally, our model straightforwardly and parsimoniously allows for individuals that might rank the middle ranks randomly, but both the top and bottom ranks accurately. This possibility was already provided as an extension in Fok et al. (2012) but requires work on top of the basic model specification.

We illustrate the usefulness of the newly proposed hidden Markov model specifications using a Monte Carlo study and two empirical applications. In the Monte Carlo study, we find that our proposed model works well and that the estimator seems unbiased in various settings. Furthermore, this study clearly illustrates the bias in the estimator for the preference parameters for the standard heteroscedastic logit model. Depending on the data generating process, the bias is either towards zero due to neglecting individual differences, or away from zero due to scaling the preference parameters based on the minimum of the estimated scale parameters. Furthermore, the estimator for the standard MNL is biased towards zero in case heteroscedasticity is present, because heteroscedasticity leads to more random-looking choice-making of respondents. Our proposed estimator and model alleviate these biases.

In the first empirical application, we consider binomial choices during a discrete choice experiment on healthy food choices. We allow for multiple phases to capture possible learning and fatigue effects. We find that accounting for individual differences in learning and fatigue leads to a much better fit of the data, while needing less free model parameters than the standard heteroscedastic logit model. In the second empirical application, we consider rank-ordered data from a survey on political preferences to capture possible differential capabilities in ranking. Again, allowing for individual differences in the dynamics of the heteroscedasticity leads to a much better

fit of the data.

This paper is set up as follows. In Section 4.2, we discuss the background and related literature. In Section 3.3, we develop the hidden Markov MNL and ROL, and discuss identification and estimation. In Section 3.4, we report the results of a Monte Carlo study. In Sections 3.5 and 3.6, we report the results of the two empirical applications. Finally, we provide a discussion and conclude.

3.2 Background

In this section, we discuss the additive random utility framework (ARUM) we employ in our paper. This framework is central in deciding how to model individual-specific dynamics in the signal-to-noise ratio. We illustrate the identification problem that may arise, and discuss related papers that have proposed solutions to this. We also indicate how our approach differs from current specifications dealing with individual-specific dynamics in the signal-to-noise ratio.

The additive random utility framework of Manski (1977) is a useful and popular tool to model choices of individuals. It relies on the assumption that an individual obtains utility from a certain alternative and that an individual chooses the alternative that gives the highest utility. The utility is assumed to be an additive function of the signal (based on observed variables and unobserved parameters) and some noise

$$\text{Utility} = \text{Signal} + \text{Noise}.$$

Mathematically, we can write this utility specification in a general form as

$$U_{itj} = x'_{itj}\beta_{it} + \sigma_{it}\varepsilon_{itj}, \quad (3.1)$$

where U_{itj} is the (unobserved) utility that individual i obtains from choosing alternative j at time t , x_{itj} is a vector of covariates representing the attributes of alternative j , β_{it} is a vector with preference parameters of individual i at time t , $\sigma_{it} > 0$ is a scale parameter for individual i at time t , and ε_{itj} is an i.i.d. error term with fixed variance.² The individual chooses the alternative that gives the highest utility. The multinomial logit and probit models are special cases of the ARUM.

Because only choices are observed and not utility, and the scale parameter does not

²A more general specification can be obtained by allowing for correlation across the error terms ε_{itj} over individuals, time periods, and/or alternatives.

vary over alternatives, we obtain an equivalent model for choices by rescaling the utility

$$U_{itj}^* = x'_{itj} \frac{\beta_{it}}{\sigma_{it}} + \varepsilon_{itj}, \quad (3.2)$$

where now the alternative is chosen with the highest scaled utility $U_{itj}^* = U_{itj}/\sigma_{it}$. The equivalence between the utility specifications in Equations (3.1) and (3.2) imply that only the signal-to-noise ratio (β_{it}/σ_{it}) is identified, and not the absolute values of the signal and the noise. Hence, if we would allow for both $\beta_{it} = \beta_i$ and $\sigma_{it} = \sigma_i$ to be individual-specific, separate identification of the two parameters can only come from distributional assumptions on these two parameters (Hess & Rose, 2012). The same holds when we allow both $\beta_{it} = \beta_t$ and $\sigma_{it} = \sigma_t$ to be time-dependent.

Therefore, for identification, the proposed models in the literature often allow for heterogeneity and time-variation in either β_{it} or σ_{it} . For example, the heteroscedastic multinomial and rank-ordered logit models allow for a (possibly) individual-specific β_i and a time-dependent scale σ_t (Hausman and Ruud, 1987, Bradley and Daly, 1994, DeSarbo et al., 2004).

We generalize the heteroscedastic multinomial logit model by allowing for the time-variation in the scale to be different across individuals (σ_{it}). We allow for individual-specific preference parameters in β_i , but exclude time-variation in this parameter. We ensure identification by letting the sequence of scale parameters of an individual $\{\sigma_{i1}, \sigma_{i2}, \dots\}$ be governed by a Markov process with the scale of one state normalized at one, as will be shown later.³

Alternatively, one can allow for individual-specific heteroscedasticity by letting β_{it} be individual- and time-specific, and keeping $\sigma_{it} = \sigma$ constant over individuals and time. An advantage of this alternative approach is that it can model the change in the signal-to-noise ratio to be different across attributes, and can thus capture choice strategies where choices are made based on different subsets of attributes as time progresses or where preferences change over time. The main disadvantage is that, in small T settings, estimation uncertainty and overfitting become problematic.

There are three papers that propose discrete choice models with individual-specific

³Bhat and Castelar (2002) propose a multinomial logit model with individual-specific preference parameters β_i and individual- and time-specific scale parameters σ_{it} . However, their formulation is highly restrictive. The scale parameter σ_{it} can take on only one of two values and which of the two values it takes on is determined deterministically: $\sigma_{it} = 1$ in case observation t of individual i corresponds to a revealed preference observation, and $\sigma_{it} = \lambda$ in case it corresponds to a stated preference observation, with λ a parameter to be estimated. Hence, the variation in scale parameters only allows for the scale to be different across different types of data.

time-variation in β_{it} : Hess and Rose (2009), Bhat and Sidharthan (2011), and Danaf et al. (2020). These three papers all propose a model with constant scale σ and preference parameters of the form $\beta_i + \beta_{it}$. Both β_i and β_{it} are allowed to follow arbitrary distributions, with the restriction that the unconditional mean of β_{it} is zero. The papers differ in the type of model (logit versus probit), estimation approach and the distributional form used for β_i and β_{it} . These approaches are quite general, but have the main disadvantage that they assume an additive specification $\beta_i + \beta_{it}$. For individual-specific heteroscedasticity in discrete choice models, a multiplicative specification via $\beta_i\beta_{it}$ (or β_i/σ_{it}) is more suitable, as choices becoming more random directly affect the signal-to-noise *ratio*. That is, increased randomness in choice-making leads to signal-to-noise ratios that become closer to zero. In a multiplicative specification, this can be modeled by a low β_{it} (or a high σ_{it}). Instead, with an additive specification, a given β_{it} could shrink the $\beta_i + \beta_{it}$ of one individual to zero, whereas for another individual it can make it more extreme or let it flip signs. Due to the additive nature of these approaches, they are less suited to model heterogeneous heteroscedasticity. Instead, we use a multiplicative specification.

A related strand of literature considers (time-invariant) scale heterogeneity: some individuals may choose more randomly throughout the observed period than others. Fiebig et al. (2010) propose a so-called generalized multinomial logit model that includes both individual-specific preferences β_i and an individual-specific scale parameter σ_i . Separate identification of the two parameter is achieved by imposing parametric population distributions on β_i and σ_i (Hess & Rose, 2012). Our approach differs crucially as we focus on the time-variation in the scale parameter, to allow for changes in individual behavior over time.

3.3 Methodology

In this section, we develop the hidden Markov multinomial logit model and the hidden Markov rank-ordered logit model to capture inter- and intra-individual heteroscedasticity. The methods are highly similar, the main difference is that for the MNL the heteroscedasticity refers to the change in the scale parameter as time progresses, and for the ROL the heteroscedasticity refers to the change in the scale parameter across consecutive ranks for the same ranking task.

Let us introduce some basic notation. We index the individuals by $i = 1, \dots, N$, the observations for individual i by $t = 1, \dots, T$, and the alternatives that individual i can

choose between, or need to rank, at time t by $j = 1, \dots, J$.⁴ Furthermore, we denote by x_{itj} a $(K \times 1)$ vector of covariates representing the attributes of alternative j at time t for alternative j , and by β_i a $(K \times 1)$ vector with individual-specific preference parameters corresponding to x_{itj} .

3.3.1 Hidden Markov multinomial logit model

For the multinomial logit model, we let the scalar $y_{it} \in \{1, 2, \dots, J\}$ denote the alternative that individual i chooses at time t , and let Y_{it} denote the corresponding random variable. The latent utility that individual i obtains from choosing alternative j at time t is given by

$$U_{itj} = x'_{itj}\beta_i + \sigma_{it}\varepsilon_{itj}, \quad (3.3)$$

where $\sigma_{it} > 0$ is an individual- and time-specific scale parameter and the error terms ε_{itj} follow independent type I extreme value distributions with location 0 and scale 1. In case the scale parameter is equal across individuals ($\sigma_{it} = \sigma_t$), we obtain the heteroscedastic multinomial logit model. In case the scale parameter is also equal over time ($\sigma_{it} = \sigma = 1$), we obtain the standard (mixed) multinomial logit model.

At each time t , an individual chooses the alternative that yields the highest utility. Given the utility specification in Equation (3.3), it follows that the conditional probability that individual i chooses alternative j at time t is given by (McFadden, 1973)

$$\Pr[Y_{it} = j | \beta_i, \sigma_{it}] = \frac{\exp\left(\frac{1}{\sigma_{it}}(x'_{itj}\beta_i)\right)}{\sum_{l=1}^J \exp\left(\frac{1}{\sigma_{it}}(x'_{itl}\beta_i)\right)}. \quad (3.4)$$

The scale parameter σ_{it} captures heteroscedasticity. The higher σ_{it} , the lower the signal-to-noise ratio and the more random the choice of individual i at time t becomes. For example, when an individual becomes tired during a survey and starts to answer more randomly, this can be modeled by a sequence of scales $\{\sigma_{it}\}_{t=1}^T$ that increases over time. In the extreme case that σ_{it} tends to infinity, the choice becomes completely random. In the other extreme case that σ_{it} is close to 0, the choice can be perfectly explained by the signal $x'_{itj}\beta_i$.

We let the time variation in the sequence of an individual's scale parameters $\{\sigma_{it}\}_{t=1}^T$ be governed a Markov process. Such a process assumes that while an individual is

⁴In this notation, the number of observations T is equal across individuals and the number of alternatives J is equal across observations and individuals. These assumptions can be easily relaxed.

making choices, she can go through a number of phases. Each phase is marked by a different scale parameter. When an individual moves to a phase with higher scale, the choices become more random. When an individual moves to a phase with lower scale, the choices become more predictable and more in line with the underlying preferences. For the example of fatigue during surveys, there may exist phases with a rather high scale. Respondents who become fatigued, enter these phases with high scale as they answer more randomly as the survey proceeds. Respondents that do not become fatigued, remain in a phase with low scale.

Let M denote the number of possible phases an individual can go through, with M set by the researcher. The number of different scale parameters is equal to M : $\sigma_{it} \in \{\tilde{\sigma}_1, \tilde{\sigma}_2, \dots, \tilde{\sigma}_M\}$. Let s_{it} denote the phase that an individual i is in at time t . Then we have

$$\sigma_{it} = \tilde{\sigma}_{s_{it}}. \quad (3.5)$$

For parameter identification, the scale parameter of one of the phases needs to be fixed. This fixed scale parameter can be set to 1, such that the preference parameters β_i can be interpreted with respect to the corresponding phase.

The phase indicators $\{s_{it}\}_{t=1}^T$ describe how individual i moves through the M phases. These indicators are unobserved. We let the time variation in $\{s_{it}\}_{t=1}^T$ follow a first-order Markov process (Goldfeld & Quandt, 1973). Such a process describes how individuals move from one phase to another using transition probabilities. We denote the transition probabilities by

$$q_{mnt} \equiv \Pr[S_{i,t+1} = n | S_{it} = m], \quad (3.6)$$

which is the probability that individual i is in phase n at time $t + 1$ given that she was in phase m at time t , and where S_{it} denotes the random variable associated with outcome s_{it} , for $m, n = 1, \dots, M$ and $t = 1, \dots, T - 1$. We have that $0 \leq q_{mnt} \leq 1$ and $\sum_{n=1}^M q_{mnt} = 1$. Finally, we denote the initial phase probabilities by

$$\pi_m \equiv \Pr[S_{i1} = m], \quad (3.7)$$

with $0 \leq \pi_m \leq 1$ and $\sum_{m=1}^M \pi_m = 1$.

Depending on the information in the data and the type of application, it may be desired to impose restrictions on the parameters of the Markov process. For example, one can restrict the transition probabilities such that an individual can either stay

in the current phase or move one phase up, also known as a change-point model (Chib, 1998). Furthermore, one may wish to restrict (some of) the values of the M different scale parameters. In case the dataset contains relatively few observations per individual, it is important to have only few phases M , e.g. $M \leq 4$. This helps to avoid overfitting, in particular finding a perfect fit phase in which the choices can be seemingly perfectly explained by the signal.

Finally, the parameters in β_i capture the preferences of individual i for the attributes in x_{itj} . We let β_i follow some distribution with density $f(\beta_i|\theta)$ where θ denotes the set of population parameters to be estimated. The type of distribution for β_i should be set by the practitioner. Examples are the multivariate normal distribution (where θ represents the mean and covariance matrix), lognormal distribution, a mixture of discrete distributions, and a mixture of normal distributions. The parameters of the distribution could also be allowed to depend on individual-specific characteristics. Moreover, in case one has large T per individual, one can directly estimate β_i without imposing a population distribution.

For identification, with sufficient variation in the variables x_{itj} , a sufficient condition for θ to be identified is that the total number of observations in the phase with fixed variance exceeds the number of parameters in θ . Additional observations in each phase are needed to identify the parameters of the Markov process.

3.3.2 Hidden Markov rank-ordered logit model

Next, we generalize the hidden Markov multinomial logit model in Section 3.3.1 to allow for rank-ordered choices. That is, we now model a partial or complete ranking over the alternatives instead of only the most preferred alternative. As the ROL uses more information than the MNL, the ROL allows for more efficient use of data. We provide the model specification for panel data but the model can also be used for cross-sectional data with $T = 1$.

Let the vector y_{it} denote the complete ranking provided by individual i at time t out of the J alternatives, and Y_{it} the corresponding random variable.⁵ That is, $y_{it} = (y_{it1}, y_{it2}, \dots, y_{itJ})'$, and y_{itj} denotes the alternative that was ranked j^{th} . For example, in case alternative three was ranked first, we have that $y_{it1} = 3$. Note that the MNL only models the first-ranked alternative y_{it1} .

⁵The specification can be easily extended to problems in which only the top J^* alternatives out of J alternatives need to be ranked.

In the rank-ordered logit model, jointly modeling the complete ranking of alternatives (y_{it}) is equivalent to modeling the sequential ranking from the highest rank (y_{it1}) to the lowest rank (y_{itJ}) (Beggs et al., 1981, Chapman and Staelin, 1982). For each rank h , the choice between the “remaining” alternatives $\{y_{itl}\}_{l=h}^J$ can be modeled with a multinomial logit model. That is, the probability of observing y_{it} has the form

$$\begin{aligned} \Pr[Y_{it} = y_{it} | \beta_i] &= \prod_{h=1}^{J-1} \Pr[Y_{ith} = y_{ith} | y_{it1}, \dots, y_{it,h-1}, \beta_i] \\ &= \prod_{h=1}^{J-1} \frac{\exp(x'_{ity_{ith}} \beta_i)}{\sum_{l=h}^J \exp(x'_{ity_{itl}} \beta_i)}. \end{aligned}$$

To allow for intra-individual heteroscedasticity — individuals may be more or less capable to assign the top ranks as compared to the middle and bottom ranks — Hausman and Ruud (1987) propose a heteroscedastic ROL. For this purpose, they introduce a scale parameter σ_h that may differ over ranks h .⁶ More specifically, in the heteroscedastic ROL the probability of observing y_{it} is given by

$$\Pr[Y_{it} = y_{it} | \beta_i, \sigma_1, \sigma_2, \dots, \sigma_{J-1}] = \prod_{h=1}^{J-1} \frac{\exp\left(\frac{1}{\sigma_h} (x'_{ity_{ith}} \beta_i)\right)}{\sum_{l=h}^J \exp\left(\frac{1}{\sigma_h} (x'_{ity_{itl}} \beta_i)\right)}.$$

The higher σ_h , the more random the assignment to rank h . Hence, a high σ_h indicates that individuals find it relatively difficult to assign an alternative to rank h .

We extend the approach of Hausman and Ruud (1987) to additionally allow for inter-individual heteroscedasticity: the ranking capabilities may differ across individuals. More specifically, we let the probability of observing a complete ranking y_{it} be given by

$$\Pr[Y_{it} = y_{it} | \beta_i, \sigma_{i1}, \sigma_{i2}, \dots, \sigma_{i,J-1}] = \prod_{h=1}^{J-1} \frac{\exp\left(\frac{1}{\sigma_{ih}} (x'_{ity_{ith}} \beta_i)\right)}{\sum_{l=h}^J \exp\left(\frac{1}{\sigma_{ih}} (x'_{ity_{itl}} \beta_i)\right)}, \quad (3.8)$$

where intra- and inter-individual heteroscedasticity is allowed for via the rank- and individual-specific scale parameter σ_{ih} .

As with the hidden Markov MNL in Section 3.3.1, we let the sequence of an indi-

⁶In their paper, Hausman and Ruud (1987) use the notation σ_h to denote the inverse of the scale parameter.

vidual's scale parameter $\{\sigma_{ih}\}_{h=1}^{J-1}$ be governed by a Markov process. This implies that we explicitly allow for "blocks" of consecutive ranks to be assigned based on the same amount of randomness. The sizes and locations of these blocks may differ across individuals. For example, some individuals may assign the top three and bottom three ranks accurately and the middle ranks more randomly. Others, may assign the top two ranks and the lowest rank accurately, and the remainder more randomly. Our model allows for all these individual differences.

We let β_i follow a distribution with density $f(\beta_i|\theta)$ and use a Markov process with M phases to govern the dynamics in $\{\sigma_{ih}\}_{h=1}^{J-1}$. That is, we have the latent phase indicator s_{ih} denoting the phase that individual i is in when assigning rank h , and we have that $\sigma_{ih} = \tilde{\sigma}_{s_{ih}}$. We let s_{ih} follow a first-order Markov process with transition and initial phase probabilities

$$q_{mnh} \equiv \Pr[S_{i,h+1} = n | S_{ih} = m], \quad (3.9)$$

$$\pi_m \equiv \Pr[S_{i1} = m], \quad (3.10)$$

with $0 \leq q_{mnh} \leq 1$, $\sum_{n=1}^M q_{mnh} = 1$ for $m = 1, \dots, M$ and $h = 1, \dots, J-2$, $0 \leq \pi_m \leq 1$ and $\sum_{m=1}^M \pi_m = 1$.

Our hidden Markov ROL generalizes the latent class ROL of Fok et al. (2012). To see the equivalence: the latent class ROL has a parameter p_j denoting the proportion of individuals that can rank exactly the first j alternatives correctly and the remaining $J-j$ alternatives randomly. Hence, the hidden Markov ROL is equivalent to the latent class ROL in case we take two phases with $\tilde{\sigma}_1 = 1$ and $\tilde{\sigma}_2 = \infty$, and do not allow individuals to move from phase two to phase one ($q_{22h} = 1$ for all h). Then $p_0 = \pi_2$, $p_1 = \pi_1 q_{121}$ and $p_j = \pi_1 q_{12j} \prod_{h=1}^{j-1} q_{11h}$ for $j = 2, \dots, J-1$. Also, equivalently to testing for an empty class in the latent class ROL ($p_j = 0$) one can test $\pi_1 = 0$ (class 0) or $q_{11,j-1} = 1$ (classes 1 up to $J-1$). Moreover, with the hidden Markov ROL one can test for equal transition probabilities across ranks ($q_{11,j} = q_{11,j+1}$).

3.3.3 Parameter estimation

To estimate the parameters of the hidden Markov MNL (HM-MNL) in Equations (3.3)-(3.7) and of the hidden Markov ROL (HM-ROL) in Equations (3.8)-(3.10), we rely on maximum simulated likelihood estimation. The likelihood functions of the

models are given by

$$\begin{aligned}
 p(y|\theta, q, \pi, \tilde{\sigma}) &= \prod_{i=1}^N p(y_i|\theta, q, \pi, \tilde{\sigma}) \\
 &= \prod_{i=1}^N \left[\int p(y_i|\beta_i, q, \pi, \tilde{\sigma}) f(\beta_i|\theta) d\beta_i \right] \\
 &= \prod_{i=1}^N \left[\int \left(\sum_{s_i^* \in \mathcal{S}} \Pr[S_i = s_i^* | q, \pi] p(y_i|\beta_i, \tilde{\sigma}, s_i^*) \right) f(\beta_i|\theta) d\beta_i \right], \quad (3.11)
 \end{aligned}$$

where $y_i = \{y_{it}\}_{t=1}^T$, $y = \{y_i\}_{i=1}^N$, $s_i = \{s_{it}\}_{t=1}^T$ (HM-MNL), $s_i = \{s_{ih}\}_{h=1}^{J-1}$ (HM-ROL), \mathcal{S} is a set of all possible sequences of phases $s_i \in \mathcal{S}$, and

$$p(y_i|\beta_i, \tilde{\sigma}, s_i) = \begin{cases} \prod_{t=1}^T \frac{\exp\left(\frac{1}{\tilde{\sigma}_{s_{it}}}(x'_{itj}\beta_i)\right)}{\sum_{l=1}^J \exp\left(\frac{1}{\tilde{\sigma}_{s_{it}}}(x'_{itl}\beta_i)\right)}, & \text{(HM-MNL),} \\ \prod_{t=1}^T \prod_{h=1}^{J-1} \frac{\exp\left(\frac{1}{\tilde{\sigma}_{s_{ih}}}(x'_{ity_{ith}}\beta_i)\right)}{\sum_{l=h}^J \exp\left(\frac{1}{\tilde{\sigma}_{s_{ih}}}(x'_{ity_{itl}}\beta_i)\right)}, & \text{(HM-ROL).} \end{cases}$$

The expression to sum over all possible sequences $s_i^* \in \mathcal{S}$ in Equation (3.11) seems computationally intensive. However, it can be rewritten as a sequential filter which is computationally efficient (Hamilton, 1989), see Equations (3.12) and (3.13) in Appendix 3.A. Moreover, the probability of observing a sequence s_i^* is a straightforward function of q and π .

For a general density $f(\beta_i|\theta)$, the integral in the likelihood function in Equation (3.11) cannot be solved analytically. To approximate the integral, we use Monte Carlo integration. That is, we obtain R draws $\beta_i^{(r)}$ from a distribution with density $f(\beta_i|\theta)$ and approximate the integral by the average of $p(y_i|\beta_i^{(r)}, q, \pi, \tilde{\sigma})$ over these R draws, see appendix 3.A for more details. We use scrambled Halton draws to ensure good coverage of $f(\beta_i|\theta)$ (Bhat, 2003, Bhat, 2001, Braaten and Weller, 1979). In case the distribution over β_i is taken to be discrete, the integral over β_i can be written as a sum and the log-likelihood function can be directly maximized without needing Monte Carlo integration.

The use of the sequential filter allows us to directly maximize the (simulated) log-likelihood function without needing to augment the likelihood function with s_{it} (or s_{ih}) to enable an Expectation Maximization (EM) type of algorithm (Dempster et

al., 1977, Hamilton, 1990). This direct maximization requires less computations in a single iteration of the optimization than an EM algorithm, and also does not depend on a given draw of s_{it} (or s_{ih}) which may possibly slow down convergence due to the extra iterations needed.

Specialized code is written in C++ and R (R Core Team, 2013, Eddelbuettel and François, 2011) to obtain the scrambled Halton draws and to evaluate the (simulated) log-likelihood function and compute its analytic gradients. The details are given in Appendix 3.A. We take the standard errors equal to the square root of the diagonal elements of the inverse of the negative Hessian of the log-likelihood function. We approximate the Hessian using the outer-product-of-gradients approximation.

The probability that an individual i is in a phase m at time t conditional on observed choices y_i , $\Pr[S_{it} = m|y_i, \theta, q, \pi, \tilde{\sigma}]$, can be computed after the maximum likelihood estimates have been obtained. Details are in Appendix 3.B.

3.4 Monte Carlo study

In this section, we illustrate the performance of our hidden Markov multinomial logit model with a Monte Carlo study. The study consists of two parts. We first evaluate the small-sample performance of the model and estimator under correct model specification. Next, we evaluate the performance of the model under model misspecification.

For the first part of the study, we consider three data generating processes (DGPs). We use 1,000 Monte Carlo replications per DGP. For each DGP, we consider 1,000 individuals, 15 observations per individual, and 2 alternatives per observation. We consider three explanatory variables: x_{1itj} , x_{2itj} from a standard normal distribution and x_{3itj} from a Bernoulli distribution with probability 0.5 of outcome one. Furthermore, in the DGPs we draw the individual-specific preference parameters from a multivariate normal distribution

$$\beta_i \sim MVN(b, \Sigma_\beta),$$

where Σ_β is a positive definite covariance matrix.

In the first DGP, the HM-MNL is the true model. In this DGP, we aim to mimic the possible learning and fatigue behavior that individuals may experience when completing a survey. We use three phases $\tilde{\sigma} = (\infty, 1, \infty)$. Individuals in the first phase

still need to learn (e.g. about their preferences) and answer randomly, individuals in the second phase answer most accurately according to their true preferences (the *minimum variance phase*), and individuals in the third phase answer randomly due to fatigue. We consider initial phase probabilities $\pi = (0.2, 0.7, 0.1)$, and transition probabilities $q_{11t} = 0.50$ and $q_{22t} = 0.99$ for all t . Based on π and q , the percentage of observations in phases one to three are 2.7%, 81.6%, and 15.8%, respectively. Furthermore, 21.5% of individuals reach phase three. At $t = 5$, the percentage of individuals in the minimum variance phase (phase 2) is largest: 85.6%. Finally, we take Σ_β diagonal.

The second and third DGPs are altered versions of the first DGP. In DGP two, the true model is the MNL: we set $\pi_2 = 1$ and $q_{22t} = 1$ for all t . In DGP three, instead of a diagonal covariance matrix as in DGP one, we add correlation across the preference parameters by letting Σ_β be a full positive definite covariance matrix with implied correlations ρ_β . These correlations can capture time-invariant scale heterogeneity: some individuals may choose more randomly throughout the survey than others. Time-invariant scale-heterogeneity shows itself in preference parameters of the same individual to either all tend to more extreme values than b or to all tend to 0. To incorporate time-invariant scale-heterogeneity in the DGP, we let the implied correlations have an absolute level of 0.7. That is, when two b parameters are both positive or both negative we set the correlation to 0.7, when one of them is positive and the other negative we set the correlation to -0.7.

For each replication, we estimate the parameters of three models: (i) a MNL, (ii) a heteroscedastic MNL (H-MNL)⁷, and (iii) our HM-MNL. For all three models, we let $\beta_i \sim MVN(b, \Sigma_\beta)$ and use 250 scrambled Halton draws per individual. For the H-MNL, we fix $\sigma_1 = 1$ during estimation and, for each replication, after estimation we scale b and Σ_β such that the lowest scale parameter is equal to one. For the HM-MNL, in estimation we use three phases with known scales $\tilde{\sigma} = (\infty, 1, \infty)$. We restrict the transition probabilities such that individuals can only stay in the current phase or move one phase up and let the transition probabilities be equal over time.

In the second part of the study, we check the robustness of our model to misspecification of the Markov process. We consider three extra DGPs (DGPs four to six) in which there are more phases in the DGP's Markov process to capture more complex forms of heterogeneous heteroscedasticity. The details of the DGPs and the results

⁷The heteroscedastic MNL (Bradley & Daly, 1994) incorporates time-dependent scale parameters $\{\sigma_t\}_{t=1}^T$ that are freely estimated except for one. We set the first scale parameter equal to one: $\sigma_1 = 1$.

are in Appendix 3.C.

3.4.1 Results

The results of the first part of the Monte Carlo study are given in Table 3.1. We report the mean across replications of the parameter estimates and the corresponding root mean squared error (RMSE) in parentheses. We do this for the three models estimated: the standard (mixed) MNL, the heteroscedastic MNL and the hidden Markov MNL.

In the first DGP, the HM-MNL is the true model with Σ_β diagonal. For this DGP, the standard MNL underestimates the mean of the preference parameters b . This is as expected, as heteroscedasticity leads to more random-looking choice making of individuals. The H-MNL also slightly underestimates the preference parameters. Hence, the bias towards zero, due to 14.4% of individuals choosing randomly at the minimum variance task 5, seems stronger than the bias away from zero, due to scaling back to the minimum variance task. The parameter estimator for the HM-MNL seems to have negligible small sample bias. Hence, the model seems well able to capture and distinguish between the individual-specific preferences and heteroscedasticity.

Interestingly, the MNL and H-MNL spuriously find a positive correlation between b_1 and b_2 ($\rho_{\beta,12} > 0$) and negative correlations $\rho_{\beta,13}$ and $\rho_{\beta,23}$. This implies that individuals with an extremal value for b_1 also tend to have extremal values for b_2 and b_3 , and vice versa. These correlations thus try to capture part of the individual-specific time-variation in the scale parameter via (spurious) individual-specific time-invariant correlations.

In the second DGP, the standard MNL is the true model. The estimator for the standard MNL seems to be unbiased. In contrast, the H-MNL clearly overestimates the preference parameters b . This illustrates that the parameter estimator for the H-MNL is biased away from zero due to estimation uncertainty in $\{\sigma_t\}_{t=1}^T$ of which the lowest is used to scale the preference parameters. The estimator for the HM-MNL seems almost unbiased: there is a slight bias away from zero. This is because the model assigns, on average, a small 2.2% fraction of individuals to start in phases one and three. The mean of the estimated probability of staying in the minimum variance phase (q_{22}) is close to the true 1. The RMSEs indicate that the loss in efficiency in estimating the HM-MNL instead of the correctly specified MNL is almost negligible.

Finally, in DGP 3, the HM-MNL is the true model and there is correlation across the

Table 3.1: Mean and RMSE (in parentheses) of the parameter estimates for the Monte Carlo study. Based on 1,000 Monte Carlo replications per DGP.

Parameter	DGP 1: HM-MNL ($\rho = 0$)			DGP 2: MNL			DGP 3: HM-MNL ($\rho \neq 0$)		
	True	MNL	HM-MNL	True	MNL	HM-MNL	True	MNL	HM-MNL
b_1	1.00	0.76 (0.24)	0.94 (0.09)	1.00	1.00	1.02	1.00	0.76 (0.24)	0.94 (0.08)
b_2	0.30	0.23 (0.07)	0.28 (0.03)	0.30	0.30	0.30	0.30	0.22 (0.08)	0.27 (0.03)
b_3	-0.50	-0.38 (0.12)	-0.48 (0.06)	-0.50	-0.50	-0.51	-0.50	-0.37 (0.13)	-0.46 (0.06)
$\sigma_{\beta,1}$	0.50	0.52 (0.03)	0.65 (0.15)	0.50	0.50	0.49	0.50	0.52 (0.03)	0.65 (0.15)
$\sigma_{\beta,2}$	0.40	0.33 (0.07)	0.41 (0.04)	0.40	0.40	0.40	0.40	0.33 (0.07)	0.41 (0.04)
$\sigma_{\beta,3}$	0.70	0.59 (0.12)	0.72 (0.07)	0.70	0.70	0.70	0.70	0.58 (0.13)	0.71 (0.07)
$\rho_{\beta,12}$	0.00	0.21 (0.22)	0.21 (0.22)	0.00	0.00	-0.03	0.00	0.69 (0.07)	0.69 (0.07)
$\rho_{\beta,13}$	0.00	-0.20 (0.21)	-0.20 (0.22)	0.00	0.00	0.03	0.00	-0.68 (0.08)	-0.68 (0.08)
$\rho_{\beta,23}$	0.00	-0.10 (0.14)	-0.10 (0.15)	0.00	-0.01	0.00	0.00	-0.72 (0.10)	-0.72 (0.10)
π_1	0.200		0.204 (0.047)	0.000		0.014 (0.027)	0.200		0.207 (0.046)
π_2	0.700		0.699 (0.055)	1.000		0.978 (0.035)	0.700		0.696 (0.054)
π_3	0.100		0.097 (0.042)	0.000		0.008 (0.015)	0.100		0.097 (0.042)
q_{11}	0.500		0.477 (0.170)	-		0.003	0.500		0.483 (0.163)
q_{22}	0.990		0.990 (0.003)	1.000		0.999 (0.001)	0.990		0.990 (0.004)

preference parameters to allow for time-invariant scale heterogeneity. The estimator for the HM-MNL seems to be unbiased for both the mean of the preference parameters (b) and the covariance matrix (Σ_β), indicating that this model can distinguish between time-invariant scale heterogeneity and time-varying scale heterogeneity.

To summarize, the parameter estimator for the preference parameters in the H-MNL specification seems biased for all three DGPs. Depending on the DGP and whether individual differences exist, the bias can be towards zero or away from zero. Also, the estimator for the standard MNL specification is biased towards zero in case heteroscedasticity is present. The HM-MNL alleviates these biases.

The results of the second part of the study, the performance of the HM-MNL under misspecification of the Markov process, are in Appendix 3.C. In the three DGPs considered, the true Markov process contains more phases than the three used in estimation, and in many of the phases the scale parameter is between one and infinity. Such a Markov process could be more realistic than the process assumed in estimation. We find that the estimators for the three models all underestimate the preference parameters. The bias towards zero is largest for the standard MNL, followed by the H-MNL. The estimator for the HM-MNL is most accurate in estimating the preference parameters. This indicates that our proposed HM-MNL works comparatively well when the true underlying Markov process is more complex than assumed in the model.

3.5 Case study I: learning and fatigue during discrete choice experiments

In this section, we illustrate our hidden Markov multinomial logit model with data obtained from a discrete choice experiment (DCE). During DCEs, respondents are repeatedly asked to make a hypothetical choice among a set of alternatives, where each alternative is described by a number of attributes (Green, 1974, Louviere and Woodworth, 1983). These experiments are used to elicit the preferences of respondents. The results can be used in product design and in predicting product demand (Rao, 2014). During DCEs, respondents might still need to learn about their preferences or the choice task at hand (Plott, 1993, Braga and Starmer, 2005), or may become tired, bored, or irritated while completing the choice tasks (Lavrakas, 2008). This latter process is known as *fatigue*. Due to learning and fatigue, a respondent may respond more randomly at some tasks. This randomness will lead to unpopular products to

be more often selected and, if unaccounted for in the model, overestimation of their potential demand.

The papers that have examined the presence of learning and fatigue during discrete choice experiments have so far only used population-level approaches for the learning and fatigue process.⁸ Using different datasets, they find mixed results: some find evidence of learning (DeSarbo et al., 2004, T.P. Holmes and Boyle, 2005, Czajkowski et al., 2014), some of fatigue (Bradley and Daly, 1994, Koppelman and Sethi, 2005, Savage and Waldman, 2008) and some of neither (Savage and Waldman, 2008, Hess et al., 2012). Because of the population-level approaches, these papers only provide insight into the aggregate scale per choice task, and thus cannot distinguish between different respondents at the same choice task: those that answer accurately, those that need to learn, and those that are fatigued. Hence, findings based on an individual-level model may totally differ.

To examine learning and fatigue during DCEs, we use data obtained from a discrete choice experiment on food choices conducted in the Netherlands (Koç & van Kippersluis, 2017).⁹ During the experiment, the respondents had to complete 18 choice tasks. At each choice task, a respondent was asked to choose between two meals: “Which of the two meals would you eat regularly (at least twice a week)?”.

The meals were described by the attributes price, taste, cooking time, and health consequences. Each attribute could take on three levels, with a clear ordering between the levels. For example, the price of the meal was either 2 Euro, 6 Euro, or 10 Euro. The respondents were divided into three groups. The groups differed in the attributes and information they obtained during the DCE about the health consequences of the meal. For the first respondent group, the health consequences of the meal were described by one explicit attribute: a meal could either be healthy, health neutral, or

⁸The only exception is the individual-level model of Campbell et al. (2015), which is a rather restrictive model. Campbell et al. (2015) *a priori* divide the choice tasks into early (E), middle (M) and late (L) tasks. To model the choices for the three different types of tasks, they specify a latent class model with seven classes. There are three different vectors of preference parameters β_E , β_M , β_L and three scale parameters σ_E , σ_M and σ_L . The first class of the latent class model has constant preferences β_M and constant scale σ_M for early, middle and late tasks (hence, no learning and fatigue). Classes 2 to 4 have a constant σ_M but different combination of β 's: class 2 has β_E for early tasks and β_M for the remaining tasks (only learning), class 3 has β_L for late tasks and β_M for the remaining tasks (only fatigue), and class 4 has β_E for early tasks, β_M for middle tasks and β_L for late tasks (learning and fatigue). Classes 5 to 7 have a constant β_M and a similar combination of σ as classes 2 to 4 have for β . This model does not allow for unobserved preference heterogeneity and the timing of learning and fatigue is fixed across respondents by *a priori* dividing the tasks into three sets.

⁹The dataset was obtained from the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands).

unhealthy. For the second and third group, the health consequences were described by three implicit health attributes: number of calories, grams of saturated fat, and grams of sodium. Furthermore, group 2 obtained health information describing what levels of these attributes constitute a healthy meal. Group 3 did not obtain this health information. For an overview of the attributes and corresponding levels per respondent group, see Table 3.2. The ordering of the tasks within each respondent group were randomized over the respondents and there was no overlap of respondents across groups.

Table 3.2: Attributes and attribute levels for the three respondent groups. The final column indicates which respondents groups (1,2 or 3) obtained which attributes in the choice experiment.

Attribute	Attribute levels			Groups
Price	2 Euro	6 Euro	10 Euro	1, 2, 3
Cooking time	10 min	30 min	50 min	1, 2, 3
Taste	OK	Good	Very good	1, 2, 3
Health consequence	Unhealthy	Health neutral	Healthy	1
Number of kilocalories	800	1,100	1,400	2, 3
Grams of saturated fat	10	20	30	2, 3
Milligrams of sodium	900	1,200	1,500	2, 3

We retain all respondents who filled in at least two choice tasks, also when a respondent dropped out. The three respondent groups contain the responses of 1,206, 1,154 and 1,185 respondents, respectively. In the model, we include the attribute levels as different dummy variables. For each attribute, we take the baseline level to be the first attribute level.

We consider three models: (1) a MNL, (2) a heteroscedastic MNL (H-MNL) (Bradley & Daly, 1994), and (3) our HM-MNL. For all three models, we take a multivariate normal distribution for β_i as given by

$$\beta_i \sim MVN(b, \Sigma_\beta),$$

where Σ_β is a full positive definite covariance matrix. In estimation, we use 250 scrambled Halton draws per respondent and 30 starting values per model.¹⁰

For the HM-MNL, we consider two specifications. Both specifications have three phases $\tilde{\sigma} = (\infty, 1, \infty)$: respondents in the first phase still need to learn and answer randomly, respondents in in the second phase (the *minimum variance phase*) answer

¹⁰The starting values for b and Σ_β for the H-MNL and HM-MNL are set equal to the maximum likelihood estimates of the MNL.

most accurately, and respondents in the third phase answer randomly due to fatigue. We restrict the transition probabilities such that a respondent can either stay in the current phase or move one phase up. For the first specification, we let the transition probabilities be equal over tasks: $q_{11t} = q_{11}$ and $q_{22t} = q_{22}$ for all t . For the second specification, we allow for the transition probabilities from the minimum variance phase to the fatigue phase to be different over tasks.

For the H-MNL, we fix $\sigma_1 = 1$ during estimation. After estimation, we scale b , Σ_β and $\{\sigma_t\}_{t=1}^T$ such that the minimum variance task has variance one, that is, $\min\{\sigma_t\}_{t=1}^T = 1$.

3.5.1 Results

The results for the first respondent group are shown in Table 3.3.¹¹ For this group, the meals were described by four attributes, explicit health information was given in the final attribute ‘health consequences’. With the MNL, we find that individuals, on average, positively value a low price and cooking time, a good taste, and a healthy meal.

The H-MNL and HM-MNL find similar patterns as the MNL in the preference parameters, although they are estimated further away from zero. Hence, learning and fatigue seem to both be present. For the population-level H-MNL, this is clearly indicated in the time-variation in the scale parameters. The variance increases at task 2, then decreases, and in the final couple of tasks increases again. The minimum variance task is estimated to be 14. Hence, in the first 14 tasks learning seems more prevalent than fatigue, whereafter fatigue seems more prevalent. The standard errors do imply that there is quite some estimation uncertainty and the minimum variance task could be anywhere from task 7 to 16.

The first HM-MNL, with transition probabilities restricted over tasks, also finds evidence of learning and fatigue. An estimated 17.4% of respondents start in the learning phase in which they reside on average five tasks.¹² Fatigue also occurs: 1.2% of respondents are estimated to answer randomly throughout the survey, and at each task an estimated 0.3% of respondents in the minimum variance phase gets fatigued. Based on the estimated initial and transition probabilities, 5.2% of respondents is fatigued at the final choice task.

¹¹The detailed results for the covariance matrix Σ_β are available upon request.

¹²The average number of tasks that someone who starts in the learning phase will remain in the learning phase is equal to $1/(1 - q_{11})$.

Table 3.3: Parameter estimates and standard errors (in parentheses) for group 1 from the food choice dataset. Baseline levels are price 2 euro, time 10 minutes, taste OK, and health unhealthy.

	MNL		H-MNL		HM-MNL ^a		HM-MNL ^b	
Price 6 euro	-0.74	(0.04)	-0.96	(0.07)	-0.98	(0.06)	-0.99	(0.06)
Price 10 euro	-2.15	(0.08)	-2.75	(0.18)	-2.76	(0.12)	-2.77	(0.13)
Time 30 min	-0.35	(0.04)	-0.44	(0.04)	-0.46	(0.05)	-0.46	(0.05)
Time 50 min	-1.23	(0.06)	-1.57	(0.11)	-1.59	(0.08)	-1.60	(0.08)
Taste good	0.66	(0.04)	0.83	(0.06)	0.85	(0.05)	0.85	(0.05)
Taste very good	1.18	(0.05)	1.51	(0.10)	1.50	(0.07)	1.51	(0.07)
Health neutral	3.50	(0.09)	4.49	(0.28)	4.55	(0.16)	4.58	(0.17)
Healthy	4.96	(0.13)	6.36	(0.39)	6.43	(0.23)	6.48	(0.24)
π_1					0.174	(0.024)	0.170	(0.024)
π_2					0.814	(0.033)	0.830	(0.039)
π_3					0.012	(0.015)	0.000	(0.025)
q_{11}					0.816	(0.043)	0.796	(0.046)
q_{22}					0.997	(0.001)		
σ_1			1.59	(0.13)				
σ_2 or $q_{22,1}$			1.76	(0.14)			0.990	(0.032)
σ_3 or $q_{22,2}$			1.59	(0.13)			0.968	(0.023)
σ_4 or $q_{22,3}$			1.64	(0.15)			1.000	(0.027)
σ_5 or $q_{22,4}$			1.33	(0.11)			1.000	(0.026)
σ_6 or $q_{22,5}$			1.51	(0.14)			1.000	(0.019)
σ_7 or $q_{22,6}$			1.16	(0.11)			1.000	(0.018)
σ_8 or $q_{22,7}$			1.21	(0.12)			1.000	(0.024)
σ_9 or $q_{22,8}$			1.14	(0.11)			1.000	(0.027)
σ_{10} or $q_{22,9}$			1.01	(0.10)			1.000	(0.018)
σ_{22} or $q_{22,10}$			1.21	(0.11)			0.990	(0.012)
σ_{12} or $q_{22,11}$			1.13	(0.11)			1.000	(0.014)
σ_{13} or $q_{22,12}$			1.22	(0.13)			1.000	(0.014)
σ_{14} or $q_{22,13}$			1.00	-			1.000	(0.019)
σ_{15} or $q_{22,14}$			1.10	(0.11)			1.000	(0.022)
σ_{16} or $q_{22,15}$			1.20	(0.11)			0.985	(0.015)
σ_{17} or $q_{22,16}$			1.25	(0.12)			1.000	(0.021)
σ_{18} or $q_{22,17}$			1.24	(0.11)			0.987	(0.021)
average σ_β	1.5		1.9		1.9		1.9	
# free parameters	44		61		48		64	
log-likelihood	-10,560		-10,524		-10,483		-10,477	
BIC	21,559		21,658		21,446		21,593	
AIC3	21,252		21,232		21,111		21,146	
AIC2	21,208		21,171		21,063		21,082	

^a HM-MNL with 3 phases $\tilde{\sigma} = (\infty, 1, \infty)$: equal transition probabilities over tasks.

^b HM-MNL with 3 phases $\tilde{\sigma} = (\infty, 1, \infty)$: transition probability to fatigue different per task.

The less restrictive second HM-MNL finds that an estimated 17.0% of respondents start in the learning phase and none of the respondents start in the fatigue phase. Furthermore, mainly after the first and second task, respondents seem to become tired. For tasks 3 up to 15, most respondents in the minimum variance phase seem to stay there, and at the final three tasks 16 to 18, more respondents seem to become fatigued. At the final choice task, an estimated 7.2% of respondents is fatigued.

According to the information criteria, both HM-MNL models are preferred over the standard MNL and the H-MNL. The first HM-MNL model seems the most preferred. Even though this HM-MNL has 13 parameters less to estimate than the H-MNL, the likelihood value indicates that it better fits the data. Thus, there seems to be quite some heterogeneity in learning and fatigue across respondents.

The results for the HM-MNL models indicate that at each task, a number of respondents answer randomly. This causes a bias towards zero in the preference parameters b in the H-MNL next to the bias away from zero due to scaling back to the minimum variance task. For this food choice dataset for the first respondent group, the biases seem to almost cancel each other with the bias towards zero seeming just a bit more dominant: the preference parameters b estimated by the H-MNL are slightly closer to zero than those estimated by the HM-MNL.

The results for the second and third respondent groups are in Tables 3.4 and 3.5, respectively. For these two groups, health consequences were described by three attributes: (i) number of calories, (ii) amount of saturated fat, and (iii) amount of sodium. For group two, health information on the attributes was provided in the text, for group three no health information was provided.

The HM-MNL models find that as the amount of information decreases from group one to three, the percentage of respondents that start in the learning phase increases from 17% to 22% to 27%. Moreover, according to the first HM-MNL specification, the probability to become fatigued once in the minimum variance phase increases from group one to three, from 0.3% to 0.4% to 0.5% per choice task, although there is some uncertainty in these estimates. Hence, summarizing health information in one attribute or providing health information in the text seems to reduce the need for learning and the risk of fatigue.

With the H-MNL we also find evidence of learning for respondent groups two and three. Remarkably, for all three groups, we find an initial increase in the variance from task 1 to 2, after which the variance decreases. This is not due to the order of

Table 3.4: Parameter estimates and standard errors (in parentheses) for group 2 from the food choice dataset. Baseline levels are price 2 euro, time 10 minutes, taste OK, calories 800, saturated fat 10 gram, and sodium 900 mg.

	MNL		H-MNL		HM-MNL ^a		HM-MNL ^b	
Price 6 euro	-0.61	(0.04)	-0.86	(0.06)	-0.91	(0.06)	-0.92	(0.06)
Price 10 euro	-1.54	(0.08)	-2.16	(0.13)	-2.19	(0.13)	-2.21	(0.13)
Time 30 min	-0.08	(0.04)	-0.13	(0.03)	-0.15	(0.05)	-0.16	(0.05)
Time 50 min	-0.65	(0.06)	-0.93	(0.07)	-0.92	(0.09)	-0.93	(0.09)
Taste good	0.69	(0.04)	0.99	(0.06)	0.98	(0.06)	0.98	(0.06)
Taste very good	1.19	(0.06)	1.69	(0.10)	1.64	(0.09)	1.65	(0.09)
Calories 1100	-0.81	(0.04)	-1.10	(0.07)	-1.11	(0.07)	-1.11	(0.07)
Calories 1400	-1.59	(0.06)	-2.20	(0.13)	-2.19	(0.11)	-2.20	(0.11)
Sat fat 20 gram	-0.66	(0.04)	-0.94	(0.06)	-0.94	(0.07)	-0.95	(0.07)
Sat fat 30 gram	-1.28	(0.06)	-1.78	(0.11)	-1.79	(0.09)	-1.81	(0.10)
Sodium 1200mg	-0.37	(0.04)	-0.52	(0.05)	-0.55	(0.06)	-0.55	(0.06)
Sodium 1500mg	-0.87	(0.05)	-1.21	(0.08)	-1.22	(0.08)	-1.23	(0.08)
π_1					0.218	(0.030)	0.217	(0.030)
π_2					0.765	(0.036)	0.782	(0.042)
π_3					0.017	(0.021)	0.002	(0.030)
q_{11}					0.816	(0.042)	0.802	(0.044)
q_{22}					0.996	(0.001)		
σ_1			1.80	(0.14)				
σ_2 or $q_{22,1}$			1.96	(0.16)			0.973	(0.040)
σ_3 or $q_{22,2}$			1.85	(0.15)			0.983	(0.035)
σ_4 or $q_{22,3}$			1.60	(0.15)			0.999	(0.037)
σ_5 or $q_{22,4}$			1.21	(0.10)			0.999	(0.035)
σ_6 or $q_{22,5}$			1.57	(0.13)			0.998	(0.034)
σ_7 or $q_{22,6}$			1.48	(0.12)			0.982	(0.024)
σ_8 or $q_{22,7}$			1.30	(0.11)			0.999	(0.028)
σ_9 or $q_{22,8}$			1.35	(0.11)			0.999	(0.026)
σ_{10} or $q_{22,9}$			1.16	(0.10)			0.997	(0.025)
σ_{22} or $q_{22,10}$			1.51	(0.13)			0.996	(0.026)
σ_{12} or $q_{22,11}$			1.32	(0.12)			0.999	(0.021)
σ_{13} or $q_{22,12}$			1.19	(0.11)			0.999	(0.026)
σ_{14} or $q_{22,13}$			1.00	-			0.995	(0.023)
σ_{15} or $q_{22,14}$			1.40	(0.12)			0.983	(0.022)
σ_{16} or $q_{22,15}$			1.10	(0.11)			0.998	(0.025)
σ_{17} or $q_{22,16}$			1.22	(0.11)			0.999	(0.025)
σ_{18} or $q_{22,17}$			1.18	(0.12)			0.999	(0.039)
average σ_β	1.1		1.5		1.4		1.4	
# free parameters	90		107		94		110	
log-likelihood	-10,989		-10,946		-10,900		-10,897	
BIC	22,872		22,955		22,734		22,887	
AIC3	22,248		22,212		22,082		22,124	
AIC2	22,158		22,105		21,988		22,014	

^a HM-MNL with 3 phases $\tilde{\sigma} = (\infty, 1, \infty)$: equal transition probabilities over tasks.

^b HM-MNL with 3 phases $\tilde{\sigma} = (\infty, 1, \infty)$: transition probability to fatigue different per task.

Table 3.5: Parameter estimates and standard errors (in parentheses) for group 3 from the food choice dataset. Baseline levels are price 2 euro, time 10 minutes, taste OK, calories 800, saturated fat 10 gram, and sodium 900 mg.

	MNL		H-MNL		HM-MNL ^a		HM-MNL ^b	
Price 6 euro	-0.71	(0.04)	-1.03	(0.07)	-1.03	(0.07)	-1.06	(0.07)
Price 10 euro	-1.82	(0.09)	-2.64	(0.16)	-2.68	(0.14)	-2.61	(0.15)
Time 30 min	-0.09	(0.04)	-0.14	(0.03)	-0.12	(0.05)	-0.15	(0.06)
Time 50 min	-0.73	(0.07)	-1.08	(0.08)	-1.09	(0.09)	-1.04	(0.10)
Taste good	0.89	(0.05)	1.30	(0.08)	1.39	(0.08)	1.40	(0.08)
Taste very good	1.40	(0.07)	2.08	(0.12)	2.19	(0.12)	2.26	(0.12)
Calories 1100	-0.58	(0.04)	-0.85	(0.06)	-0.81	(0.06)	-0.92	(0.07)
Calories 1400	-1.29	(0.06)	-1.89	(0.11)	-1.86	(0.10)	-2.04	(0.11)
Sat fat 20 gram	-0.43	(0.04)	-0.61	(0.04)	-0.60	(0.06)	-0.68	(0.06)
Sat fat 30 gram	-0.91	(0.05)	-1.31	(0.08)	-1.30	(0.08)	-1.44	(0.09)
Sodium 1200mg	-0.33	(0.04)	-0.48	(0.04)	-0.52	(0.06)	-0.55	(0.07)
Sodium 1500mg	-0.72	(0.05)	-1.04	(0.07)	-1.07	(0.07)	-1.11	(0.08)
π_1					0.273	(0.033)	0.269	(0.032)
π_2					0.713	(0.035)	0.706	(0.047)
π_3					0.014	(0.027)	0.025	(0.045)
q_{11}					0.838	(0.036)	0.831	(0.037)
q_{22}					0.995	(0.002)		
σ_1			1.90	(0.14)				
σ_2 or $q_{22,1}$			2.13	(0.17)			0.993	(0.064)
σ_3 or $q_{22,2}$			1.85	(0.14)			0.990	(0.052)
σ_4 or $q_{22,3}$			1.88	(0.16)			0.998	(0.046)
σ_5 or $q_{22,4}$			1.40	(0.12)			0.992	(0.030)
σ_6 or $q_{22,5}$			1.17	(0.10)			0.995	(0.033)
σ_7 or $q_{22,6}$			1.70	(0.14)			0.996	(0.029)
σ_8 or $q_{22,7}$			1.63	(0.13)			0.997	(0.029)
σ_9 or $q_{22,8}$			1.44	(0.13)			0.999	(0.028)
σ_{10} or $q_{22,9}$			1.29	(0.12)			0.987	(0.026)
σ_{22} or $q_{22,10}$			1.30	(0.12)			0.990	(0.028)
σ_{12} or $q_{22,11}$			1.17	(0.10)			0.990	(0.022)
σ_{13} or $q_{22,12}$			1.25	(0.11)			1.000	(0.022)
σ_{14} or $q_{22,13}$			1.26	(0.11)			0.999	(0.031)
σ_{15} or $q_{22,14}$			1.13	(0.09)			0.982	(0.022)
σ_{16} or $q_{22,15}$			1.50	(0.12)			0.993	(0.020)
σ_{17} or $q_{22,16}$			1.33	(0.12)			0.999	(0.018)
σ_{18} or $q_{22,17}$			1.00	-			1.000	(0.033)
average σ_β	1.1		1.6		1.5		1.6	
# free parameters	90		107		94		110	
log-likelihood	-11,474		-11,433		-11,360		-11,354	
BIC	23,844		23,933		23,656		23,804	
AIC3	23,217		23,188		23,001		23,037	
AIC2	23,127		23,081		22,907		22,927	

^a HM-MNL with 3 phases $\tilde{\sigma} = (\infty, 1, \infty)$: equal transition probabilities over tasks.

^b HM-MNL with 3 phases $\tilde{\sigma} = (\infty, 1, \infty)$: transition probability to fatigue different per task.

the choice tasks as they are randomized over respondents. Hence, it seems that there is a relatively large group of respondents who try to answer accurately at task 1, but from task 2 onwards do not so anymore. This behavior can be seen more explicitly by the second HM-MNL specification, where we find that there is relatively large group of respondents that move to the fatigue phase after tasks one and two. Combined with the reduction of respondents that need to learn in these two tasks, the variation in the scale parameters in the H-MNL can be explained.

In summary, the hidden Markov MNLs provide the best fit of the data for all three respondent groups. These models also provide interesting and plausible insights into the presence of learning and fatigue during the discrete choice experiment. Furthermore, the hidden Markov MNLs allow for an analysis per individual: given the individual's choices, what is the probability that she was affected by learning or fatigue during the experiment? For this purpose, one can compute the conditional probabilities that an individual is in a certain phase at a given choice task using the formulas in Appendix 3.B. Such an individual-level analysis is not possible with the heteroscedastic MNL.

3.6 Case study II: differential capabilities in ranking

In this section, we illustrate our hidden Markov rank-ordered logit model with rankings obtained from a survey on cultural opinions conducted in the Netherlands (Sociaal en Cultureel Planbureau, 2004, Fok et al., 2012). One of the questions in the survey asked the respondents to rank 16 political goals from most to least desired. In total, 2,261 individuals aged sixteen years and older completed the ranking. The initial presented ordering of the political goals in the survey was randomized over respondents.

We estimate and compare three models: (1) a ROL, (2) a heteroscedastic ROL (H-ROL) (Hausman & Ruud, 1987) and (3) our HM-ROL. For all models, we take a multivariate normal distribution for β_i as given by

$$\beta_i \sim MVN(b, \Sigma_\beta),$$

with full positive definite covariance matrix Σ_β .¹³ In estimation, we use 30 starting

¹³For the political preferences ranking data, we have only one observation per individual and quite

values per model and 250 scrambled Halton draws per respondent. For the H-ROL, we fix $\sigma_1 = 1$.

For the HM-ROL, we consider three specifications. The first specification is equivalent to the latent class ROL in Fok et al. (2012), with the addition of allowing for individual-specific preference parameters. That is, we use two phases $\tilde{\sigma} = (1, \infty)$ and restrict the transition probabilities such that, between consecutive ranks, a respondent can only move from phase 1 to phase 2 and once in phase 2 stays there. In other words, a respondent can assign all ranks accurately (all choices based on phase 1), all ranks randomly (all choices based on phase 2), or the top j ranks accurately and the bottom $J - j$ randomly (in phase 1 for rank one until j , in phase 2 for ranks $j + 1$ and higher) for any j .

For the second specification, we also allow for bottom ranks to be assigned accurately. We use three phases $\tilde{\sigma} = (1, \infty, 1)$ and restrict the transition probabilities such that, between consecutive ranks, a respondent can only move from phase 1 to phase 2 or from phase 2 to phase 3, and once in phase 3 stays there. Also, a respondent is restricted to start (assign the top rank) in either phase 1 or 2. In the third specification, we include a middle phase with scale to be estimated: $\tilde{\sigma} = (1, \tilde{\sigma}_2, \infty, 1)$ to allow for a decrease in the ability of respondents to assign lower ranks. We restrict the transition and initial phase probabilities such that a respondent can only move one phase up and can only start in the first and third phase. For further parsimony, we restrict the transition probabilities to be equal to each other over ranks h , except for moving from the first to the second phase in the first two tasks (assign top ranks accurately) and from the third to the final phase in the last two tasks (assign bottom ranks accurately).

3.6.1 Results

The results for the political preferences ranking data are given in Table 3.6. With the standard ROL we find that individuals seem to attach most value to goals as ‘maintain order’, ‘stable economy’, ‘fight crime’, ‘freedom of speech’, and ‘social security’. The estimated correlations across the individual-specific preference parameters in Σ_β indicate which goals are often ranked close by (results not shown).¹⁴ We find that this holds most strongly for (i) ‘maintain order’ and ‘fight crime’, (ii) ‘more say politics’ and ‘more say community’, (iii) ‘economic growth’ and ‘stable economy’,

a number of free parameters in the (15×15) matrix Σ_β . Therefore, using a low-rank approximation of Σ_β or a Bayesian approach with informative priors might be useful to reduce the risk of overfitting.

¹⁴The detailed results for Σ_β are available upon request.

(iv) ‘defence forces’ and ‘cities and countryside’, and (v) ‘humane society’ and ‘ideas > money’.

We next consider the H-ROL which accounts for the behavior that individuals cannot rank all alternatives accurately in a homogeneous way across individuals. The estimates for the mean preference parameters b are further away from zero than for the ROL, and the scale parameters σ_h show a gradual increase as the rank h increases. Hence, individuals seem to be unable to assign all ranks accurately, rendering the estimator for the standard ROL biased. Moreover, individuals seem to most accurately assign the top ranks, followed by the middle and then the bottom ranks. The preference ordering of the political goals stays roughly the same.

The HM-ROL allows for individual differences in the ranking capabilities. The first HM-ROL specification is equivalent to the latent class ROL in Fok et al. (2012), with the addition of allowing for preference heterogeneity. We find that all individuals are able to rank the first alternative accurately ($\pi_1 = 1.0$). The probabilities of staying in the minimum variance phase 1 for consecutive ranks (q_{11h}) are mostly smaller than one, indicating that quite some individuals find it rather difficult to assign the middle and bottom ranks. Moreover, there seem to be individual differences in the number of top ranks that can be assigned accurately. The probabilities of staying in the minimum variance phase one are especially low for the final couple of ranks. This indicates that a large proportion of respondent who are able to accurately assign ranks 1 to 9, have more trouble assigning the lower ranks. These findings mostly agree with the findings Fok et al. (2012), with the exception that they find that 4% of individuals cannot rank the first alternative accurately. This difference suggests that it is important to allow for preference heterogeneity when allowing for differential capabilities in ranking.

The second HM-ROL specification also allows for bottom ranks to be assigned accurately. The probability of staying in the first phase (q_{11h}) are close to one for the first six ranks, and quite a bit lower for the subsequent ranks. This indicates that quite some individuals can accurately assign ranks one to six, but find it more difficult to assign middle ranks from rank seven onwards. The probability of staying in the second phase (q_{22h}) differ quite a bit over ranks. These probabilities are often quite low, indicating that indeed some individuals are able to assign the bottom ranks accurately. Because of the uncertainty in these estimates, it might be sensible to add restrictions to the transition probabilities. For example, one can impose them equal across certain (middle) ranks.

Table 3.6: Parameter estimates and standard errors (in parentheses) for the political preferences ranking dataset. Baseline level is ‘take good care of immigrants’.

	ROL	H-ROL	HM-ROL ^a	HM-ROL ^b	HM-ROL ^c
maintain order	2.45 (0.05)	4.04 (0.10)	2.82 (0.07)	3.03 (0.08)	3.98 (0.15)
more say politics	1.20 (0.05)	2.32 (0.06)	1.51 (0.06)	1.67 (0.07)	2.29 (0.12)
fight rising prices	1.50 (0.05)	2.82 (0.08)	1.84 (0.07)	2.02 (0.08)	2.78 (0.13)
freedom of speech	2.31 (0.04)	3.84 (0.10)	2.64 (0.06)	2.86 (0.07)	3.77 (0.15)
economic growth	1.44 (0.05)	2.66 (0.07)	1.74 (0.07)	1.92 (0.08)	2.62 (0.13)
defence forces	-0.77 (0.05)	-2.22 (0.07)	-1.39 (0.11)	-1.74 (0.17)	-1.86 (0.15)
more say community	1.14 (0.05)	2.21 (0.06)	1.43 (0.06)	1.60 (0.07)	2.18 (0.12)
cities and countryside	0.17 (0.04)	0.44 (0.03)	0.29 (0.06)	0.36 (0.07)	0.46 (0.09)
stable economy	2.40 (0.05)	3.97 (0.11)	2.75 (0.06)	2.96 (0.08)	3.92 (0.15)
fight crime	2.40 (0.05)	3.99 (0.11)	2.77 (0.07)	2.99 (0.08)	3.92 (0.15)
humane society	1.67 (0.05)	3.04 (0.08)	1.97 (0.06)	2.17 (0.07)	2.99 (0.13)
ideas > money	0.91 (0.04)	1.82 (0.06)	1.15 (0.06)	1.31 (0.07)	1.81 (0.11)
fight unemployment	2.10 (0.05)	3.64 (0.10)	2.47 (0.06)	2.68 (0.08)	3.59 (0.15)
fight pollution	1.05 (0.04)	2.13 (0.06)	1.33 (0.05)	1.50 (0.07)	2.07 (0.11)
social security	2.31 (0.05)	3.85 (0.10)	2.65 (0.06)	2.88 (0.08)	3.79 (0.15)
δ_2					2.29 (0.17)
π_1			1.00 (0.02)	1.00 (0.02)	0.98 (0.02)
π_2			0.00 (0.01)	0.00 (0.01)	
π_3					0.02 (0.01)
π_4					
σ_1		1.00			
σ_2 or $q_{11,1}$		1.10 (0.04)	0.99 (0.01)	0.98 (0.01)	
σ_3 or $q_{11,2}$		1.23 (0.04)	1.00 (0.01)	0.99 (0.02)	
σ_4 or $q_{11,3}$		1.21 (0.04)	1.00 (0.01)	1.00 (0.01)	
σ_5 or $q_{11,4}$		1.36 (0.04)	0.99 (0.01)	0.98 (0.01)	
σ_6 or $q_{11,5}$		1.58 (0.05)	0.99 (0.01)	0.97 (0.02)	
σ_7 or $q_{11,6}$		1.63 (0.05)	0.98 (0.01)	0.96 (0.02)	
σ_8 or $q_{11,7}$		1.86 (0.06)	0.99 (0.02)	0.94 (0.03)	
σ_9 or $q_{11,8}$		1.90 (0.06)	0.98 (0.02)	0.92 (0.03)	
σ_{10} or $q_{11,9}$		1.99 (0.06)	0.94 (0.03)	0.89 (0.03)	
σ_{11} or $q_{11,10}$		2.44 (0.07)	0.96 (0.03)	0.91 (0.05)	
σ_{12} or $q_{11,11}$		2.78 (0.08)	0.94 (0.03)	0.83 (0.05)	
σ_{13} or $q_{11,12}$		2.59 (0.08)	0.97 (0.04)	0.94 (0.06)	
σ_{14} or $q_{11,13}$		3.27 (0.10)	0.87 (0.04)	0.78 (0.07)	

Table 3.6: (Continued)

	ROL	H-ROL	HM-ROL ^a	HM-ROL ^b	HM-ROL ^c
σ_{15} or $q_{1,1,14}$		4.54 (0.15)	0.77 (0.05)	0.61 (0.11)	
$q_{22,1}$			0.96 (5.63)		
$q_{22,2}$			0.96 (0.68)		
$q_{22,3}$			0.68 (0.28)		
$q_{22,4}$			0.78 (0.41)		
$q_{22,5}$			0.98 (0.35)		
$q_{22,6}$			0.96 (0.26)		
$q_{22,7}$			0.74 (0.19)		
$q_{22,8}$			0.41 (0.15)		
$q_{22,9}$			0.61 (0.17)		
$q_{22,10}$			1.00 (0.18)		
$q_{22,11}$			0.83 (0.12)		
$q_{22,12}$			0.72 (0.10)		
$q_{22,13}$			1.00 (0.15)		
$q_{22,14}$			0.99 (0.15)		
$q_{11,1}$					0.90 (0.05)
$q_{11,2}$					1.00 (0.06)
$q_{11,3:14}$					0.83 (0.02)
q_{22}					0.95 (0.01)
$q_{33,1:12}$					0.87 (0.04)
$q_{33,13}$					1.00 (0.15)
$q_{33,14}$					1.00 (0.16)
average σ_β	1.3	2.5	1.6	1.7	2.3
# free parameters	135	149	150	164	144
log-likelihood	-62,369	-62,193	-62,207	-62,172	-62,152
BIC	125,781	125,537	125,573	125,610	125,417
AIC3	125,144	124,833	124,865	124,835	124,736
AIC2	125,009	124,684	124,715	124,671	124,592

^aHM-ROL with 2 phases $\bar{\sigma} = (1, \infty)$: accurately assign top ranks.

^bHM-ROL with 3 phases $\bar{\sigma} = (1, \infty, 1)$: accurately assign top and bottom ranks.

^cHM-ROL with 4 phases $\bar{\sigma} = (1, \bar{\sigma}_2, \infty, 1)$: accurately assign top and bottom ranks + decrease in accuracy.

In the third HM-ROL specification, we allow for a decrease in accuracy for assigning alternatives for consecutive ranks, as well as for bottom ranks to be assigned accurately. The estimated scale parameter for the second phase is equal to 2.3. We find that 10% of respondent in the minimum variance phase move to the second phase after rank one, 0% after rank two, and 17% of respondents after each remaining rank. Hence, the information content in the ranks assigned seems to highly differ across respondents.

The main difference between the three HM-ROL specifications is that the final specification allows for a decrease in accuracy in the alternatives assigned to consecutive ranks, whereas the first two specifications assume that an individual either completely accurately assigns a rank or completely randomly. According to the three information criteria, the third HM-ROL specification should be preferred. Hence, for this ranking dataset, it seems more likely that individuals do not completely randomly assign middle ranks, but that the choice is more random compared to top ranks.

When comparing all five models, the information criteria indicate that the third HM-ROL specification should be most preferred. Even though this model has five parameters less to estimate than the H-ROL, the likelihood value indicates that it much better fits the data. The standard ROL should be the least preferred model, followed by the first HM-ROL specification. The H-ROL and second HM-ROL specification are at a shared second place.

3.7 Conclusion

The heteroscedastic logit model is useful to describe repeated choices of individuals when randomness in the choice-making varies over time. For example, due to fatigue, individuals may respond more randomly to survey questions as the survey progresses. Or when asked to give a complete ranking amongst multiple alternatives, individuals may more accurately assign top ranks than middle and bottom ranks.

In this paper, we generalize the standard heteroscedastic logit model to allow for individual differences in the dynamics in this randomness. In case individual differences exist, this individual-level approach has three main advantages: (i) it alleviates biases in the preference parameters, (ii) makes more efficient use of data, and (iii) allows for an analysis of individual behavior. The generalization amounts to adding an individual- and time/rank-specific scale parameter to the multinomial and rank-ordered logit model. We let the dynamics in the sequence of an individual's scale

parameters be governed by a Markov process. Additionally, we allow for unobserved preference heterogeneity. For inference, we develop a simulated maximum likelihood estimation approach.

In a Monte Carlo study, we find that our proposed model works well and the proposed estimator seems unbiased in various settings. For the standard heteroscedastic logit model, the biases in the estimator for the preference parameters are clearly illustrated: the bias towards zero due to neglecting individual differences in the dynamics in the scale parameter, and the bias away from zero due to scaling the preference parameters based on the minimum of the estimated scale parameters. Depending on the data generating process, one of these biases may dominate the other. In case of heteroscedasticity, the estimator for the preference parameters of the standard MNL is clearly biased towards zero, because heteroscedasticity leads to more random-looking choice-making of respondents. Our proposed model and estimator eliminate these biases. Furthermore, when allowing for preference heterogeneity via a multivariate normal distribution, both the standard MNL and the heteroscedastic MNL tend to spuriously capture individual differences in the dynamics in the scale parameter in time-invariant correlations between preference parameters.

We also illustrate our model with two empirical applications: one using multinomial choice data from a discrete choice experiment on food choices to model learning and fatigue effects, and one on rank-ordered data from a survey to model differential capabilities in ranking. For the multinomial choices, we find that accounting for individual differences in learning and fatigue leads to a much better fit of the data, while needing less model parameters. The same holds for the rank-ordered data.

Our approach has one main limitation: each variable gets scaled with the same factor. Hence, the model cannot capture choice strategies where choices are made based on different subsets of attributes as time progresses, or where preferences change over time. The model could be extended to allow for a different scale parameter per variable, for example, by letting each scale parameter be governed by its own Markov process. However, for datasets with limited information per individual, such an approach would be susceptible to overfitting and estimation uncertainty can become problematic.

We provide three venues for future research. First, in case one wants to impose restrictions on the minimum number of tasks an individual should be in a phase, one can use a second- or higher-order Markov process. By using suitable restrictions on the transition probabilities, no extra parameters need to be estimated. Of course, if

desired, one can also allow the transition probabilities to depend on the duration in a phase using such a higher-order Markov process. Second, for rank-ordered data, the Markov process over time and over ranks can be combined, to simultaneously allow for learning and fatigue and for differential capabilities in ranking.

Third, for the applications, we recommend to use more flexible forms of preference heterogeneity than the used multivariate normal. This especially holds when one wants to include scale parameters between one and infinity. A mixture of multivariate normal distributions might be able to capture more realistically the differences in preferences across individuals. One way in which individuals may differ is that some individuals may answer more randomly throughout the observed period than others, also known as time-invariant scale heterogeneity. In the multivariate normal, such behavior is partly captured in the correlations in the covariance matrix. Using a more flexible form than one multivariate normal could further reduce the way in which the Markov process can capture part of the time-invariant scale heterogeneity in case one of the scale parameters is allowed to be between one and infinity.

Appendix

3.A Maximum simulated likelihood estimation

We use maximum simulated likelihood estimation to estimate the parameters of the hidden Markov multinomial and rank-ordered logit model. For this purpose, we maximize the (approximated) likelihood function directly with respect to θ , q , π , and $\tilde{\sigma}$ using a quasi Newton-Raphson algorithm. During optimization, we use analytic gradients of the simulated log-likelihood function and approximate the Hessian with the BFGS algorithm. In this appendix, we provide more details for the estimation approaches including the explicit likelihood functions for multinomial choices in Section 3.A.1 and for rank-ordered choices in Section 3.A.2.

3.A.1 Hidden Markov multinomial logit model

The likelihood function of the HM-MNL can be written as

$$\begin{aligned}
 p(y|\theta, q, \pi, \tilde{\sigma}) &= \prod_{i=1}^N \left[\int \left(\sum_{s_i^* \in \mathcal{S}} \Pr[S_i = s_i^* | q, \pi] p(y_i | \beta_i, \tilde{\sigma}, s_i^*) \right) f(\beta_i | \theta) d\beta_i \right] \\
 &= \prod_{i=1}^N \left[\int \left\{ \sum_{s_i^* \in \mathcal{S}} \left(\Pr[S_i = s_i^* | q, \pi] \prod_{t=1}^T \frac{\exp\left(\frac{1}{\tilde{\sigma}_{s_{it}}}(x'_{itj}\beta_i)\right)}{\sum_{l=1}^J \exp\left(\frac{1}{\tilde{\sigma}_{s_{it}}}(x'_{itl}\beta_i)\right)} \right) \right\} f(\beta_i | \theta) d\beta_i \right] \\
 &= \prod_{i=1}^N \left[\int \left\{ \pi' f_{i1} \left(\prod_{t=2}^{T-1} Q_{t-1} \tilde{f}_{it} \right) Q_{T-1} \tilde{f}_{iT} \right\} f(\beta_i | \theta) d\beta_i \right], \tag{3.12}
 \end{aligned}$$

where Q_t is a $(M \times M)$ transition probability matrix with element (m, n) equal to q_{mnt} , and \tilde{f}_{it} is a $(M \times 1)$ vector with the likelihood contribution of task t of respondent i given β_i and s_{it} , with element m equal to

$$\tilde{f}_{itm} \equiv \Pr[Y_{it} = y_{it} | \beta_i, \tilde{\sigma}, s_{it} = m] = \frac{\exp\left(\frac{1}{\tilde{\sigma}_m}(x'_{itj}\beta_i)\right)}{\sum_{l=1}^J \exp\left(\frac{1}{\tilde{\sigma}_m}(x'_{itl}\beta_i)\right)}.$$

Furthermore, f_{it} is a diagonal $(M \times M)$ matrix with the diagonal equal to \tilde{f}_{it} .

We approximate the likelihood function using Monte Carlo integration:

$$\begin{aligned}
 p(y|\theta, q, \pi, \tilde{\sigma}) &\approx \prod_{i=1}^N \left[\frac{1}{R} \sum_{r=1}^R p(y_i | \beta_i^{(r)}, q, \pi, \tilde{\sigma}) \right] \\
 &= \prod_{i=1}^N \left[\frac{1}{R} \sum_{r=1}^R \left(\pi' f_{i1}^{(r)} \left(\prod_{t=2}^{T-1} Q_{t-1} f_{it}^{(r)} \right) Q_{T-1} \tilde{f}_{iT}^{(r)} \right) \right],
 \end{aligned}$$

where $\beta_i^{(r)}$ is a draw from a distribution with density $f(\beta_i | \theta)$ and $f_{it}^{(r)}$ has m^{th} element $\Pr[Y_{it} = y_{it} | \beta_i^{(r)}, \tilde{\sigma}, s_{it} = m]$ for $r = 1, \dots, R$. The corresponding simulated log-likelihood function is given by

$$\log p(y|\theta, q, \pi, \tilde{\sigma}) \approx \sum_{i=1}^N \log \left[\frac{1}{R} \sum_{r=1}^R \left(\pi' f_{i1}^{(r)} \left(\prod_{t=2}^{T-1} Q_{t-1} f_{it}^{(r)} \right) Q_{T-1} \tilde{f}_{iT}^{(r)} \right) \right].$$

3.A.2 Hidden Markov rank-ordered logit model

For the HM-ROL, the likelihood function can be written as

$$\begin{aligned}
p(y|\theta, q, \pi, \tilde{\sigma}) &= \prod_{i=1}^N \left[\int \left(\sum_{s_i^* \in \mathcal{S}} \Pr[S_i = s_i^* | q, \pi] p(y_i | \beta_i, \tilde{\sigma}, s_i^*) \right) f(\beta_i | \theta) d\beta_i \right] \\
&= \prod_{i=1}^N \left[\int \left\{ \sum_{s_i^* \in \mathcal{S}} \left(\Pr[S_i = s_i^* | q, \pi] \prod_{t=1}^T \prod_{h=1}^{J-1} \frac{\exp\left(\frac{1}{\tilde{\sigma}_{s_{ih}}} (x'_{ity_{ith}} \beta_i)\right)}{\sum_{l=h}^J \exp\left(\frac{1}{\tilde{\sigma}_{s_{ih}}} (x'_{ity_{iul}} \beta_i)\right)} \right) \right\} f(\beta_i | \theta) d\beta_i \right] \\
&= \prod_{i=1}^N \left[\int \left\{ \sum_{s_i^* \in \mathcal{S}} \left(\Pr[S_i = s_i^* | q, \pi] \prod_{h=1}^{J-1} \prod_{t=1}^T \frac{\exp\left(\frac{1}{\tilde{\sigma}_{s_{ih}}} (x'_{ity_{ith}} \beta_i)\right)}{\sum_{l=h}^J \exp\left(\frac{1}{\tilde{\sigma}_{s_{ih}}} (x'_{ity_{iul}} \beta_i)\right)} \right) \right\} f(\beta_i | \theta) d\beta_i \right] \\
&= \prod_{i=1}^N \left[\int \left\{ \pi' f_{i1} \left(\prod_{h=2}^{J-2} Q_{h-1} f_{ih} \right) Q_{J-2} \tilde{f}_{i, J-1} \right\} f(\beta_i | \theta) d\beta_i \right], \tag{3.13}
\end{aligned}$$

where Q_h is a $(M \times M)$ transition probability matrix, and \tilde{f}_{ih} is a $(M \times 1)$ vector with the likelihood contribution of respondent i at rank h given β_i and s_{ih} with m^{th} element equal to

$$\tilde{f}_{ihm} = \prod_{t=1}^T \frac{\exp\left(\frac{1}{\tilde{\sigma}_m} (x'_{ity_{ith}} \beta_i)\right)}{\sum_{l=h}^J \exp\left(\frac{1}{\tilde{\sigma}_m} (x'_{ity_{iul}} \beta_i)\right)}.$$

Furthermore, f_{ih} is a diagonal $(M \times M)$ matrix with the diagonal equal to \tilde{f}_{ih} .

We approximate the likelihood function using Monte Carlo integration:

$$\begin{aligned}
p(y|\theta, q, \pi, \tilde{\sigma}) &\approx \prod_{i=1}^N \left[\frac{1}{R} \sum_{r=1}^R p(y_i | \beta_i^{(r)}, q, \pi, \tilde{\sigma}) \right] \\
&= \prod_{i=1}^N \left[\frac{1}{R} \sum_{r=1}^R \left(\pi' f_{i1}^{(r)} \left(\prod_{h=2}^{J-2} Q_{h-1} f_{ih}^{(r)} \right) Q_{J-2} \tilde{f}_{i, J-1}^{(r)} \right) \right],
\end{aligned}$$

where $\beta_i^{(r)}$ is a draw from a distribution with density $f(\beta_i | \theta)$ and $f_{ih}^{(r)}$ has m^{th} element $\prod_{t=1}^T \Pr[Y_{ith} = y_{ith} | y_{it1}, \dots, y_{it, h-1}, \beta_i^{(r)}, \tilde{\sigma}, s_{it} = m]$ for $r = 1, \dots, R$. The corresponding simulated log-likelihood function is given by

$$\log p(y|\theta, q, \pi, \tilde{\sigma}) \approx \sum_{i=1}^N \log \left[\frac{1}{R} \sum_{r=1}^R \left(\pi' f_{i1}^{(r)} \left(\prod_{h=2}^{J-2} Q_{h-1} f_{ih}^{(r)} \right) Q_{J-2} \tilde{f}_{i, J-1}^{(r)} \right) \right].$$

3.A.3 Miscellaneous details

The parameters q , π , and $\tilde{\sigma}$ are constrained, as are possibly several parameters in θ . To ensure unconstrained optimization of the simulated log-likelihood function, we reparametrize the constrained parameters in terms of parameters that are unconstrained and optimize over these unconstrained parameters. Furthermore, to increase the probability of finding a global maximum, we recommend using multiple starting values and picking the solution with gives the highest log-likelihood value.

Finally, we compute standard errors using the square root of the diagonal elements of the inverse of the negative Hessian of the log-likelihood function. We approximate the Hessian using the outer-product-of-gradients approximation based on the analytic gradient of the log-likelihood function. For this purpose, we consider the Hessian with respect to the untransformed, (possibly) constrained parameters in θ , q , π , and $\tilde{\sigma}$. Moreover, the log-likelihood function is again approximated using the same draws as used for the optimization.

3.B Conditional distribution of S_{it}

For the hidden Markov multinomial logit model, the distribution of S_{it} conditional on the choices y_i of individual i is a multinomial distribution with outcomes $1, \dots, M$ with corresponding probabilities that can be computed as follows. It holds that

$$\begin{aligned} \Pr[S_{it} = m | y_i, \theta, q, \pi, \tilde{\sigma}] &= \int \Pr[S_{it} = m, \beta_i | y_i, \theta, q, \pi, \tilde{\sigma}] d\beta_i \\ &= \int \Pr[S_{it} = m | y_i, \beta_i, \theta, q, \pi, \tilde{\sigma}] f(\beta_i | y_i, \theta, q, \pi, \tilde{\sigma}) d\beta_i \\ &= \int \Pr[S_{it} = m | y_i, \beta_i, q, \pi, \tilde{\sigma}] \frac{p(y_i | \beta_i, q, \pi, \tilde{\sigma})}{p(y_i | \theta, q, \pi, \tilde{\sigma})} f(\beta_i | \theta) d\beta_i \\ &= \frac{1}{p(y_i | \theta, q, \pi, \tilde{\sigma})} \int \Pr[S_{it} = m | y_i, \beta_i, q, \pi, \tilde{\sigma}] p(y_i | \beta_i, q, \pi, \tilde{\sigma}) f(\beta_i | \theta) d\beta_i, \end{aligned}$$

which can be approximated by

$$\begin{aligned} \Pr[S_{it} = m | y_i, \theta, q, \pi, \tilde{\sigma}] &\approx \frac{1}{p(y_i | \theta, q, \pi, \tilde{\sigma})} \frac{1}{R} \sum_{r=1}^R \Pr[S_{it} = m | y_i, \beta_i^{(r)}, q, \pi, \tilde{\sigma}] p(y_i | \beta_i^{(r)}, q, \pi, \tilde{\sigma}) \\ &= \frac{\sum_{r=1}^R \Pr[S_{it} = m | y_i, \beta_i^{(r)}, q, \pi, \tilde{\sigma}] \times p(y_i | \beta_i^{(r)}, q, \pi, \tilde{\sigma})}{\sum_{n=1}^M \sum_{r=1}^R \Pr[S_{it} = n | y_i, \beta_i^{(r)}, q, \pi, \tilde{\sigma}] \times p(y_i | \beta_i^{(r)}, q, \pi, \tilde{\sigma})}, \end{aligned}$$

where for the second equality we use that $\sum_m \Pr[S_{it} = m | y_i, \theta, q, \pi, \tilde{\sigma}] = 1$, and we let $\beta_i^{(r)}$ be a draw from a distribution with density $f(\beta_i | \theta)$ for $r = 1, \dots, R$. The probability $\Pr[S_{it} = m | y_i, \beta_i^{(r)}, q, \pi, \tilde{\sigma}]$ can be computed with the Hamilton filter (Hamilton, 1989) and a smoother (C.-J. Kim, 1994).

The Hamilton filter sequentially computes the filtered probabilities ($\xi_{itm|t} \equiv \Pr[S_{it} = m | \{y_{il}\}_{l=1}^t, \beta_i, q, \pi, \tilde{\sigma}]$) and predicted probabilities ($\xi_{i,t+1,m|t} \equiv \Pr[S_{i,t+1} = m | \{y_{il}\}_{l=1}^t, \beta_i, q, \pi, \tilde{\sigma}]$) using

$$\xi_{itm|t} = \frac{\xi_{itm|t-1} \Pr[Y_{it} = y_{it} | \beta_i, \sigma_{it} = \tilde{\sigma}_m]}{\sum_{n=1}^M \xi_{itn|t-1} \Pr[Y_{it} = y_{it} | \beta_i, \sigma_{it} = \tilde{\sigma}_n]},$$

$$\xi_{i,t+1,m|t} = \sum_{n=1}^M Q_{nm} \xi_{itn|t},$$

for $m = 1, \dots, M$ and $t = 1, \dots, T$. The filter is initialised by $\xi_{i1m|0} = \Pr[S_{i1} = m | \beta_i, q, \pi, \tilde{\sigma}] = \pi_m$. Given the filtered and predicted probabilities up to $t = T$, the required smoothed estimates can be computed sequentially using (C.-J. Kim, 1994)

$$\Pr[S_{it} = m | y_i, \beta_i, q, \pi, \tilde{\sigma}] = \sum_{n=1}^M \xi_{i,t+1,n|T} \frac{Q_{m,n} \xi_{itm|t}}{\xi_{i,t+1,n|t}},$$

for $t = T - 1, T - 2, \dots, 1$.

For the hidden Markov rank-ordered logit model, the conditional probabilities that S_{ih} is equal to a phase m can be computed in a similar fashion.

3.C Monte Carlo study: results DGPs 4-6

Table 3.7: Mean and RMSE (in parentheses) of the parameter estimates for the Monte Carlo study for DGPs 4 to 6. Based on 1,000 Monte Carlo replications per DGP.

Parameter	True	DGP 4 ^a		DGP 5 ^b		DGP 6 ^c				
		MNL	H-MNL	HM-MNL	MNL	H-MNL	HM-MNL	MNL	H-MNL	HM-MNL
b_1	1.00	0.65 (0.35)	0.84 (0.17)	0.97 (0.07)	0.67 (0.33)	0.86 (0.15)	0.91 (0.10)	0.43 (0.57)	0.67 (0.34)	0.86 (0.16)
b_2	0.30	0.20 (0.11)	0.25 (0.05)	0.29 (0.03)	0.20 (0.10)	0.26 (0.05)	0.27 (0.04)	0.13 (0.17)	0.20 (0.10)	0.26 (0.05)
b_3	-0.50	-0.33 (0.17)	-0.43 (0.09)	-0.48 (0.06)	-0.34 (0.16)	-0.43 (0.08)	-0.46 (0.07)	-0.22 (0.28)	-0.34 (0.17)	-0.43 (0.09)
$\sigma_{\beta,1}$	0.50	0.49 (0.03)	0.64 (0.15)	0.48 (0.07)	0.46 (0.05)	0.59 (0.10)	0.46 (0.06)	0.35 (0.16)	0.52 (0.06)	0.41 (0.12)
$\sigma_{\beta,2}$	0.40	0.30 (0.11)	0.39 (0.04)	0.39 (0.04)	0.30 (0.11)	0.38 (0.05)	0.37 (0.05)	0.21 (0.19)	0.32 (0.10)	0.34 (0.08)
$\sigma_{\beta,3}$	0.70	0.52 (0.19)	0.68 (0.08)	0.68 (0.08)	0.52 (0.19)	0.66 (0.08)	0.64 (0.09)	0.37 (0.34)	0.55 (0.17)	0.60 (0.14)
$\rho_{\beta,12}$	0.00	0.25 (0.26)	0.26 (0.27)	-0.02 (0.17)	0.20 (0.22)	0.20 (0.22)	0.00 (0.13)	0.26 (0.29)	0.26 (0.29)	-0.05 (0.27)
$\rho_{\beta,13}$	0.00	-0.25 (0.26)	-0.25 (0.27)	0.01 (0.17)	-0.19 (0.21)	-0.20 (0.22)	0.00 (0.15)	-0.26 (0.30)	-0.26 (0.30)	0.03 (0.26)
$\rho_{\beta,23}$	0.00	-0.12 (0.17)	-0.12 (0.17)	0.01 (0.14)	-0.09 (0.15)	-0.09 (0.16)	0.00 (0.14)	-0.13 (0.23)	-0.13 (0.23)	0.01 (0.20)
π_1			0.280	0.257			0.257			0.344
π_2			0.593	0.664			0.664			0.600
π_3			0.127	0.079			0.079			0.056
q_{11}			0.378	0.397			0.397			0.586
q_{22}			0.981	0.984			0.984			0.938

For DGPs 4 to 6, we simulate data from the HM-MNL in Equations (3.3)-(3.7) with $\beta_i \sim MVN(b, \Sigma_\beta)$. In the DGPs, the transition probabilities are set equal over tasks, and respondents can only move one phase up. More details:

^a DGP 4: 5 phases, $\bar{\sigma} = (\infty, 2, 1, 2, \infty)$, $\pi = (0.2, 0.2, 0.4, 0.1, 0.1)$, $q_{11} = 0.20$, $q_{22} = 0.35$, $q_{33} = 0.98$, and $q_{44} = 0.80$.

^b DGP 5: 5 phases, $\bar{\sigma} = (10, 2, 1, 2, 10)$, $\pi = (0.2, 0.2, 0.4, 0.1, 0.1)$, $q_{11} = 0.20$, $q_{22} = 0.35$, $q_{33} = 0.98$, and $q_{44} = 0.80$.

^c DGP 6: 9 phases, $\bar{\sigma} = (\infty, 10, 5, 2, 1, 2, 5, 10, \infty)$, $\pi = (0.1, 0.1, 0.1, 0.1, 0.4, 0.05, 0.05, 0.05)$, $q_{11} = q_{22} = q_{33} = q_{44} = 0.20$, $q_{55} = 0.90$, and $q_{66} = q_{77} = q_{88} = 0.70$

Chapter 4

A dynamic model of clickthrough and conversion probabilities of paid search advertisements

4.1 Introduction

Search engine advertising (SEA) has become an important marketing channel for firms (Ryan, 2016). SEA is an advertising form that allows firms to place advertisements on the search results pages of search engines such as Google, Yahoo! and Bing. Search engines select the ads to be shown based on an individual's search, enabling advertisers to target individuals. Most search engines use an auction to select the ads. In general, the advertiser who sets the highest (quality adjusted) bid obtains the most prominent position. Usually, up to three ads are shown both on top and on the bottom of the search results page for a given search query.

In designing a SEA strategy, an advertiser has to create text ads and landing pages, determine the search phrases for which an ad is eligible to show up (the *keywords*), and set a bid on each keyword. When the performances of ads change over time, a SEA strategy requires regular revision to be effective. Dynamics in ad performance

can result from e.g. the introduction of new products, changes in consumer tastes or populations, the entering of new competitors, or seasonality. Yet, there is little empirical research on dynamics in ad performance.

In this paper, we develop a dynamic Bayesian model for the performance of paid search ads in Google. The proposed model is especially suited to deal with dynamic SEA environments. It allows for dynamics through seasonal effects and time-varying parameters, and discriminates between permanent and transitory dynamics. Especially when shocks are long lasting, dynamic SEA strategies are required for long-term profitability. In our empirical application, we find evidence of substantial persistent time variation in ad performance, emphasizing the importance of addressing dynamics in ad performance models.

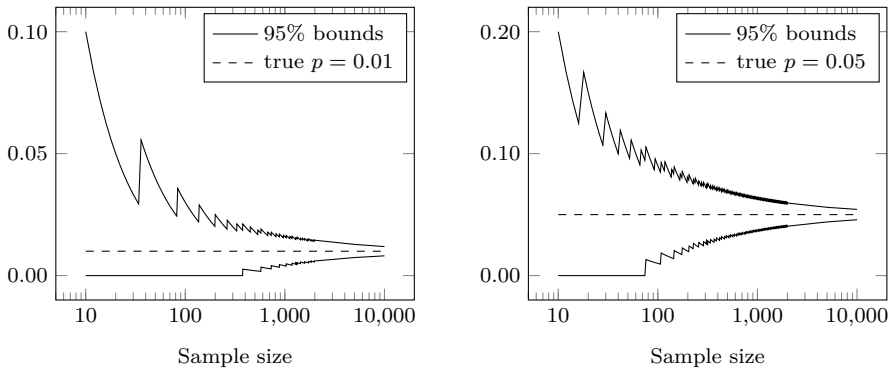
We model ad performance in terms of the clickthrough and conversion probabilities of keywords. In the context of SEA, the term "keyword" refers to a string of words, e.g. a keyword can be quite generic ("laptop") or more specific ("laptop acer vn7 571g"). An advertiser links each of her ads to a set of keywords, and sets a bid on each keyword. When a consumer's search query matches a keyword, the associated ad will be eligible for the auction. The clickthrough probability is the probability that a consumer who gets served an ad due to the keyword, clicks on the ad. The conversion probability is the probability that a consumer who has already clicked on the ad converts, that is, buys a product or service.

A number of studies have constructed models for clickthrough and conversion probabilities of keywords (see Rutz and Bucklin, 2007, Ghose and Yang, 2009, Agarwal et al., 2011, and Rutz et al., 2012). Our study differs from these papers by focusing on the *dynamics* of keyword performance.

Despite the availability of advertiser-level ad performance data, a number of challenges arise when estimating clickthrough and conversion probabilities of keywords. Next to the challenge of potential dynamics in ad performance, a second challenge is data sparsity. The majority of keywords in an advertiser's portfolio generate only little traffic, that is, few users search for that keyword. For these so-called sparse keywords, taking simple averages of realized clicks and conversions in the past is insufficient to estimate the clickthrough and conversion probabilities as these estimates can be highly inaccurate. We illustrate this in Figure 4.1, which shows the proportion of clicks (conversions) we can expect to observe given a certain sample size (number of impressions¹ or clicks) for two realistic probabilities: 1% and 5%. The figure

¹The number of impressions is the number of times an ad is shown on the search results page.

Figure 4.1: Conservative 95% bounds of the observed proportion of clicks/conversions as a function of the sample size on a logarithmic scale (true probability equals 1% (left) or 5% (right)).



shows that one needs at least a couple of hundred impressions (clicks) for the sample average to be a reliable estimator of the true clickthrough (conversion) probability. If a keywords generates little traffic, it may take a long time before this number of impressions and clicks are collected. Within this time frame the true probability may, in practical situations, already have changed.

The third challenge is estimating the causal effect of ad position on ad performance, as the position is endogenously related to clickthrough and conversion probabilities. There are three sources of endogeneity:

- i There is a potential reversed causality relationship due to strategic bidding behavior: an advertiser might bid more for keywords with a high expected clickthrough and conversion probability, to obtain a favorable position for these keywords.
- ii There is a reversed causality relationship due to a quality adjustment in the keyword auction (Google uses the so-called quality score): position might not only affect clickthrough probabilities, but reversely, the previous clickthrough rates affect the position through their impact on the quality adjustment.
- iii There is a potential confounding factor: competition is likely to affect both keyword performance and ad position, but is unobserved by the advertiser.

In this paper, we propose a dynamic model that addresses all above challenges by allowing for explained and unexplained dynamics, data sparsity, missing data, position endogeneity and unobserved heterogeneity across keywords. The model captures unexplained dynamics through time-varying parameters that follow either stationary

or nonstationary AR(1) processes to distinguish between transitory and permanent dynamics. The model addresses the sparsity problem by linking keywords to each other based on common factors such as semantic keyword characteristics. Finally, the model accounts for position endogeneity in the manner proposed by Ghose and Yang (2009). That is, we simultaneously model the consumers' clickthrough and conversion behavior, the search engine's position allocating behavior, and the firm's bidding behavior, and correlate the error terms of the equations with each other.

The resulting model is estimated using a Bayesian approach. We develop an efficient Gibbs sampler with Polya-Gamma data augmentation for the logit part of the model (Geman and Geman, 1987, Tanner and Wong, 1987, Polson et al., 2013) in which we draw the time-varying parameters using the forward-filtering backward-sampling algorithm of Durbin and Koopman (2002). This efficient approach is crucial to be able to use the methodology at a daily frequency for a realistically large number of keywords. It also deals naturally with missing data.

We illustrate the model using a unique dataset from a Dutch online retailer that sells laptops and advertises on Google. The data consists of the historical performance of 14,710 keywords measured at a daily frequency over the period January 2014 until March 2016. We find substantial time variation in clickthrough and conversion probabilities, indicating that a dynamic SEA strategy is required. Furthermore, we find that shocks mostly have a permanent or highly persistent effect on clickthrough probabilities; this holds for market-level shocks and most brand-level shocks. For conversion probabilities the shocks have different effects on different type of ads. Whereas market shocks permanently affect conversion probabilities, most brand-level shocks have a more transitory effect. Finally, Bayes factors indicate that the dynamic model is substantially better in forecasting ad performance than the static model.

We also find evidence of position and bidding endogeneity, indicating that purely predictive models are unable to capture causal relationships between ad position and clickthrough and conversion probabilities.

The managerial implications of this paper are threefold. First, advertisers can use the model to obtain accurate daily estimates of clickthrough and conversion probabilities of individual keywords. These estimates can be used to set bids, adjust text ads and landing pages, and to identify keywords whose performance is divergent from similar keywords. Second, advertisers can examine the extent of dynamics in their SEA environment, to determine how often their bidding strategy should be revised. In doing so, advertisers can discriminate between keywords by using the persistence and

influence of shocks on different types of keywords. Finally, advertisers can use the model to track the performance of keywords to timely identify when this performance changes.

The remainder of this paper is organized as follows. In Section 4.2 we discuss the background for this research. We explain how the mechanism underlying SEA works and discuss related work on modeling clickthrough and conversion probabilities. In Section 4.3 we briefly discuss the data generally made available by search engines to further discuss the context of SEA. Section 4.4 is devoted to a detailed discussion of the methodology. We show empirical results in Section 4.5 including an analysis of the model's predictive performance against a static model. Section 4.6 discusses the managerial implications of this research. We conclude with a summary and a critical discussion. Finally, Appendix 4.A documents our efficient Gibbs sampler in detail.²

4.2 Background

4.2.1 The mechanism underlying search engine advertising

From the search engine's perspective, much literature has focused on the mechanism design of the keyword auction (see e.g. Borgs et al., 2007, Cary et al., 2007, Edelman et al., 2007, and Yao and Mela, 2011). The design Google and Yahoo! use is formally known as a generalized, second-price, sealed-bid auction (Edelman et al., 2007).

This real-time keyword auction works as follows. Advertisers link their ads to keywords and place a bid on each keyword. The bid indicates the maximum amount the advertiser is willing to pay for a click. Some search engines such as Google also assign a quality score to an ad, to adjust the bids for relevance of the advertised website. Next, when a consumer enters the search query at a search engine, the engine considers all advertisers' ads for which the associated keywords match the consumer's search. The available ad slots are allocated according to the advertisers' quality adjusted bids. The search engine only charges the advertiser a fee when a consumer clicks on the ad; this fee is known as the cost-per-click (CPC). The CPC is based on the bid of the ad that is ranked just below (the second price), corrected for the quality scores of these two ads. The CPC is thus not necessarily equal to the bid, but it is never higher.

The distinct feature of the generalized, second-price auction is that bidders pay a price

²The Supplementary Materials, containing all results of the empirical application as well as trace plots and effective sample sizes of the MCMC output, are available upon request.

based on the bid of the advertiser ranked below. Hereby, search engines avoid that advertisers use cycling bidding strategies to optimize profits, that is, that advertisers continuously decrease their bids until they obtain a less prominent ad position after which they increase their bids again (Borgs et al., 2007).

4.2.2 Modeling clickthrough and conversion probabilities of keywords

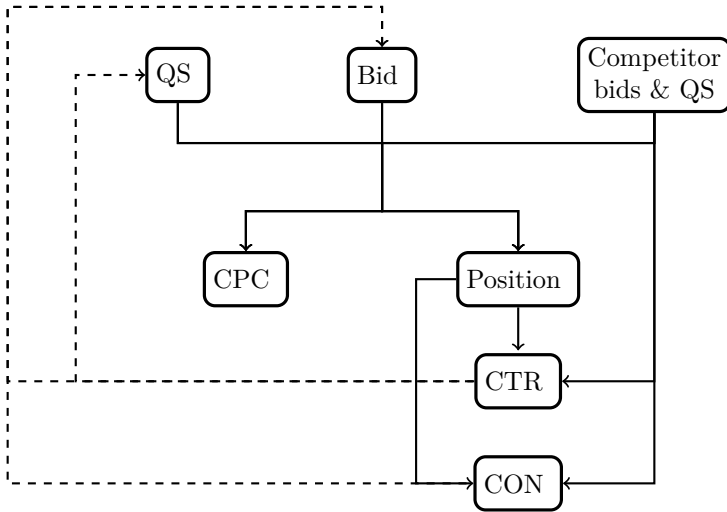
From the advertisers' perspective, some literature has focused on modeling clickthrough and conversion probabilities of keywords (see e.g. Ghose and Yang, 2009, Agarwal et al., 2011, and Rutz et al., 2012). The key focus of these studies is addressing position endogeneity.

To better understand the sources of position endogeneity, we conceptualize the mechanism underlying the keyword auction in Figure 4.2. Based on the inputs of the keyword auction (the advertiser's bid and quality score and the competitors' bids and quality scores) the search engine determines the ad position and the cost-per-click. The ad position then potentially affects consumers' clickthrough and conversion behavior.

There are three potential sources of endogeneity. First, competition can be a confounding factor as it is unobserved and both enters the keyword auction to determine the ad position as well as potentially affects clickthrough and conversion probabilities through consumer behavior. Second, there is a potential reversed causality problem as some search engines, including Google, use the past clickthrough rates to assign quality scores to keywords to determine the ad position. Finally, there is a second potential reversed causality problem due to strategic bidding behavior. An advertiser might set bids based on expected clickthrough and conversion probabilities for different ad positions (we refer to this as *bidding endogeneity*).

To correct for all these sources of endogeneity, the earlier mentioned studies use parametric simultaneous equations models of the clickthrough and conversion probabilities and the ad's position plus a specific strategy to solve for bidding endogeneity. Agarwal et al. (2011) use data on randomized bids to explain the position. Alternatively, Rutz et al. (2012) use latent instrumental variables (LIVs) to explain the position. In the LIV approach, the endogenous variable (ad position) is split into a part that is uncorrelated with the error terms of the clickthrough and conversion equations, the latent instruments, and a part that is potentially correlated. Finally,

Figure 4.2: Conceptual model of keyword performance. Solid lines represent contemporary causal effects, dashed lines represent future causal effects.



QS: quality score,
CPC: cost-per-click,
CTR: clickthrough rate,
CON: conversion rate.

Ghose and Yang (2009) simultaneously model the firm’s bid with the clickthrough and conversion probabilities and the ad’s position.

In this paper we opt for the approach by Ghose and Yang (2009). Although randomized bids as used by Agarwal et al. (2011) yield a better source of variation to identify the causal impact of position, it is rare to find firms who actually practice randomized bidding. A drawback of the LIV approach of Rutz et al. (2012) is that it relies on the existence of latent “groups” that are correlated with position and uncorrelated with the unexplained parts of the clickthrough and conversion probabilities. In general it is unknown whether such groups exist.

The above mentioned studies find mixed results regarding the drivers of keyword performance. Generally, they agree that the more prominent the position, the higher the clickthrough probability (Ghose and Yang, 2009, Agarwal et al., 2011, and Rutz et al., 2012). Furthermore, Agarwal et al. (2011) and Ghose and Yang (2009) find that profits are usually not highest in the top positions. Instead, profits increase until some position when going down the search results page after which they decrease again. These studies ignore dynamics other than day-of-the-week effects.

From the search engine’s perspective there is also literature on modeling clickthrough probabilities. These models are used to estimate quality scores of ads to help allocate ads on the search results pages. The models proposed in this literature are predictive models, no steps are taken to account for position or bidding endogeneity. One such model that allows for dynamic performance is proposed in Graepel et al. (2010), who develop a Bayesian model for clickthrough probabilities. This model allows for dynamics by adjusting the parameters as new data comes in through a Bayesian learning algorithm that gives higher weight to more recent observations.

4.3 General structure of data

Google provides advertisers with a number of ad performance metrics. These metrics are aggregated on the level of the keyword and some time period, such as the hour of the day or day of the week. The performance metrics include the number of obtained impressions, clicks, and conversions, the average position over the impressions, and the average cost-per-click (CPC). Google also provides four metrics related to the quality score: quality score (ranging from 1 to 10), landing page experience, ad relevance, and expected clickthrough rate. The quality score metric is, however, not the actual measure used by Google in real-time to assign positions to ads.³

Based on the words in the keyword, an advertiser can construct semantic characteristics of keywords. These characteristics might be useful in estimating keyword performance. They can include the number of words in the keyword, or the specificity of the keyword (e.g. generic, branded or retailer-specific search like in Ghose and Yang, 2009). In addition, the advertiser knows the match type of each keyword. The match type of a keyword refers to how “well” the keyword must match the consumers search in order to be eligible to show, and is one of ‘exact’, ‘broad’, or ‘phrase’. The broader the match type, the more divergent the search phrase and the keyword may be.

4.4 Methods

In this section, we discuss the statistical model we propose for keyword performance. We consider model specification, parameter identification, and model inference.

³For more information on the quality score, see <https://support.google.com/googleads/answer/7050591?hl=en>.

The model we propose is a dynamic Bayesian model of the consumers' clickthrough and conversion behavior, the search engine's position allocating behavior, and the firm's bid behavior. The model is a time-varying parameters model, also known as a model in state-space representation (e.g. Hamilton, 1994, Chapter 13).

4.4.1 Model specification

We index the keywords by $i = 1, \dots, N$ and time periods by $t = 1, \dots, T$. We denote by I_{it} , N_{it} , and M_{it} the number of impressions, clicks, and conversions of keyword i at time t , respectively. By definition, $I_{it} \geq N_{it} \geq M_{it}$. Furthermore, we denote by p_{it}^{CTR} the unobserved clickthrough probability on keyword i at time t (the probability of a click given an impression) and by p_{it}^{CON} the conversion probability (the probability of a conversion given a click). Let POS_{it} denote the average ad position over the impressions on keyword i at time t , BID_{it} the bid, CPC_{it} the cost-per-click, and QS_{it} the quality score. Finally, let x_i denote a $(K \times 1)$ vector of (semantic) characteristics of keyword i , and s_t a vector of seasonal dummies.

We assume that, conditional on $\{p_{it}^{CTR}\}_{t=1}^T$ and $\{p_{it}^{CON}\}_{t=1}^T$, the impressions and clicks on keyword i at time t are independent across ads served, and that the impressions and clicks on keyword i are independent across time and independent of other keywords $j \neq i$. Then, the number of clicks on keyword i at time t and the number of conversions are conditionally binomially distributed. That is,

$$\begin{aligned} N_{it}|I_{it}, p_{it}^{CTR} &\sim \text{BIN}(I_{it}, p_{it}^{CTR}), \\ M_{it}|N_{it}, p_{it}^{CON} &\sim \text{BIN}(N_{it}, p_{it}^{CON}), \end{aligned}$$

for $i = 1, \dots, N$ and $t = 1, \dots, T$.

Next, we propose a dynamic simultaneous equations model of the clickthrough and conversion probabilities, the ad's position, and the firm's bid. The model is given by

$$p_{it}^{CTR} = \Lambda(\alpha_i^{CTR} + x_i' \beta_t^{CTR} + \lambda_i^{CTR} \ln(POS_{it}) + s_t' \gamma^{CTR} + \eta_{it}^{CTR}), \quad (4.1)$$

$$p_{it}^{CON} = \Lambda(\alpha_i^{CON} + x_i' \beta_t^{CON} + \lambda_i^{CON} \ln(POS_{it}) + s_t' \gamma^{CON} + \eta_{it}^{CON}), \quad (4.2)$$

$$\ln(POS_{it}) = \alpha_i^{POS} + x_i' \beta_t^{POS} + \lambda_i^{POS} \ln(BID_{it}) + \psi^{POS} \ln(QS_{it}) + s_t' \gamma^{POS} + \eta_{it}^{POS}, \quad (4.3)$$

$$\ln(BID_{it}) = \alpha_i^{BID} + x_i' \beta_t^{BID} + q_{it}' \delta_i^{BID} + s_t' \gamma^{BID} + \eta_{it}^{BID}, \quad (4.4)$$

for keywords $i = 1, \dots, N$ and time periods $t = 1, \dots, T$, where $\Lambda(\theta) \equiv 1/(1 + \exp(-\theta))$ is the standard logistic link function.

The key parts of the model are the clickthrough and conversion equations (4.1) and

(4.2). These two equations have an equivalent functional form with different parameters. The logistic link function is used to map real-valued numbers to probabilities between 0 and 1. The baseline levels of the clickthrough and conversion probabilities are captured in the keyword-specific intercepts α_i , discussed in detail in Section 4.4.1.2. Stochastic dynamics are captured in the term $x'_i\beta_t$, which captures the time variation in clickthrough and conversion probabilities for different types of keywords. The process for β_t captures the carryover effects of shocks to subsequent periods, and is discussed in Section 4.4.1.1. Deterministic seasonal effects are captured in the term $s'_t\gamma$. The effect of ad position is captured in the keyword-specific parameters λ_i and is discussed in Section 4.4.1.2.

The position equation (4.3) is included to correct for position endogeneity and the bid equation (4.4) to correct for bidding endogeneity. For both equations we use a linear specification for the log transformed variables. They also deviate from the clickthrough and conversion equations in that we let the position depend on the bid and quality score, and let the bid depend on q_{it} , a vector of instrumental variables not included in the other equations. We discuss these instruments in Section 4.4.1.3. Note that the position equation (4.3) can be rewritten as

$$\text{POS}_{it} = g_{it}^{POS} \text{QS}_{it}^{\psi^{POS}} \text{BID}_{it}^{\lambda^{POS}},$$

where the multiplication factor g_{it}^{POS} depends on x_i , s_t and η_{it} in a potentially time-varying way. Hence, we assume that the position depends on the bid and the quality score in a multiplicative way. The parameters ψ^{POS} and λ^{POS} are elasticity parameters; if the bid increases by 1%, then the position increases by $\lambda^{POS}\%$.

The key elements in correcting for position endogeneity are the keyword- and time-specific error terms in $\eta_{it} = (\eta_{it}^{CTR}, \eta_{it}^{CON}, \eta_{it}^{POS}, \eta_{it}^{BID})'$. We assume that η_{it} is multivariate normally distributed for keyword i and time t and independent across keywords and time. That is,

$$\eta_{it} \sim MVN(0, \Sigma_{\eta}), \quad (4.5)$$

where all elements of the positive definite matrix Σ_{η} are allowed to be non-zero.

Even when no position endogeneity is present, it is important to include η_{it} into the clickthrough and conversion equations. The model is based on the aggregation of choices on the keyword- and time-level. The parameter η_{it} captures keyword- and time-specific deviations that are not captured by other model parameters. In

case a keyword i receives many observations and clicks in a given time period t , the likelihood of p_{it}^{CTR} and p_{it}^{CON} given the observed data is highly peaked at the observed fractions of clicks/conversions. Hence, the estimation procedure will model the clickthrough and conversion probabilities to be (almost) equal to the realized proportions in the data. In case η_{it}^{CTR} and η_{it}^{CON} are included, they can capture potential deviations between expected and realized proportions. In case they are not included, the estimates of the parameters will become such that they mainly fit these few observations with many impressions, instead of representing general patterns across the whole set of keywords.

4.4.1.1 Time-varying parameters: the dynamic impact of shocks

The impact of changes in the environment on ad performance is captured in the time-varying parameters $\beta_t = (\beta_t^{CTR'}, \beta_t^{CON'}, \beta_t^{POS'}, \beta_t^{BID'})'$. Changes in the environment can result from changes in macroeconomic conditions, in the firm (e.g. changing reputation), in the market competitiveness (e.g. new competitor or the launch of a new product), in the search engine's position-allocating mechanism, or in consumers (e.g. changing tastes and attitudes). The effect of changes on ad performance can be transitory or permanent.

To capture the dynamics in SEA environments, we take independent AR(1) processes for the time-varying parameters. That is,

$$\beta_{t+1} = \Phi\beta_t + \nu_t, \quad \nu_t \sim MVN(0, \Sigma_\beta), \quad \beta_1 \sim MVN(0, 5\Sigma_\beta), \quad (4.6)$$

for $t = 1, \dots, T$, where Φ and Σ_β are $(4K \times 4K)$ diagonal matrices. These AR(1) processes can capture a wide variety of paths for the time-varying parameters (Van Heerde et al., 2004).

The autoregressive parameters $\{\phi_k\}_{k=1}^{4K}$ on the diagonal of Φ measure the persistence of the impact of shocks on future values of β_{kt} . In case $\phi_k = 1$, shocks are permanent. In case $\phi_k = 0$, shocks do not impact future clickthrough and conversion probabilities, and a static model would do. In case $0 < \phi_k < 1$, the effects of shocks carry over to next periods but the process is mean-reverting: shocks die out geometrically with decay rate ϕ_k .

The AR(1) processes in Equation (4.6) have no intercept. If the β_t series is stationary, an intercept captures the unconditional mean of the series. If the series is nonstationary, the intercept would either capture the level at time $t = 1$ (in deviation-

from-mean form), or a drift parameter (in regular form). We move this intercept to the mean of the α_i parameter as will be explained next. This ensures that we can interpret the intercept as in the deviations-from-mean form and thus have equal interpretation of the parameter in case of stationarity and nonstationarity of the β_t series. Moreover, it helps improve the mixing rates of the sampler.

4.4.1.2 Unobserved heterogeneity across keywords

The model captures unobserved heterogeneity across keywords through the keyword-specific parameters $\alpha_i = (\alpha_i^{CTR}, \alpha_i^{CON}, \alpha_i^{POS}, \alpha_i^{BID})'$ and $\lambda_i = (\lambda_i^{CTR}, \lambda_i^{CON}, \lambda_i^{POS})'$. We shrink α_i and λ_i to common means across similar keywords.

The parameters in α_i capture common baseline levels of ad performance as well as keyword-specific deviations. We take α_i to be independently normally distributed across keywords,

$$\begin{pmatrix} \alpha_i^{CTR} \\ \alpha_i^{CON} \\ \alpha_i^{POS} \\ \alpha_i^{BID} \end{pmatrix} \sim MVN \left(\begin{pmatrix} x_i' \tilde{\alpha}^{CTR} \\ x_i' \tilde{\alpha}^{CON} \\ x_i' \tilde{\alpha}^{POS} \\ x_i' \tilde{\alpha}^{BID} \end{pmatrix}, \begin{bmatrix} \sigma_{\alpha,CTR}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\alpha,CON}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\alpha,POS}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\alpha,BID}^2 \end{bmatrix} \right), \quad (4.7)$$

for $i = 1, \dots, N$, where $\tilde{\alpha} = (\tilde{\alpha}^{CTR}, \tilde{\alpha}^{CON}, \tilde{\alpha}^{POS}, \tilde{\alpha}^{BID})$ captures the common baseline levels.

The parameters in λ_i capture the impact of ad position on clickthrough and conversion probabilities $(\lambda_i^{CTR}, \lambda_i^{CON})$, and the effect of bid on ad position (λ_i^{POS}) . We take λ_i to be independently normally distributed across keywords. That is,

$$\begin{pmatrix} \lambda_i^{CTR} \\ \lambda_i^{CON} \\ \lambda_i^{POS} \end{pmatrix} \sim MVN \left(\begin{pmatrix} x_i' \tilde{\lambda}^{CTR} \\ x_i' \tilde{\lambda}^{CON} \\ x_i' \tilde{\lambda}^{POS} \end{pmatrix}, \begin{bmatrix} \sigma_{\lambda,CTR}^2 & 0 & 0 \\ 0 & \sigma_{\lambda,CON}^2 & 0 \\ 0 & 0 & \sigma_{\lambda,POS}^2 \end{bmatrix} \right), \quad (4.8)$$

for $i = 1, \dots, N$.

4.4.1.3 Instrumental variables

The instruments q_{it} in the bid equation (4.4) are necessary for identification of the dynamic simultaneous equations model. They must be excluded from the other equations (4.1)-(4.3). A researcher can take any set of valid instruments: the instruments should be correlated with the bid, but uncorrelated with ad position and clickthrough

and conversion probabilities after correcting for bid/position.

We propose to use previous performance indicators as instruments, as these indicators capture the potential strategic bidding behavior of advertisers that causes the bidding endogeneity. We consider the previous clickthrough rate and the previous number of impressions obtained.⁴ For the previous clickthrough rate to be a valid instrument, we assume that the quality score measure we include in the position equation is a sufficient statistic for the previous clickthrough rate in explaining ad position.

To allow for heterogeneity in the effect of the instruments we consider keyword-specific parameters

$$\delta_i^{BID} \sim MVN(x_i' \bar{\delta}^{BID}, \Sigma_{\delta}^{BID}), \quad (4.9)$$

where Σ_{δ}^{BID} is a positive definite diagonal matrix.

4.4.2 Parameter identification

To ensure that the parameters in the model in (4.1)-(4.9) are identified, we have to consider two issues. First, for the stochastic dynamics part $x_i' \beta_t$, a researcher may wish to include many characteristics such that the matrix $X = (x_1, x_2, x_3, \dots, x_N)'$ is not of full column rank. For example, a researcher may want to include an intercept and all dummies for a categorical variable, to distinguish between market-level shocks and the lower level shocks for different categories. In this case, where X is not of full column rank, identification restrictions need to be imposed. More specifically, a set of variables k^* has to be selected, such that the matrix X without the columns corresponding to these variables in k^* is of full column rank. For these variables in k^* , the following restrictions are sufficient for identification: (i) $\beta_1 = 0$, (ii) $\tilde{\alpha} = 0$, and (iii) $\tilde{\lambda} = 0$. Note that these variables will still have a non-zero effect for $t > 1$.

Second, the simultaneous equations model in (4.1)-(4.4) is identified as the model is a triangular system (Greene, 2012, Ghose and Yang, 2009): the bid equation depends only on exogenous variables, the position equation depends only on the endogenous variable bid, and the clickthrough and conversion equations depend only on the endogenous variable ad position. Identification in this triangular system is ensured through the exclusion restrictions that the instrumental variables in the bid equation are excluded in the clickthrough, conversion, and position equations, and the

⁴Ghose and Yang (2009) use the lagged ad position as instrument in the bid equation. Exogeneity of this instrument depends on the assumption that the error terms in the position equation are serially uncorrelated. This assumption might very well be unrealistic, rendering lagged ad position invalid as instrument.

bid variable in the position equation is excluded in the clickthrough and conversion equations. Hence, the model is identified and we do not need to impose restrictions on the covariance matrix Σ_η .

4.4.3 Bayesian inference

We perform Bayesian inference for the dynamic simultaneous equations model in (4.1)-(4.9). We use Markov Chain Monte Carlo (MCMC) techniques and rely on a Gibbs sampler with Polya-Gamma data augmentation (Geman and Geman, 1987, Tanner and Wong, 1987, Polson et al., 2013). The advantage of using a Bayesian estimation approach is that we can use informative priors for keyword characteristics that are very rare. The Gibbs sampler also deals naturally with missing values.

The Polya-Gamma data augmentation scheme is suitable for binomial logistic regression models (Polson et al., 2013). It allows for exact inference by introducing one layer of Polya-Gamma distributed latent variables, where the latent variables are drawn at the level of the keyword and time period. Alternative MCMC approaches for Bayesian inference for logistic regression models are (i) data augmentation schemes where the logistic distributed error terms are approximated by mixtures of normals (C.C. Holmes and Held, 2006, Frühwirth-Schnatter and Frühwirth, 2010) or (ii) an independence or random walk Metropolis-Hastings (MH) algorithm without data augmentation (Rossi et al., 2005). The disadvantages of the alternative data augmentation schemes are that they are not exact, require two layers of auxiliary variables, and require much more memory storage as augmentation is performed on the level of an impression or click and not on the total number of impressions and clicks (this is especially relevant for the SEA application). The disadvantage of the MH algorithms is that they often have poor mixing rates and that tuning may be required (Frühwirth-Schnatter & Frühwirth, 2010). This is especially important when dynamic states are involved.

The Gibbs sampler we use is outlined in Appendix 4.A. In this Gibbs sampler, we subsequently draw the auxiliary variables from the Polya-Gamma distribution, the time-varying parameters from a multivariate normal distribution using the forward-filtering backward-sampling (FFBS) algorithm (Carter and Kohn, 1994, Frühwirth-Schnatter, 1994) of Durbin and Koopman (2002) and collapsed filtering (Durbin and Koopman, 2012, Jungbacker and Koopman, 2015), the time-invariant parameters from multivariate normal distributions, and the covariance matrices from inverse Wishart distributions. Specialized code is written in R (R Core Team, 2013) and

C++ (Eddelbuettel & François, 2011).

4.5 Empirical application

In this section, we apply the proposed dynamic Bayesian model to data of a Dutch online retailer. We present the data in Section 4.5.1 and discuss the in-sample results in Section 4.5.2. Finally, in Section 4.5.3, we compare the performance of our dynamic model to a static model without time-varying parameters.

4.5.1 Data

The data contains the historical performance of 14,710 keywords related to laptops measured at the daily frequency over the period January 1, 2014 until March 31, 2016.⁵ The data contains information on the daily number of impressions, clicks, and conversions⁶, and the daily average cost-per-click, ad position, and quality score⁷. We consider all data for model inference. In total, the keywords obtained 47.0 million impressions, 1.6 million clicks and 33.0 thousand conversions. The average clickthrough rate was 3.4% and the average conversion rate was 2.0%. Moreover, the top 5% of keywords based on impressions accounted for 92.5% of total impressions, whereas the bottom 50% accounted for 0.2% of total impressions.

We also use semantic characteristics of keywords. Each keyword is assigned to one of four categories indicating the specificity of the keyword: (i) ‘generic’, (ii) ‘brand only’, (iii) ‘brand and series’, or (iv) ‘retailer’. The ‘brand only’ keywords are keywords that include the brand name of a laptop but not the name of a specific series or model (e.g. ‘asus laptop’), whereas the ‘brand and series’ keywords include at least a brand’s series name (e.g. ‘asus vivobook’). We divide the keywords in the ‘brand only’ and ‘brand and series’ categories into the eight brands available at the retailer: Acer, Apple, Asus, HP, Lenovo, Microsoft, MSI, and Toshiba.

⁵Google provides data aggregated on the device used by the consumer (computer, tablet, or mobile device). We only include data on searches made via the computer, as consumer behavior might differ for the three electronic devices and the comparative usage of the three devices might have changed over time.

⁶Conversions are measured based on the keyword associated with the *last clicked ad* by the consumer as tracked by Google. Conversions are counted when the consumers makes a purchase within 30 days of clicking on the last clicked ad.

⁷We do not have data on the landing page experience, ad relevance, and expected click-through rate. Furthermore, we impute missing quality scores with a 6. Quality scores are missing when there were insufficient previous impressions and clicks for Google to determine the quality score. In these cases, Google uses a quality score of 6 in the keyword auction, see <https://searchengineland.com/google-adwords-keyword-quality-score-reporting-update-226355>.

Furthermore, we know the match type of the keyword, which is either broad or exact, and the number of words in the keyword. For these two variables we consider time-invariant parameters, that is, $\beta_t = 0$. We include seasonal dummies for the day of the week.

As instruments, we use the previous clickthrough rate and the natural logarithm of the previously obtained number of impressions. We compute the previous clickthrough rate by taking the realized clickthrough rate over the previous month.⁸ Once the advertiser has implemented the model to set the bid, the previous clickthrough rate can be estimated using Equation (4.1) instead. Furthermore, for the previous impressions we consider the average daily number of impressions obtained on a keyword in the previous month.

Finally, we use the CPC as a proxy for the bid as done in Ghose and Yang (2009) and Skiera and Abou Nabout (2013). Data on historical bids are not provided by Google, and have not been stored by the retailer. Using the CPC instead of the bid is justified for competitive keywords, where the difference between the CPC and the bid is small (Abou Nabout et al., 2012). A disadvantage is that we do not always observe the CPC when we observe the position. We therefore impute the missing CPCs for explaining position, using a stochastic local level model.⁹

4.5.2 Baseline results

In this section, we discuss the in-sample results for the proposed dynamic Bayesian model. Posterior results are obtained using 35,000 simulations after 5,000 burn-in draws. We keep every 4th draw to deal with the correlation in the chain. Here, we show the most important results.

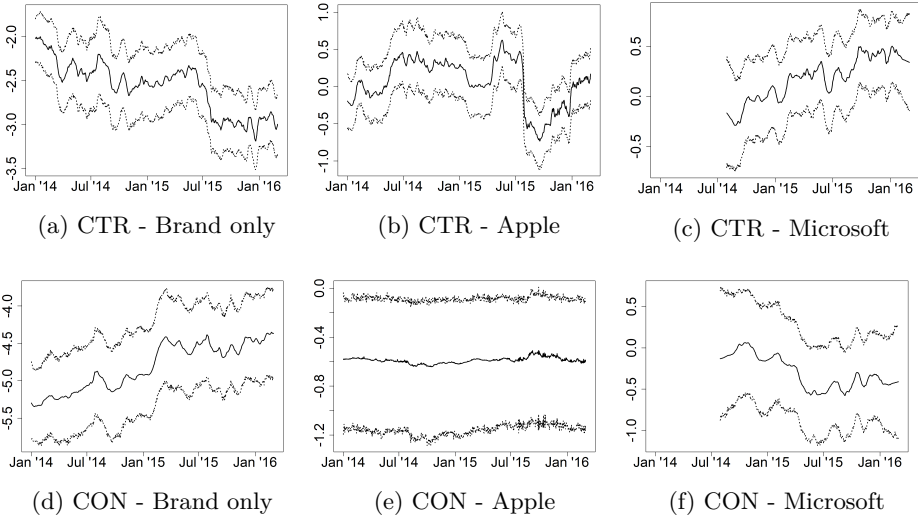
We find that clickthrough and conversion probabilities have substantially changed over time. Figure 4.3 shows illustrative examples of the smoothed estimates and 95% pointwise highest posterior density intervals (HPDIs) of the time-varying parameters. For the brand only keywords, we find that clickthrough probabilities have substantially decreased over time, whereas conversion probabilities have increased.

⁸In case a keyword obtained at least 5,000 impressions in the previous month, we take the clickthrough rate (CTR) of that *specific keyword*. Otherwise, we take the CTR over the *campaign group* the keyword was assigned to or, if that campaign group received less than 5,000 impressions in the previous month, the *specificity category* the keyword was assigned to.

⁹The local level model is given by $CPC_{i,t+1} = \mu_{it}$ with $\mu_{i,t+1} = \mu_{it} + \varepsilon_{it}$, $\mu_{i1} \sim N(\overline{CPC}, 0.5)$, and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$, where \overline{CPC} is the mean of all CPCs in the dataset and we estimate σ_ε^2 with maximum likelihood ($\hat{\sigma}_\varepsilon^2 = 0.007$ based on a set of popular keywords). We use the smoothed estimates of CPC_{it} .

The conversion performance of Apple keywords was stable, whereas the clickthrough performance was volatile. For Microsoft laptops, introduced at the retailer in August 2014, we see quite some time variation in ad performance with alternating periods of high and low clickthrough and conversion probabilities.

Figure 4.3: Posterior means and 95% highest posterior density intervals of $\{\beta_t\}_{t=1}^T + \tilde{\alpha}$ for the brand only keywords, retailer-specific keywords, and Microsoft keywords. For both the clickthrough (CTR) and conversion (CON) probabilities.



The 95% point-wise HPDIs in Figure 4.3 are quite wide. This is not so much caused by uncertainty in the dynamics in the time-varying parameter series (as the different smoothed draws follow similar dynamics), but is mainly caused by uncertainty in how the absolute levels should be attributed to the higher-level brand only effect and the lower-level brand effects (e.g. Apple, Microsoft). Adding the brand only effect to any of the brand effects, we find much smaller 95% HPDIs.

To assess the persistence of shocks — how long shocks carryover to next periods — we consider the posterior results for the autoregressive parameters in Φ in Table 4.1. We also compute the half-life of shocks. The half-life is the number of weeks before the effect of the shock is below 50% from its original value.¹⁰ For both the clickthrough and conversion probabilities, we find that shocks at the specificity level are permanent or highly persistent. The brand-level shocks on clickthrough perform-

¹⁰The half-life of shocks (in days) d can be computed by equating $\phi^d = 0.5$, that is, $d = \ln(0.5)/\ln(\phi)$ where ϕ is the posterior mean of the autoregressive parameter.

Table 4.1: Posterior means, 95% highest posterior density intervals, and the half-life of shocks (based on the posterior means) for the autoregressive parameters Φ in the AR(1) process for $\{\beta_t\}_{t=1}^T$ in Equation (4.6) for the clickthrough (CTR) and conversion (CON) probabilities.

	CTR			CON				
	Mean	2.5th percentile	97.5th percentile	Half-life (in weeks)	Mean	2.5th percentile	97.5th percentile	Half-life (in weeks)
Generic	1.00	0.99	1.00	-	1.00	0.98	1.00	-
Brand only	1.00	0.99	1.00	-	1.00	0.98	1.00	-
Brand & series	1.00	0.99	1.01	-	1.00	0.99	1.00	-
Retailer	1.00	0.99	1.00	-	0.99	0.96	1.01	10.1
Microsoft	0.99	0.97	1.00	11.6	0.98	0.94	1.01	4.6
Toshiba	0.99	0.97	1.00	7.2	0.30	-0.83	1.00	0.1
HP	1.00	0.99	1.00	-	0.75	-0.54	1.01	0.3
Acer	0.97	0.92	1.00	3.2	0.52	-0.73	1.01	0.2
Asus	0.68	-0.11	1.01	0.3	0.43	-0.74	1.00	0.1
Apple	0.99	0.98	1.00	11.5	0.30	-0.82	1.00	0.1
Lenovo	0.99	0.98	1.00	14.8	0.35	-0.70	0.99	0.1
MSI	0.99	0.96	1.00	6.9	0.35	-0.84	1.00	0.1

ance are generally also persistent, with permanent shocks for HP ads and a half-life ranging from 0.3 weeks for Asus ads to 14.8 weeks for Lenovo ads. The effects of brand-level shocks on the conversion performance are more transitory; the half-life ranges from 0.1 weeks to 4.6 weeks.

To compare the relative performance of keywords, in Figure 4.4 we plot the time-varying parameter series including baseline levels for the different specificity groups (left) and brands (right). Clickthrough probabilities are highest for retailer-specific keywords, followed by generic, brand only, and brand & series keywords. Conversion probabilities are also highest for retailer-specific keywords followed by brand & series, generic and brand only keywords. The clickthrough probabilities of different brands are volatile, whereas the conversion probabilities are relatively stable. Conversion probabilities are lowest for Apple keywords, followed by Microsoft, MSI, and Toshiba keywords. All graphs again show quite some time variation.

Figure 4.4: Plots of the series $\{\beta_t\}_{t=1}^T + \bar{\alpha}$ for the specificity and brand series, for the clickthrough (CTR) and conversion (CON) probabilities.

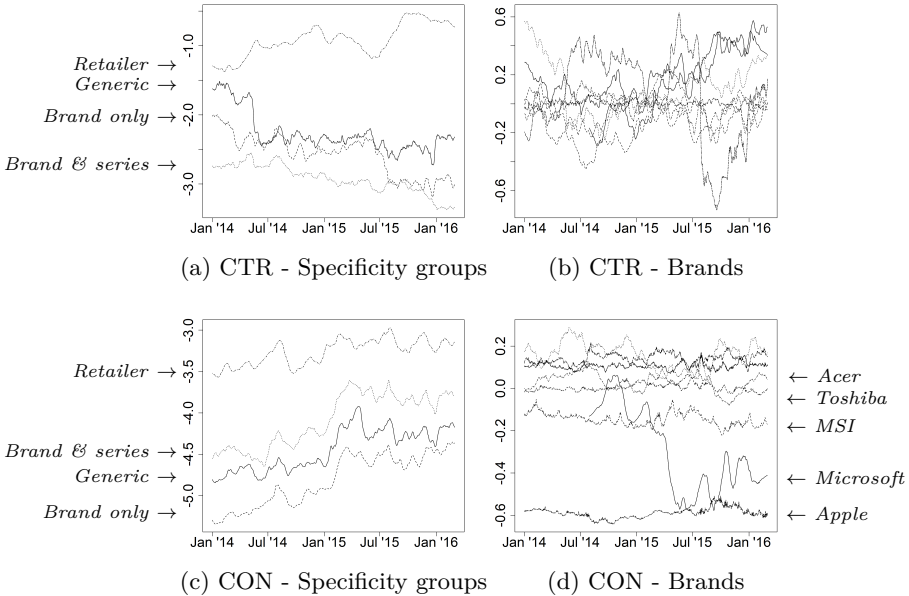


Table 4.2 displays day-of-the-week effects and the effects of keyword length and match type. Clickthrough probabilities are highest on Mondays to Wednesdays and for shorter and exact keywords. Conversion probabilities are lowest on Saturdays, and

highest for longer and exact keywords. The posterior estimates for the standard deviations σ_α show that there is substantial variation across keywords in the baseline level of clickthrough and conversion probabilities.

Table 4.2: Posterior means and standard deviations (in parentheses) for the seasonal effects (γ), the time-invariant parameters in $\tilde{\alpha}$ and σ_α (the square root of the diagonal of Σ_α).

	CTR	CON
Tuesday	0.00 (0.01)	0.01 (0.03)
Wednesday	-0.01 (0.01)	-0.03 (0.03)
Thursday	-0.02 (0.01)	-0.03 (0.03)
Friday	-0.03 (0.01)	-0.01 (0.04)
Saturday	-0.06 (0.01)	-0.05 (0.03)
Sunday	-0.06 (0.01)	-0.03 (0.03)
ln (# words)	-0.04 (0.02)	0.16 (0.04)
Exact match	0.48 (0.02)	0.25 (0.04)
σ_α	0.65 (0.01)	0.43 (0.02)

Next, we consider the effect of ad position. Table 4.3 displays the estimated effect of ad position (columns 2 and 3) and of bid/CPC (column 4). We find that the more prominent the ad — the lower the position number — the higher the clickthrough probability. This holds in general for all types of keywords, although the relationship is strongest for retailer-specific and exact keywords, and weakest for Apple, HP, and MSI keywords. For the conversion probabilities we do not find strong evidence that ad position affects conversion probabilities in general. Furthermore, high bids/CPCs are mostly associated with more prominent ads. This holds strongest for long and Microsoft keywords. The estimates for the standard deviations σ_λ show that the effect of position on clickthrough and conversion probabilities varies substantially over similar keywords.

Table 4.3, columns 5 and 6, show the posterior results for the instruments in the bid equation. The results indicate that the bid/CPC is associated with past performance. The instruments seem strong enough to identify the other parameters. In general, the advertiser sets higher bids on keywords that previously obtained a high number of impressions and a high clickthrough rate. Given the size of the variance across keywords (σ_δ), the reverse relationship also seems to hold for a number of keywords. This implies that the advertiser may not always bid strategically based on previous clickthrough rates and impressions obtained.

Table 4-3: Posterior means and standard deviations (in parentheses) for the position parameters in the clickthrough and conversion equations (columns 2-3), the bid parameters in the position equation ($\tilde{\lambda}$) (column 4), the instruments in the bid equation ($\tilde{\delta}$) (columns 5-6) and σ_λ and σ_δ (the square root of diagonal of Σ_λ and Σ_δ respectively). Base categories are specificity ‘Generic’, brand ‘Acer’ and match type ‘Broad’.

	Impact POS on		Impact BID/CPC on		Impact lagged CTR on		Impact lagged IMP on	
	CTR	CON	POS		BID	BID	BID	BID
Intercept	-1.01 (0.04)	0.07 (0.07)	-0.04 (0.02)		0.02 (0.00)		0.07 (0.01)	
Brand only	-0.15 (0.05)	0.06 (0.11)	0.05 (0.02)		-0.01 (0.01)		0.03 (0.02)	
Brand & series	0.03 (0.04)	-0.05 (0.10)	0.15 (0.02)		-0.01 (0.01)		-0.03 (0.01)	
Retailer	-0.46 (0.11)	-0.05 (0.15)	0.22 (0.05)		-0.01 (0.01)		-0.22 (0.03)	
Microsoft	0.08 (0.08)	0.29 (0.15)	-0.14 (0.03)		-0.02 (0.01)		0.00 (0.02)	
Toshiba	-0.02 (0.04)	0.02 (0.12)	-0.08 (0.02)		-0.02 (0.01)		0.04 (0.01)	
HP	0.12 (0.04)	-0.04 (0.10)	-0.01 (0.01)		0.00 (0.01)		0.04 (0.01)	
Asus	0.09 (0.04)	-0.16 (0.10)	-0.01 (0.01)		0.02 (0.01)		0.01 (0.01)	
Apple	0.21 (0.05)	0.07 (0.13)	-0.04 (0.02)		-0.01 (0.01)		-0.03 (0.02)	
Lenovo	0.00 (0.04)	-0.30 (0.11)	-0.02 (0.02)		0.03 (0.01)		0.05 (0.01)	
MSI	0.14 (0.06)	0.17 (0.17)	0.02 (0.02)		-0.03 (0.01)		0.01 (0.01)	
ln(# words)	-0.02 (0.03)	0.03 (0.06)	-0.07 (0.01)		0.00 (0.00)		0.00 (0.01)	
Exact match	-0.22 (0.02)	0.07 (0.05)	-0.05 (0.01)		-0.01 (0.00)		-0.01 (0.01)	
$\sigma_\lambda/\sigma_\delta$	0.43 (0.01)	0.18 (0.02)	0.28 (0.00)		0.05 (0.00)		0.13 (0.00)	

Table 4.4 shows the posterior mean of the covariance matrix of the error terms, Σ_η : the variances are displayed on the diagonal, the covariances on the upper diagonal, and the correlations on the lower diagonal. Position endogeneity seems present, as the unexplained parts of the clickthrough probabilities are positively correlated with the unexplained parts of ad position. For conversion probabilities we find no strong evidence of position endogeneity.

Table 4.4: Posterior means and standard deviations (in parentheses) for the variances (diagonal), covariances (upper diagonal) and correlations (lower diagonal) of η_{it}

	CTR	CON	POS	BID/CPC
CTR	0.137 0.002	-0.003 0.004	0.027 0.002	0.003 0.001
CON	-0.045 0.067	0.026 0.005	0.005 0.005	-0.003 0.003
POS	0.222 0.016	0.099 0.103	0.106 0.000	-0.030 0.001
BID	0.018 0.008	-0.048 0.049	-0.240 0.010	0.144 0.001

Finally, bidding endogeneity also seems present, as there is a negative correlation of -0.240 between the position and bid error terms. Part of this correlation can be explained because we use the CPC to proxy the bid; the ad position and CPC are both influenced by unobserved competitive behavior. These findings reinforce that it is important to account for these forms of endogeneity.

4.5.3 Model comparison

In this section, we compare the performance of the dynamic model to a static model with seasonality, that is, setting all $\beta_t = 0$. We compute log Bayes factors to evaluate the models' in-sample and out-of-sample performance. In case the log Bayes factor is greater than $\log(3)$ we have sufficient evidence to favor the null model (the static model), in case it is smaller than $\log(1/3)$ we have evidence to favor the alternative model (the dynamic model) (Kass & Raftery, 1995).

We compute the in-sample log Bayes factors with the Savage-Dickey density ratio, using the estimates obtained from the dynamic model only (Dickey, 1971).¹¹ We

¹¹The in-sample log Bayes factor of the static model against the dynamic model can be computed by the Savage-Dickey density ratio

$$\ln BF = \ln p(\Phi|y)|_{\Phi=O} - \ln p(\Phi)|_{\Phi=O}, \quad (4.10)$$

compute the predictive Bayes factors using predictions from both the dynamic and the static model for a full year. For these predictions we use a moving window of 26 weeks; after each window we make predictions for each day in the next week and then we move the window one week further. This also allows the parameters of the static model to change. We make predictions for each day in the period March 30, 2015 until March 27, 2016. For each model, we use 8.000 simulations after 2.000 burn-in draws and we keep each 4th draw.

Table 4.5: Log Bayes factors for the dynamic model (alternative model) against a static model with $\beta_t = 0$ (null model).

	CTR	CON
In-sample log Bayes factor	-474 364	-39 710
Predictive log Bayes factor	-6 790	-87

Table 4.5 shows the log Bayes factors separately for the clickthrough and conversion equation. The log Bayes factors are all highly negative and much smaller than $\log(1/3)$ ($= -0.48$). Hence, the dynamic model is superior to a static model in terms of both in- and out-of-sample performance. There is thus substantial evidence of dynamics in the clickthrough and conversion probabilities in the dataset, indicating that a dynamic SEA strategy is to be preferred over a static strategy.

Illustrative examples of the dynamic and static models' predictions are given in Figure 4.5. The clickthrough predictions are given in the top three figures, the conversion predictions in the bottom three figures. Overall, the predictions of the dynamic model (solid lines) are more volatile than those of the static model (dashed lines). For retailer-specific keywords we find that the dynamic model's clickthrough forecasts fluctuate around the static model's forecasts. Hence, where the dynamic model is able to capture short-term fluctuations, the static model with a moving window is not. For the conversion predictions we find that, in the period April 2015 until July 2015, the dynamic predictions are substantially higher than the static predictions. The reason

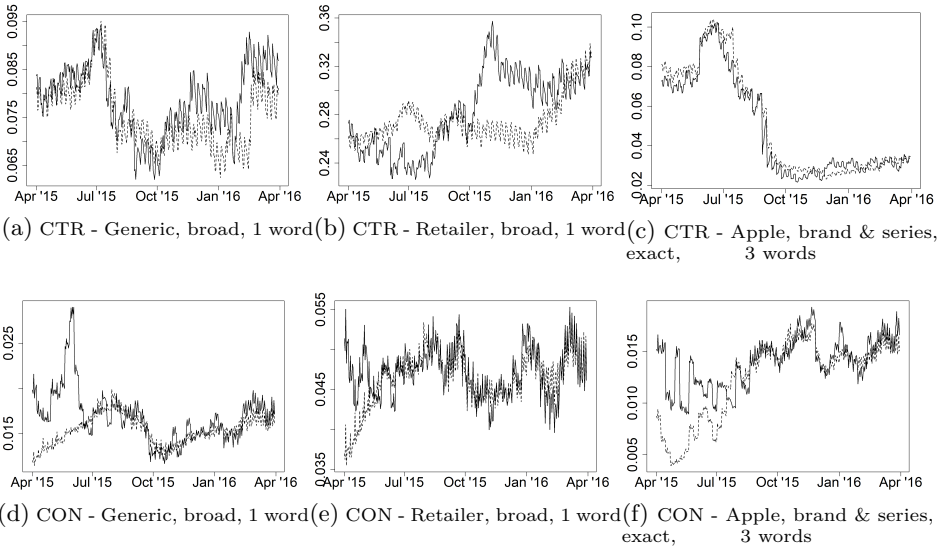
where $p(\Phi|y)$ denotes the posterior marginal pdf of Φ from the dynamic model, and $p(\Phi)$ denotes the prior pdf of Φ . We approximate the first term on the right hand side in Equation (4.10) using Rao-Blackwellization based on the full conditional posterior distribution of Φ (Gelfand & Smith, 1990). That is,

$$\ln p(\Phi|y)|_{\Phi=O} \approx \ln \left(\frac{1}{S} \sum_{s=1}^S p(\Phi|\beta^{(s)}, \Sigma_{\beta}^{(s)})|_{\Phi=O} \right), \quad (4.11)$$

where S is the number of Monte Carlo simulations, and $\beta^{(s)}$ and $\Sigma_{\beta}^{(s)}$ are the parameter draws at the s^{th} simulation.

is reflected in Figure 4.4, where we see a substantial increase in the conversion rates of all specificity groups in the period October 2014 until April 2015. The dynamic model timely captures this increase whereas the static model lags behind.

Figure 4.5: Posterior mean of predictions for clickthrough (a)-(c) and conversion (d)-(f) probabilities for three keywords for dynamic model (solid lines) and static model (dashed lines) using a moving window of 26 weeks. Ad position is set to 1.



4.6 Managerial implications

The managerial implications of this paper are threefold. First, advertisers can use the model to obtain accurate daily estimates of clickthrough and conversion probabilities of individual ads. These estimates can be used to set bids and test the performance of text ads and landing pages. These estimates can also be used to identify keywords of which the performance is divergent from similar keywords.

Second, advertisers can use the model to examine the extent of dynamics in their SEA environment. The more dynamic the environment and the higher the persistence of shocks, the more often the SEA strategy should be revised. Moreover, advertisers that manage large ad portfolios can prioritize their focus on keywords based on the expected influence and persistence of shocks on the keywords' performance.

Finally, advertisers can use the model to track the performance of ads to timely

identify when the performance of keywords changes. An advertiser can then analyze the causes of these changes and adjust, for example, the text ad, landing page, product pricing, or bid accordingly.

As a final remark, the model's predictions of clickthrough and conversion probabilities are insufficient to determine the optimal bid per keyword. To set the optimal bid, an advertiser has to know the value of each obtained impression, click and conversion, using additional information on spillover effects to future searches (see Rutz and Bucklin, 2011, Rutz et al., 2011, Agarwal et al., 2011), substitution effects across marketing channels (see S. Yang and Ghose, 2010, Dinner et al., 2014, and Blake et al., 2015), and branding profits of keywords (see Ghose and Yang, 2008). Combining the information from these sources requires the formation of an attribution strategy like in Li and Kannan (2014). An alternative is to use a bidding heuristic as given in Skiera and Abou Nabout (2013). We therefore consider the determination of the optimal bid to be outside the scope of this paper.

4.7 Summary and conclusions

In this article, we propose a dynamic Bayesian model for clickthrough and conversion probabilities of paid search advertisements. Clickthrough and conversion probabilities can be subject to changes over time, due to, for example, changes in the tastes and attitudes of consumers or the launch of a new product. Gaining insight into the dynamics of ad performance is crucial for advertisers to develop effective search engine advertising strategies.

Our main contribution is the development of a model that is especially suited to deal with dynamic SEA environments: the model allows for time-varying parameters, seasonal effects, data sparsity, missing data, position endogeneity and unobserved cross-sectional heterogeneity. Moreover, we propose AR(1) processes for the time-varying parameters, thereby allowing for shocks on different types of ads (e.g. brand-specific versus generic ads) to have different dynamic effects on ad performance (e.g. permanent versus transitory).

In the empirical application, we find evidence of substantial persistent time variation in ad performance, emphasizing the importance of addressing dynamics in SEA ad performance models. We also find evidence of position and bidding endogeneity, indicating that purely predictive models are unable to capture causal relationships between ad position and clickthrough and conversion probabilities.

We note several limitations of this study. First, a drawback of the proposed method is the large computation time involved. Especially drawing the auxiliary Polya-Gamma variables is time-consuming, due to the large number of impressions and clicks in the dataset. Second, in the empirical application we use the cost-per-click (CPC) to proxy the bid. The actual bid may contain more information on the resulting ad position than the CPC. Finally, data could be missing not at random. For example, an advertiser might not bid on keywords that are expected to perform poorly. In this case, the results of the model might not hold for the non-selected keywords.

We note two interesting ways in which this study can be extended. First, one can add correlation across the time-varying parameters of the clickthrough and conversion equations. Such correlations can capture the idea that some shocks affect both clickthrough and conversion probabilities. On the downside, allowing for these correlations will substantially increase computation time as the time-varying parameter series for the two equations then need to be drawn jointly. Finally, one can use latent factors to explain ad performance instead of pre-specified keyword characteristics. Such an analysis will aid understanding of which factors drive the difference in ad performance across keywords and will help advertisers in designing effective ad campaigns. Again, this will substantially increase computation time.

Appendix

4.A Gibbs sampler

To obtain posterior results of the dynamic Bayesian model, we use a Gibbs sampler with Polya-Gamma data augmentation (Geman and Geman, 1987, Tanner and Wong, 1987, Polson et al., 2013). The Polya-Gamma data augmentation scheme is suitable for binomial likelihoods (Polson et al., 2013). This scheme involves introducing one layer of auxiliary latent variables that follows a Polya-Gamma distribution. Conditional on these latent variables, the posterior distribution of the parameters of interest has the same functional form as the posterior distribution of parameters from a linear regression model with normally distributed error terms. This approach is similar to the data augmentation scheme for probit models of Albert and Chib (1993), but requires less memory storage as latent variables are drawn for each observation (keyword times day) instead of for each impression or click. In a SEA application this is crucial as the number of daily impressions and clicks can be very large.

The Polya-Gamma data augmentation scheme works as follows. Suppose that we have a binomially distributed variable $y \sim \text{BIN}(N, 1/(1 + \exp(-\theta)))$. Introduce an auxiliary random variable ω that follows the Polya-Gamma distribution $\text{PG}(N, 0)$. The likelihood function $p(y|\theta)$ can be written as

$$p(y|\theta) = \left(\frac{1}{1 + \exp(-\theta)} \right)^y \left(\frac{1}{1 + \exp(\theta)} \right)^{N-y} = \int p(y|\theta, \omega) p(\omega) d\omega,$$

where Polson et al. (2013) have showed that the conditional distribution $p(y|\theta, \omega)$ is proportional to the likelihood kernel of a linear regression model

$$p(y|\theta, \omega) \propto \exp \left\{ -\frac{\omega}{2} (z - \theta)^2 \right\},$$

with pseudo-observations $z \equiv (y - N/2)/\omega$, signal θ , and independently distributed error terms with variances $1/\omega$. Thus, the full conditional distribution of θ , $p(\theta|\omega, y) \propto p(y|\theta, \omega)p(\theta)$, becomes standard. That is, the full conditional distribution of θ is the same as if we have the linear regression model $z = \theta + \nu$, $\nu \sim N(0, 1/\omega)$ with prior $p(\theta)$. Moreover, the full conditional distribution $p(\omega|\theta, y)$ is also a Polya-Gamma distribution, and ω can thus be easily sampled along in the Gibbs sampler (Polson et al., 2013).

For our dynamic model in Equations (4.1)-(4.4), we have that conditional on the auxiliary latent Polya-Gamma distributed variables for the clickthrough and conversion equations (denoted by ω_{it}^{CTR} and ω_{it}^{CON} , respectively) we have the multivariate linear regression model

$$z_{it}^{CTR} = \alpha_i^{CTR} + x'_i \beta_t^{CTR} + \lambda_i^{CTR} \ln(\text{POS}_{it}) + s'_t \gamma^{CTR} + \eta_{it}^{CTR} + \xi_{it}^{CTR}, \quad (4.12)$$

$$z_{it}^{CON} = \alpha_i^{CON} + x'_i \beta_t^{CON} + \lambda_i^{CON} \ln(\text{POS}_{it}) + s'_t \gamma^{CON} + \eta_{it}^{CON} + \xi_{it}^{CON}, \quad (4.13)$$

$$\ln(\text{POS}_{it}) = \alpha_i^{POS} + x'_i \beta_t^{POS} + \lambda_i^{POS} \ln(\text{BID}_{it}) + \psi^{POS} \ln(\text{QS}_{it}) + s'_t \gamma^{POS} + \eta_{it}^{POS}, \quad (4.14)$$

$$\ln(\text{BID}_{it}) = \alpha_i^{BID} + x'_i \beta_t^{BID} + q'_{it} \delta_i^{BID} + s'_t \gamma^{BID} + \eta_{it}^{BID}, \quad (4.15)$$

with $z_{it}^{CTR} \equiv (N_{it} - I_{it}/2)/\omega_{it}^{CTR}$, $z_{it}^{CON} \equiv (M_{it} - N_{it}/2)/\omega_{it}^{CON}$, $\xi_{it}^{CTR} \sim N(0, 1/\omega_{it}^{CTR})$, and $\xi_{it}^{CON} \sim N(0, 1/\omega_{it}^{CON})$. The equations are related through $\eta_{it} \sim \text{MVN}(0, \Sigma_\eta)$.

For ease of representation, we replace the names CTR, CON, POS and BID by the numbers 1 to 4, respectively, and rename the variables and parameters to obtain the

specific blocks for the Gibbs sampler. That is, we rewrite Equations (4.12)-(4.15) as

$$z_{1it} = w'_{1it}\pi_{1i} + x'_i\beta_{1t} + s'_{1it}\gamma_1 + \eta_{1it} + \xi_{1it}, \quad (4.16)$$

$$z_{2it} = w'_{2it}\pi_{2i} + x'_i\beta_{2t} + s'_{2it}\gamma_2 + \eta_{2it} + \xi_{2it}, \quad (4.17)$$

$$z_{3it} = w'_{3it}\pi_{3i} + x'_i\beta_{3t} + s'_{3it}\gamma_3 + \eta_{3it}, \quad (4.18)$$

$$z_{4it} = w'_{4it}\pi_{4i} + x'_i\beta_{4t} + s'_{4it}\gamma_4 + \eta_{4it}, \quad (4.19)$$

where

$$\begin{bmatrix} z_{1it} \\ z_{2it} \\ z_{3it} \\ z_{4it} \end{bmatrix} = \begin{bmatrix} z_{it}^{CTR} \\ z_{it}^{CON} \\ \ln(\text{POS}_{it}) \\ \ln(\text{BID}_{it}) \end{bmatrix}, \quad \begin{bmatrix} w'_{1it} \\ w'_{2it} \\ w'_{3it} \\ w'_{4it} \end{bmatrix} = \begin{bmatrix} 1 & \ln(\text{POS}_{it}) \\ 1 & \ln(\text{POS}_{it}) \\ 1 & \ln(\text{BID}_{it}) \\ 1 & q'_{it} \end{bmatrix}, \quad \begin{bmatrix} \pi'_{1i} \\ \pi'_{2i} \\ \pi'_{3i} \\ \pi'_{4i} \end{bmatrix} = \begin{bmatrix} \alpha_i^{CTR} & \lambda_i^{CTR} \\ \alpha_i^{CON} & \lambda_i^{CON} \\ \alpha_i^{POS} & \lambda_i^{POS} \\ \alpha_i^{BID} & \delta_i^{BID} \end{bmatrix},$$

$$\begin{bmatrix} \beta_{1t} \\ \beta_{2t} \\ \beta_{3t} \\ \beta_{4t} \end{bmatrix} = \begin{bmatrix} \beta_t^{CTR} \\ \beta_t^{CON} \\ \beta_t^{POS} \\ \beta_t^{BID} \end{bmatrix}, \quad \begin{bmatrix} s'_{1it} \\ s'_{2it} \\ s'_{3it} \\ s'_{4it} \end{bmatrix} = \begin{bmatrix} s'_t \\ s'_t \\ s'_t \\ s'_t \end{bmatrix} \ln(QS_{it}), \quad \begin{bmatrix} \gamma'_1 \\ \gamma'_2 \\ \gamma'_3 \\ \gamma'_4 \end{bmatrix} = \begin{bmatrix} \gamma'^{CTR} \\ \gamma'^{CON} \\ \gamma'^{POS} \\ \gamma'^{BID} \end{bmatrix} \psi^{POS}, \quad \begin{bmatrix} \eta_{1it} \\ \eta_{2it} \\ \eta_{3it} \\ \eta_{4it} \end{bmatrix} = \begin{bmatrix} \eta_{it}^{CTR} \\ \eta_{it}^{CON} \\ \eta_{it}^{POS} \\ \eta_{it}^{BID} \end{bmatrix},$$

and $(\xi_{1it}, \xi_{2it})' = (\xi_{it}^{CTR}, \xi_{it}^{CON})'$. We also rewrite Equations (4.6)-(4.9) in terms of j :

$$\begin{aligned} \beta_{j,t+1} &= \Phi_j \beta_{jt} + \nu_{jt}, & \nu_{jt} &\sim MVN(0, \Sigma_{\beta,j}), & \beta_{j1} &\sim MVN(0, 5\Sigma_{\beta,j}), & \text{for } j &= 1, \dots, 4, \\ \alpha_{ji} &\sim N(x'_i \tilde{\alpha}_j, \sigma_{\alpha,j}^2), & & & & & \text{for } j &= 1, \dots, 4, \\ \lambda_{ji} &\sim N(x'_i \tilde{\lambda}_j, \sigma_{\lambda,j}^2), & & & & & \text{for } j &= 1, \dots, 3, \\ \delta_{ji} &\sim MVN(x'_i \tilde{\delta}_j, \Sigma_{\delta,j}), & & & & & \text{for } j &= 4. \end{aligned}$$

For computational efficiency, in the Gibbs sampler we draw the parameters of each of the four model equations *separately* by conditioning on the η_{it} of the other equations. This also helps deal with missing values. For this, we compute

$$\begin{aligned} \bar{\eta}_{jit} &\equiv E[\eta_{jit} | \eta_{-j,it}] = \Sigma_{\eta(j,-j)} \Sigma_{\eta(-j,-j)}^{-1} \eta_{-j,it}, \\ \bar{\sigma}_{\eta,j}^2 &\equiv \text{Var}(\eta_{jit} | \eta_{-j,it}) = \Sigma_{\eta(j,j)} - \Sigma_{\eta(j,-j)} \Sigma_{\eta(-j,-j)}^{-1} \Sigma_{\eta(-j,j)}, \end{aligned}$$

for $j = 1, \dots, 4$. Here we denote by $\eta_{-j,it}$ all elements in η_{it} except for the j^{th} element and any missing elements, and by $\Sigma_{\eta(j,-j)}$ all elements of Σ_{η} related to row j and all columns except for the j^{th} . Then, we can rewrite each of the Equations (4.16)-(4.19) as a univariate regression model conditional on $\eta_{-j,it}$, as given by

$$z_{jit} = w'_{jit}\pi_{ji} + x'_i\beta_{jt} + s'_{jit}\gamma_j + \bar{\eta}_{jit} + \zeta_{jit}, \quad (4.20)$$

for $j = 1, \dots, 4$, where the introduction of $\bar{\eta}_{jit}$ ensures that the error terms ζ_{jit} are independent of each other

$$\begin{pmatrix} \zeta_{1it} \\ \zeta_{2it} \\ \zeta_{3it} \\ \zeta_{4it} \end{pmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\zeta,1it}^2 \equiv 1/\omega_{1it} + \bar{\sigma}_{\eta,1}^2 & 0 & 0 & 0 \\ 0 & \sigma_{\zeta,2it}^2 \equiv 1/\omega_{2it} + \bar{\sigma}_{\eta,2}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\zeta,3it}^2 \equiv \bar{\sigma}_{\eta,3}^2 & 0 \\ 0 & 0 & 0 & \sigma_{\zeta,4it}^2 \equiv \bar{\sigma}_{\eta,4}^2 \end{bmatrix} \right).$$

Hence, after drawing the parameters of a single equation j , we update the η_j of that equation, and condition on the new η_j in drawing the parameters of the next equations.

4.A.1 Overview Gibbs sampler

We use the rewritten model in Equation (4.20) to construct the Gibbs sampler. To improve the mixing rates of the sampler, we (i) jointly sample η_{jit} and the parameters in $(\alpha_{ji}, \beta_{jt}, \lambda_{ji}, \delta_{ji}, \gamma_j)$, and (ii) jointly sample η and Σ_η . Because of sampling jointly, we need to draw η twice within a single Gibbs step.

The Gibbs steps are given by

1. For $j = 1, \dots, 4$ do
 - i. Sample $\omega_{jit} | I_{it}, N_{it}, M_{it}, \pi_{ji}, \beta_{jt}, \gamma_j, \eta_{jit}$ (if $j \in \{1, 2\}$, for $i = 1, \dots, N$, $t = 1, \dots, T$).
 - ii. Compute $z_{jit} | \omega_{jit}, I_{it}, N_{it}, M_{it}$ (if $j \in \{1, 2\}$, for $i = 1, \dots, N$, $t = 1, \dots, T$).
 - iii. Compute $\bar{\eta}_{jit}, \bar{\sigma}_{\eta,j}^2 | \eta_{-j,it}, \Sigma_\eta$ (for $i = 1, \dots, N$ and $t = 1, \dots, T$).
 - iv. Sample $\pi_{ji} = (\alpha_{ji}, \lambda_{ji}, \delta_{ji}) | z_{ji}, \omega_{ji}, \beta_j, \gamma_j, \tilde{\alpha}_j, \tilde{\lambda}_j, \tilde{\delta}_j, \bar{\eta}_{ji}, \sigma_{\alpha,j}^2, \sigma_{\lambda,j}^2, \Sigma_{\delta,j}, \bar{\sigma}_{\eta,j}^2$ (for $i = 1, \dots, N$).
 - v. Sample $\{\beta_{jt}\}_{t=1}^T | z_j, \omega_j, \pi_j, \gamma_j, \bar{\eta}_j, \Sigma_{\beta,j}, \Phi_j, \bar{\sigma}_{\eta,j}^2$.
 - vi. Sample $\gamma_j | z_j, \omega_j, \pi_j, \beta_j, \bar{\eta}_j, \bar{\sigma}_{\eta,j}^2$.
 - vii. Sample η_{jit} (for $i = 1, \dots, N$ and $t = 1, \dots, T$):
 - a. If $j \in \{1, 2\}$, sample $\eta_{jit} | z_{jit}, \omega_{jit}, \pi_{ji}, \beta_{jt}, \gamma_j, \eta_{-j,it}, \Sigma_\eta$.
 - b. If $j \in \{3, 4\}$, compute $\eta_{jit} | z_{jit}, \pi_{ji}, \beta_{jt}, \gamma_j$.
2. Sample $\tilde{\alpha} | \alpha, \Sigma_\alpha$, sample $\tilde{\lambda} | \lambda, \Sigma_\lambda$, and sample $\tilde{\delta} | \delta, \Sigma_\delta$.
3. Sample $\Sigma_\alpha | \alpha, \tilde{\alpha}$, sample $\Sigma_\lambda | \lambda, \tilde{\lambda}$, and sample $\Sigma_\delta | \delta, \tilde{\delta}$.

4. Sample $\Phi | \beta, \Sigma_\beta$.
5. Sample $\Sigma_\beta | \beta, \Phi$.
6. Sample $\Sigma_\eta | z, \omega, \pi, \beta, \gamma$.
7. Compute $\bar{\eta}_{jit}, \bar{\sigma}_{\eta,j}^2 | \eta_{-j,it}, \Sigma_\eta$ (for $j = 1, \dots, 4, i = 1, \dots, N$ and $t = 1, \dots, T$).
8. Sample $\eta_{jit} | z_{jit}, \omega_{jit}, \pi_{ji}, \beta_{jt}, \gamma_j, \eta_{-j,it}, \Sigma_\eta$ (for $j = 1, 2, i = 1, \dots, N$ and $t = 1, \dots, T$).

4.A.2 Priors

We choose conjugate priors to ensure that the model parameters can be drawn using Gibbs steps. For the logit equations (clickthrough and conversion) we take slightly informative priors, for the normal equations (position and bid) we take rather uninformative priors. First, for the means of the keyword-specific parameters ($\tilde{\alpha}$, $\tilde{\lambda}$, and $\tilde{\delta}$) we take multivariate normal prior distributions with mean 0 and covariance matrix I (for clickthrough and conversion equations) or $100I$ (for position and bid equations). Second, for the variances of the keyword-specific parameters (diagonal elements of Σ_α , Σ_λ , and Σ_δ) we take inverse Gamma-2 prior distributions with shape parameter $\kappa_0 = 5$ and scale parameter $\kappa_1 = 5 \times 0.1$.

Third, for the time-varying parameters we take a multivariate normal prior for Φ with mean $\hat{\Phi}_0 = 0.5\iota$ and covariance matrix $\Sigma_{\Phi_0} = 0.5I$, where ι represents a vector of ones. For the diagonal elements of Σ_β we take inverse Gamma-2 prior distributions with shape parameter $\kappa_{\beta,0} = 5$ and scale parameter $\kappa_{\beta,1} = 5 \times 0.001$. Fourth, for the time-invariant parameters in (γ, ψ^{POS}) we take a multivariate normal prior distribution with mean 0 and covariance matrix I (for clickthrough and conversion equations) or $100I$ (for position and bid equations).

Finally, for the covariance matrix Σ_η we take an inverse Wishart prior with 8 degrees of freedom and scale matrix $8 \times 0.1I$.

4.A.3 Initialization

We take the following initialization. For the baseline level, we set $\Sigma_\alpha = 0.1I$ and $\tilde{\alpha} = 0$ except for the intercept in $\tilde{\alpha}$ which we set to -3 for the clickthrough and conversion equations, to 1 for the position equation, and to -1 for the bid equation. For the time-varying parameters, we set $\{\beta_t\}_{t=1}^T = 0$, $\Phi = 0.5I$, and $\Sigma_\beta = 0.001I$. Furthermore, we initialize $\gamma = 0$, $\psi^{POS} = 0$, $\tilde{\lambda} = 0$, and $\Sigma_\lambda = 0.1I$. For the

instruments, we set $\tilde{\delta} = 0$ and $\Sigma_\delta = 0.1I$. Finally, for the keyword- and time-specific shocks, we initialize $\Sigma_\eta = 0.1I$, η_{it}^{CTR} and η_{it}^{CON} to 0 for all i and t and compute η_{it}^{POS} and η_{it}^{PC} based on the other initializations.

4.A.4 Steps Gibbs sampler

4.A.4.1 Sampling Polya-Gamma variables ω

The full conditional posterior distribution of the auxiliary latent Polya-Gamma variables ω_{1it} (ω_{2it}) are independent Polya-Gamma distributions with parameters I_{it} (N_{it}) and θ_{1it} (θ_{2it}) for $i = 1, \dots, N$ and $t = 1, \dots, T$ where I_{it} (N_{it}) denotes the number impressions (clicks), and

$$\theta_{jit} = w'_{jit}\pi_{ji} + x'_i\beta_{jt} + s'_t\gamma_j + \eta_{jit},$$

for $j = 1, 2$. We draw the Polya-Gamma variables using the R package BayesLogit (Windle, Polson et al., 2014). For computational efficiency, we approximate the Polya-Gamma variable ω_{1it} (ω_{2it}) by normal variables in case $I_{it} > 170$ ($N_{it} > 170$) (Windle, Polson et al., 2014). In this approximation, we set the first and second moment of the normal distribution equal to the first and second moment of the associated Polya-Gamma distribution.

After drawing the Polya-Gamma variables, we compute the pseudo data points for the clickthrough and conversion equations

$$\begin{aligned} z_{1it} &= (N_{it} - I_{it}/2) / \omega_{1it}, \\ z_{2it} &= (M_{it} - N_{it}/2) / \omega_{2it}. \end{aligned}$$

4.A.4.2 Sampling α_i , λ_i , and δ_i

To sample α_{ji} , λ_{ji} , and δ_{ji} (collected in π_{ji}) for $i = 1, \dots, N$, note that we can write Equation (4.20) as the univariate normal regression model

$$y_{\pi,jit} \equiv z_{jit} - x'_i\beta_{jt} - s'_{jit}\gamma_j - \bar{\eta}_{jit} = w'_{jit}\pi_{ji} + \zeta_{jit}, \quad \zeta_{jit} \sim N(0, \sigma_{\zeta,jit}^2),$$

for $t = 1, \dots, T$, with a normal prior for $\pi_{ji} \sim MVN(\bar{\pi}_{j0}, \Sigma_{\pi_{j0}})$ where

$$\bar{\pi}_{j0} = \begin{bmatrix} x'_i\tilde{\alpha}_j \\ x'_i\tilde{\lambda}_j \end{bmatrix}, \quad \Sigma_{\pi_{j0}} = \begin{bmatrix} \sigma_{\alpha,j}^2 & 0 \\ 0 & \sigma_{\lambda,j}^2 \end{bmatrix}.$$

When $j = 4$, the elements $\tilde{\lambda}_j$ and $\sigma_{\lambda,j}^2$ are replaced by $\tilde{\delta}_j$ and $\Sigma_{\delta,j}$.

We draw π_{ji} from $MVN(\hat{\pi}_{ji}, \hat{\Sigma}_{\pi_{ji}})$

$$\begin{aligned}\hat{\Sigma}_{\pi_{ji}} &= \left(\sum_{t=1}^T w_{jit} w'_{jit} / \sigma_{\zeta,jit}^2 + \Sigma_{\pi_{j0}}^{-1} \right)^{-1}, \\ \hat{\pi}_{ji} &= \hat{\Sigma}_{\pi_{ji}} \left(\sum_{t=1}^T w_{jit} y_{\pi,jit} / \sigma_{\zeta,jit}^2 + \Sigma_{\pi_{j0}}^{-1} \bar{\pi}_0 \right),\end{aligned}$$

for $i = 1, \dots, N$.

4.A.4.3 Sampling β_t

To sample $\{\beta_{jt}\}_{t=1}^T$, note that we can write Equation (4.20) as the univariate normal regression model

$$y_{\beta,jit} \equiv z_{jit} - w'_{jit}\pi_{ji} - s'_{jit}\gamma_j - \bar{\eta}_{jit} = x'_i\beta_{jt} + \zeta_{jit}, \quad \zeta_{jit} \sim N(0, \sigma_{\zeta,jit}^2),$$

with

$$\beta_{j,t+1} = \Phi_j\beta_{jt} + \nu_{jt}, \quad \nu_{jt} \sim MVN(0, \Sigma_{\beta,j}), \quad \beta_{j1} \sim MVN(0, 5\Sigma_{\beta,j}).$$

We sample $\{\beta_{jt}\}_{t=1}^T$ using the simulation smoother of Durbin and Koopman (2002) (as explained in Durbin and Koopman (2012), section 4.9.2). To speed up computations, we perform collapsed filtering (Durbin and Koopman 2012, Chapter 6.5, Jungbacker and Koopman 2015).

Collapsed filtering works as follows. We have the $(N \times 1)$ vector $y_{\beta,jt}$ and the $(N \times K)$ matrix X , with $K \gg N$. We can compute a $(K \times N)$ matrix A_{jt}^* such that we can obtain the correct smoothed estimates by using the observation equation with $(K \times 1)$ observation vector

$$A_{jt}^* y_{\beta,jt} = A_{jt}^* X + A_{jt}^* \zeta_{jt},$$

where the covariance matrix of ζ_{jt} is a diagonal matrix with elements $\sigma_{\zeta,jit}^2$. Hence, this procedure allows for much lower computation times because the altered observation vector is of much smaller dimension than the original observation vector, while the covariance matrix remains diagonal. We take the i^{th} column of A_{jt}^* equal to

$$A_{jit}^* = \begin{cases} \left(\sum_{n: I_{nt} \geq 1} \frac{1}{\sigma_{\zeta,jnt}^2} x_n x'_n \right)^{-1/2} x_i / \sigma_{\zeta,jit}^2, & \text{if } j = 1, 3, \\ \left(\sum_{n: N_{nt} \geq 1} \frac{1}{\sigma_{\zeta,jnt}^2} x_n x'_n \right)^{-1/2} x_i / \sigma_{\zeta,jit}^2, & \text{if } j = 2, 4, \end{cases}$$

where for the matrix in the first terms on the right hand sides we first take the Cholesky decomposition (upper triangular) and then the inverse. This A^* is chosen because then $A_{jt}^* \zeta_{jt} \sim MVN(0, I)$.

In case X is not of full column rank (see Section 4.4.2) the above procedure needs to be slightly altered. That is, let \tilde{X} be the $(N \times K_2)$ matrix with columns of X such that \tilde{X} is of full column rank and has the same column space as X . Then, we compute A_{jt}^* using the rows of \tilde{X} instead of X .

Finally, we have to deal with missing values, which in the collapsed case refers to time periods in which there is a variable k^* in x_i which has the same value over all i . In other words, that time period has no keywords with impressions/clicks that have a specific characteristic in x_i . When such missings occur, we set element k^* in $A_{jt}^* y_{\beta, jt}$ to zero, and the elements in the $k^{*(th)}$ row and column of $A_{jt}^* X$ equal to zero.

4.A.4.4 Sampling γ

To sample γ_j note that we can write Equation (4.20) as the univariate normal regression model

$$y_{\gamma, j, it} \equiv z_{jit} - w'_{jit} \pi_{ji} - x'_i \beta_{jt} - \bar{\eta}_{jit} = s'_{jit} \gamma_j + \zeta_{jit}, \quad \zeta_{jit} \sim N(0, \sigma_{\zeta, jit}^2).$$

We draw γ_j from $MVN(\hat{\gamma}_j, \hat{\Sigma}_{\gamma_j})$ where

$$\begin{aligned} \hat{\Sigma}_{\gamma_j} &= \left(\sum_{i=1}^N \sum_{t=1}^T s_{jit} s'_{jit} / \sigma_{\zeta, it}^2 + \Sigma_{\gamma_{j0}}^{-1} \right)^{-1}, \\ \hat{\gamma}_j &= \hat{\Sigma}_{\gamma_j} \left(\sum_{i=1}^N \sum_{t=1}^T s_{jit} y_{\gamma, j, it} / \sigma_{\zeta, it}^2 \right), \end{aligned}$$

where $\Sigma_{\gamma_{j0}}$ is the diagonal covariance matrix of the normal prior for γ .

4.A.4.5 Sampling η

Next we sample $\{\{\eta_{jit}\}_{i=1}^N\}_{t=1}^T$. In case $j \in \{3, 4\}$ (position and bid equations), we see from Equations (4.18) and (4.19) that we can directly compute η_{jit} :

$$\eta_{jit} = z_{jit} - w'_{jit} \pi_{ji} - x'_i \beta_{jt} - s'_{jit} \gamma_j. \quad (4.21)$$

In case $j \in \{1, 2\}$ (clickthrough and conversion equations), we sample η_{jit} . Note that we can write both Equations (4.16) and (4.17) as the univariate normal regression

model

$$y_{\eta,jit} \equiv z_{jit} - w'_{jit}\pi_{ji} - x'_i\beta_{jt} - s'_{jit}\gamma_j = \eta_{jit} + \xi_{jit}, \quad \xi_{jit} \sim N(0, 1/\omega_{jit}),$$

with a normal prior $\eta_{jit} \sim N(\bar{\eta}_{jit}, \bar{\sigma}_{\eta,j}^2)$. We draw η_{jit} for $i = 1, \dots, N$ and $t = 1, \dots, T$ from $N(\hat{\eta}_{jit}, \hat{\Sigma}_{\eta_{jit}})$ where

$$\hat{\Sigma}_{\eta_{jit}} = (\omega_{jit} + 1/\bar{\sigma}_{\eta,j}^2)^{-1}, \quad \hat{\eta}_{jit} = \hat{\Sigma}_{\eta_{jit}} (\omega_{jit}y_{\eta,jit} + \bar{\eta}_{jit}/\bar{\sigma}_{\eta,j}^2).$$

4.A.4.6 Sampling $\tilde{\alpha}$, $\tilde{\lambda}$, and $\tilde{\delta}$

To sample $\tilde{\alpha}$, note that Equation (4.7) is a multivariate regression model given $\{\alpha_i\}_{i=1}^N$ and Σ_α . We draw $\tilde{\alpha}$ from $MVN(\hat{\tilde{\alpha}}, \hat{\Sigma}_{\tilde{\alpha}})$ where

$$\begin{aligned} \hat{\Sigma}_{\tilde{\alpha}} &= \left(\sum_{i=1}^N (I_4 \otimes x'_i)' \Sigma_\alpha^{-1} (I_4 \otimes x_i) + \Sigma_{\tilde{\alpha}_0}^{-1} \right)^{-1}, \\ \hat{\tilde{\alpha}} &= \hat{\Sigma}_{\tilde{\alpha}} \left(\sum_{i=1}^N (I_4 \otimes x'_i)' \Sigma_\alpha^{-1} \alpha_i \right), \end{aligned}$$

where $\Sigma_{\tilde{\alpha}_0}$ is the covariance matrix of the normal prior for $\tilde{\alpha}$, and \otimes denotes the Kronecker product.

To sample $\tilde{\lambda}$, note that Equation (4.8) is a multivariate regression model given $\{\lambda_i\}_{i=1}^N$ and Σ_λ . We draw $\tilde{\lambda}$ from $MVN(\hat{\tilde{\lambda}}, \hat{\Sigma}_{\tilde{\lambda}})$ where

$$\begin{aligned} \hat{\Sigma}_{\tilde{\lambda}} &= \left(\sum_{i=1}^N (I_3 \otimes x'_i)' \Sigma_\lambda^{-1} (I_3 \otimes x_i) + \Sigma_{\tilde{\lambda}_0}^{-1} \right)^{-1}, \\ \hat{\tilde{\lambda}} &= \hat{\Sigma}_{\tilde{\lambda}} \left(\sum_{i=1}^N (I_3 \otimes x'_i)' \Sigma_\lambda^{-1} \lambda_i \right), \end{aligned}$$

where $\Sigma_{\tilde{\lambda}_0}$ is the covariance matrix of the normal prior for $\tilde{\lambda}$.

To sample $\tilde{\delta}$, note that Equation (4.9) is a multivariate regression model given $\{\delta_i\}_{i=1}^N$ and Σ_δ . We draw $\tilde{\delta}$ from $MVN(\hat{\tilde{\delta}}, \hat{\Sigma}_{\tilde{\delta}})$ where

$$\begin{aligned} \hat{\Sigma}_{\tilde{\delta}} &= \left(\sum_{i=1}^N \Sigma_\delta^{-1} + \Sigma_{\tilde{\delta}_0}^{-1} \right)^{-1}, \\ \hat{\tilde{\delta}} &= \hat{\Sigma}_{\tilde{\delta}} \left(\sum_{i=1}^N \Sigma_\delta^{-1} \delta_i \right), \end{aligned}$$

where $\Sigma_{\tilde{\delta}_0}$ is the covariance matrix of the normal prior for $\tilde{\delta}$.

4.A.4.7 Sampling Σ_α , Σ_λ , and Σ_δ

To sample Σ_α , Σ_λ , and $\Sigma_{\delta,j}$, note that these covariance matrices are diagonal. Therefore, we separately draw each diagonal element.

To sample $\sigma_{\alpha,j}^2$, note we have a univariate regression model for $\alpha_{ji} \sim N(x'_i \tilde{\alpha}_j, \sigma_{\alpha,j}^2)$ for $i = 1, \dots, N$. We therefore draw $\sigma_{\alpha,j}^2$ from the inverse Gamma distribution

$$\sigma_{\alpha,j}^2 \sim IG \left(\frac{\sum_{i=1}^N (\alpha_{ji} - x'_i \tilde{\alpha}_j)^2 + \kappa_1}{2}, \frac{N + \kappa_0}{2} \right),$$

for $j = 1, \dots, 4$, with prior parameters κ_0 and κ_1 .

To sample $\sigma_{\lambda,j}^2$, note we have a univariate regression model for $\lambda_{ji} \sim N(x'_i \tilde{\lambda}_j, \sigma_{\lambda,j}^2)$ for $i = 1, \dots, N$. We therefore draw $\sigma_{\lambda,j}^2$ from the inverse Gamma distribution

$$\sigma_{\lambda,j}^2 \sim IG \left(\frac{\sum_{i=1}^N (\lambda_{ji} - x'_i \tilde{\lambda}_j)^2 + \kappa_1}{2}, \frac{N + \kappa_0}{2} \right),$$

for $j = 1, \dots, 3$, with prior parameters κ_0 and κ_1 .

To sample $\Sigma_{\delta,j,kk}$, for $j = 4$, note we have a univariate regression model for $\delta_{ki} \sim N(x'_i \tilde{\delta}_{jk}, \Sigma_{\delta,j,kk})$ for $i = 1, \dots, N$. We therefore draw $\Sigma_{\delta,j,kk}$ from the inverse Gamma distribution

$$\Sigma_{\delta,j,kk} \sim IG \left(\frac{\sum_{i=1}^N (\delta_{jki} - \tilde{\delta}_{jk})^2 + \kappa_1}{2}, \frac{N + \kappa_0}{2} \right),$$

for $k = 1, 2$, with prior parameters κ_0 and κ_1 .

4.A.4.8 Sampling Φ

To sample Φ , note that Equation (4.6) is a multivariate regression model given β and Σ_β . We draw Φ from $MVN(\hat{\Phi}, \hat{\Sigma}_\Phi)$ where

$$\begin{aligned} \hat{\Sigma}_\Phi &= \left(\sum_{t=2}^T \beta_{t-1} \Sigma_\beta^{-1} \beta_{t-1} + \Sigma_{\Phi_0}^{-1} \right)^{-1}, \\ \hat{\Phi} &= \hat{\Sigma}_\Phi \left(\sum_{t=2}^T \beta_{t-1} \Sigma_\beta^{-1} \beta_t + \Sigma_{\Phi_0}^{-1} \hat{\Phi}_0 \right), \end{aligned}$$

where $\hat{\Phi}_0$ is the mean vector and Σ_{Φ_0} is the covariance matrix of the normal prior for Φ .

4.A.4.9 Sampling Σ_β

To sample Σ_β , note that Φ is a diagonal matrix and we can therefore draw each k^{th} diagonal elements separately. For the k^{th} element, we have the univariate regression model

$$\beta_{k,t+1} = \Phi_{kk}\beta_{kt} + \nu_{kt}, \quad \nu_{kt} \sim N(0, \Sigma_{\beta,kk}), \quad \beta_{k1} \sim N(0, 5\Sigma_{\beta,kk}).$$

We therefore draw $\Sigma_{\beta,kk}$ from the inverse Gamma distribution

$$\Sigma_{\beta,kk} \sim IG\left(\frac{\sum_{t=2}^T (\beta_{kt} - \Phi_{kk}\beta_{k,t-1})^2 + \beta_{k1}^2/5 + \kappa_{\beta,1}}{2}, \frac{T + \kappa_{\beta,0}}{2}\right),$$

with prior parameters $\kappa_{\beta,0}$ and $\kappa_{\beta,1}$.

4.A.4.10 Sampling Σ_η

To sample Σ_η in a computationally efficiently manner, we use an independence Metropolis-Hastings (MH) step (Metropolis et al., 1953, Hastings, 1970). For this purpose, we first reparameterize Σ_η into elements that are unconstrained, using a Cholesky decomposition. Next, we draw a candidate for the unconstrained parameters from a multivariate normal distribution with as mean the posterior mode, and as covariance matrix the negative of the inverse of the Hessian of the log posterior at the posterior mode. To find the posterior mode and Hessian, we perform an optimization using the analytic gradient and an approximated Hessian from the outer-product-of-gradients (BHHH) method. Details are in the Supplementary Materials, available upon request.

Chapter 5

Conclusions

People differ. We have different preferences, we respond differently to health treatments, different educational formats work best for us, and so forth. To develop effective policies, it is often useful to acknowledge and account for these individual differences. That is, when one knows how individuals can differentially respond, one can gain understanding into the different effects of a policy across individuals and one can develop personalized policies, e.g. personalized health treatments, marketing or education.

To infer individual responses from a given set of data, models are a useful tool. Unfortunately, the amount of data available on a given individual is often too limited to accurately infer her responses based on her data alone. In these cases, models can be used that consider the underlying population distribution of individual responses. These models share information across all individuals in the dataset. Once the population distribution of responses has been estimated, one can infer per individual where s/he most likely is in the distribution based on the individual's data.

In this thesis, I develop approaches to accurately estimate the population distribution of responses. The proposed approaches overcome important limitations of the existing approaches by allowing for more realistic behavior. That is, they allow for many different forms of response distributions, including those where some individuals may have no response to certain variables (Chapter 2). Also, the proposed methods allow for the responses of individuals to change over time (Chapters 3 and 4). In the applications in this thesis, I find that our proposed approaches lead to improved

predictions of individual outcomes. Also, the approaches lead to interesting insights into how individuals respond. These improvements can lead to the design of more effective policies.

References

- About Nabout, N., Skiera, B., Stepanchuk, T. & Gerstmeier, E. (2012). An analysis of the profitability of fee-based compensation plans for search engine marketing. *International Journal of Research in Marketing*, 29(1), 68–80.
- Agarwal, A., Hosanagar, K. & Smith, M.D. (2011). Location, location, location: An analysis of profitability of position in online advertising markets. *Journal of Marketing Research*, 48(6), 1057–1073.
- Albert, J.H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679.
- Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 1152–1174.
- Beggs, S., Cardell, S. & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, 17(1), 1–19.
- Bhat, C.R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, 35(7), 677–693.
- Bhat, C.R. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37(9), 837–855.
- Bhat, C.R. & Castelar, S. (2002). A unified mixed logit framework for modeling revealed and stated preferences: Formulation and application to congestion pricing analysis in the San Francisco Bay area. *Transportation Research Part B: Methodological*, 36(7), 593–616.
- Bhat, C.R. & Sidharthan, R. (2011). A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multino-

- mial probit models. *Transportation Research Part B: Methodological*, 45(7), 940–953.
- Blake, T., Nosko, C. & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica*, 83(1), 155–174.
- Borgs, C., Chayes, J., Immorlica, N., Jain, K., Etesami, O. & Mahdian, M. (2007). Dynamics of bid optimization in online advertisement auctions. *Proceedings of the 16th International Conference on World Wide Web*, 531–540.
- Braaten, E. & Weller, G. (1979). An improved low-discrepancy sequence for multidimensional quasi-Monte Carlo integration. *Journal of Computational Physics*, 33(2), 249–258.
- Bradley, M. & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation*, 21(2), 167–184.
- Braga, J. & Starmer, C. (2005). Preference anomalies, preference elicitation and the discovered preference hypothesis. *Environmental and Resource Economics*, 32(1), 55–89.
- Cai, B. & Dunson, D. (2005). Variable selection in nonparametric random effects models. *Technical report, Department of Statistical Science, Duke University*.
- Campbell, D., Boeri, M., Doherty, E. & Hutchinson, W.G. (2015). Learning, fatigue and preference formation in discrete choice experiments. *Journal of Economic Behavior & Organization*, 119, 345–363.
- Campbell, D., Hensher, D.A. & Scarpa, R. (2011). Non-attendance to attributes in environmental choice analysis: A latent class specification. *Journal of Environmental Planning and Management*, 54(8), 1061–1076.
- Carter, C.K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3), 541–553.
- Cary, M., Das, A., Edelman, B., Giotis, I., Heimerl, K., Karlin, A.R., Mathieu, C. & Schwarz, M. (2007). Greedy bidding strategies for keyword auctions. *Proceedings of the 8th ACM Conference on Electronic Commerce*, 262–271.
- Chapman, R.G. & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research*, 19(3), 288–301.
- Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2), 221–241.
- Collins, A.T., Rose, J.M. & Hensher, D.A. (2013). Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological*, 56, 234–253.

-
- Cook, S.R., Gelman, A. & Rubin, D.B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3), 675–692.
- Czajkowski, M., Giergiczny, M. & Greene, W.H. (2014). Learning and fatigue effects revisited: Investigating the effects of accounting for unobservable preference and scale heterogeneity. *Land Economics*, 90(2), 324–351.
- Danaf, M., Atasoy, B. & Ben-Akiva, M. (2020). Logit mixture with inter and intra-consumer heterogeneity and flexible mixing distributions. *Journal of Choice Modelling*, 35, 100188.
- Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.
- DeSarbo, W.S., Lehmann, D.R. & Hollman, F.G. (2004). Modeling dynamic effects in repeated-measures experiments involving preference/choice: An illustration involving stated preference analysis. *Applied Psychological Measurement*, 28(3), 186–209.
- Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 204–223.
- Dinner, I.M., Van Heerde, H.J. & Neslin, S.A. (2014). Driving online and offline sales: The cross-channel effects of traditional, online display, and paid search advertising. *Journal of Marketing Research*, 51(5), 527–545.
- Dunson, D.B., Herring, A.H. & Engel, S.M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, 103(482), 534–546.
- Durbin, J. & Koopman, S.J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3), 603–616.
- Durbin, J. & Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.
- Eddelbuettel, D. & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8), 1–18. <https://doi.org/10.18637/jss.v040.i08>
- Edelman, B., Ostrovsky, M. & Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *The American Economic Review*, 97(1), 242–259.

- Fan, W. & Bouguila, N. (2013). Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection. *Pattern Recognition*, 46(10), 2754–2769.
- Ferguson, T.S. et al. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 2(4), 615–629.
- Fiebig, D.G., Keane, M.P., Louviere, J. & Wasi, N. (2010). The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. *Marketing Science*, 29(3), 393–421.
- Fok, D., Paap, R. & Van Dijk, B. (2012). A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. *Journal of Applied Econometrics*, 27(5), 831–846.
- Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2), 183–202.
- Frühwirth-Schnatter, S. & Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. *Statistical Modelling and Regression Structures* (pp. 111–132). Springer.
- Gelfand, A.E. & Smith, A.F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.
- Geman, S. & Geman, D. (1987). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Readings in computer vision* (pp. 564–584). Elsevier.
- George, E.I. & McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423), 881–889.
- George, E.I. & McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 339–373.
- Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99(467), 799–804.
- Ghose, A. & Yang, S. (2008). Analyzing search engine advertising: Firm behavior and cross-selling in electronic markets. *Proceedings of the 17th International Conference on World Wide Web*, 219–226.
- Ghose, A. & Yang, S. (2009). An empirical analysis of search engine advertising: Sponsored search in electronic markets. *Management Science*, 55(10), 1605–1622.
- Gilbride, T.J., Allenby, G.M. & Brazell, J.D. (2006). Models for heterogeneous variable selection. *Journal of Marketing Research*, 43(3), 420–430.

-
- Goldfeld, S.M. & Quandt, R.E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1(1), 3–15.
- Graepel, T., Candela, J.Q., Borchert, T. & Herbrich, R. (2010). Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 13–20.
- Green, P.E. (1974). On the design of choice experiments involving multifactor alternatives. *Journal of Consumer Research*, 1(2), 61–68.
- Greene, W.H. (2012). *Econometric Analysis* (seventh). Pearson Education International.
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, 357–384.
- Hamilton, J.D. (1990). Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1-2), 39–70.
- Hamilton, J.D. (1994). *Time Series Analysis*. Princeton university press Princeton.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.
- Hausman, J.A. & Ruud, P.A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of econometrics*, 34(1-2), 83–104.
- Hensher, D.A., Collins, A.T. & Greene, W.H. (2013). Accounting for attribute non-attendance and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: A warning on potential confounding. *Transportation*, 40(5), 1003–1020.
- Hensher, D.A. & Greene, W.H. (2010). Non-attendance and dual processing of common-metric attributes in choice analysis: A latent class specification. *Empirical Economics*, 39(2), 413–426.
- Hess, S., Hensher, D.A. & Daly, A. (2012). Not bored yet—revisiting respondent fatigue in stated choice experiments. *Transportation Research Part A: Policy and Practice*, 46(3), 626–644.
- Hess, S. & Rose, J.M. (2009). Allowing for intra-respondent variations in coefficients estimated on repeated choice data. *Transportation Research Part B: Methodological*, 43(6), 708–719.
- Hess, S. & Rose, J.M. (2012). Can scale and coefficient heterogeneity be separated in random coefficients models? *Transportation*, 39(6), 1225–1239.

- Hess, S., Stathopoulos, A., Campbell, D., O'Neill, V. & Caussade, S. (2013). It's not that I don't care, I just don't care very much: Confounding between attribute non-attendance and taste heterogeneity. *Transportation*, 40(3), 583–607.
- Hole, A.R. (2011). A discrete choice model with endogenous attribute attendance. *Economics Letters*, 110(3), 203–205.
- Hole, A.R., Kolstad, J.R. & Gyrd-Hansen, D. (2013). Inferred vs. stated attribute non-attendance in choice experiments: A study of doctors' prescription behaviour. *Journal of Economic Behavior & Organization*, 96, 21–31.
- Holmes, C.C. & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1), 145–168.
- Holmes, T.P. & Boyle, K.J. (2005). Dynamic learning and context-dependence in sequential, attribute-based, stated-preference valuation questions. *Land Economics*, 81(1), 114–126.
- Jungbacker, B. & Koopman, S.J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. *The Econometrics Journal*, 18(2), C1–C21.
- Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.
- Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60(1-2), 1–22.
- Kim, S., Dahl, D.B. & Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis (Online)*, 4(4), 707.
- Kim, S., Tadesse, M.G. & Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, 93(4), 877–893.
- Kim, S., DeSarbo, W.S. & Fong, D.K. (2018). A hierarchical Bayesian approach for examining heterogeneity in choice decisions. *Journal of Mathematical Psychology*, 82, 56–72.
- Koç, H. & van Kippersluis, H. (2017). Thought for food: Nutritional information and educational disparities in diet. *Journal of Human Capital*, 11(4), 508–552.
- Koppelman, F.S. & Sethi, V. (2005). Incorporating variance and covariance heterogeneity in the generalized nested logit model: An application to modeling long distance travel choice behavior. *Transportation Research Part B: Methodological*, 39(9), 825–853.
- Korobilis, D. (2013). Bayesian forecasting with highly correlated predictors. *Economics Letters*, 118(1), 148–150.
- Lavrakas, P.J. (2008). *Encyclopedia of survey research methods*. Sage Publications.

-
- Li, H. & Kannan, P. (2014). Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment. *Journal of Marketing Research*, 51(1), 40–56.
- Louviere, J.J. & Woodworth, G. (1983). Design and analysis of simulated consumer choice or allocation experiments: An approach based on aggregate data. *Journal of Marketing Research*, 20(4), 350–367.
- MacLehose, R.F., Dunson, D.B., Herring, A.H. & Hoppin, J.A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology*, 18(2), 199–207.
- Manski, C.F. (1977). The structure of random utility models. *Theory and Decision*, 8(3), 229.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). Academic Press: New York.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.
- Mitchell, T.J. & Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404), 1023–1032.
- Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2), 249–265.
- Plott, C.R. (1993). Rational individual behavior in markets and social choice processes.
- Polson, N.G., Scott, J.G. & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504), 1339–1349.
- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/>
- Rao, V.R. (2014). *Applied conjoint analysis*. Springer.
- Roberts, G.O., Gelman, A., Gilks, W.R. et al. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1), 110–120.
- Roberts, G.O., Rosenthal, J.S. et al. (2001). Optimal scaling for various Metropolis–Hastings algorithms. *Statistical Science*, 16(4), 351–367.

- Rossi, P. (2015). Bayesm: Bayesian inference for marketing/micro-econometrics, 2012. URL <http://CRAN.R-project.org/package=bayesm>. R package version, 2–2.
- Rossi, P., Allenby, G. & McCulloch, R. (2005). *Bayesian statistics and marketing*. New York: Wiley.
- Rossi, P. (2014). *Bayesian non-and semi-parametric methods and applications*. Princeton University Press.
- Rutz, O.J. & Bucklin, R.E. (2007). A model of individual keyword performance in paid search advertising. Available at SSRN 1024765.
- Rutz, O.J. & Bucklin, R.E. (2011). From generic to branded: A model of spillover in paid search advertising. *Journal of Marketing Research*, 48(1), 87–102.
- Rutz, O.J., Bucklin, R.E. & Sonnier, G.P. (2012). A latent instrumental variables approach to modeling keyword conversion in paid search advertising. *Journal of Marketing Research*, 49(3), 306–319.
- Rutz, O.J., Trusov, M. & Bucklin, R.E. (2011). Modeling indirect effects of paid search advertising: Which keywords lead to more future visits? *Marketing Science*, 30(4), 646–665.
- Ryan, D. (2016). *Understanding digital marketing: Marketing strategies for engaging the digital generation*. Kogan Page Publishers.
- Savage, S.J. & Waldman, D.M. (2008). Learning and fatigue during choice experiments: A comparison of online and mail survey modes. *Journal of Applied Econometrics*, 23(3), 351–371.
- Scarpa, R., Gilbride, T.J., Campbell, D. & Hensher, D.A. (2009). Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, 36(2), 151–174.
- Skiera, B. & Abou Nabout, N. (2013). PROSAD: A bidding decision support system for profit optimizing search engine advertising. *Marketing Science*, 32(2), 213–220.
- Sociaal en Cultureel Planbureau. (2004). Culturele veranderingen in Nederland 2004 - CV'04 [data retrieved from DANS, <https://doi.org/10.17026/dans-xtud36b>].
- Tang, X., Xue, F. & Qu, A. (2020). Individualized multi-directional variable selection. *Journal of the American Statistical Association*, (forthcoming).
- Tanner, M.A. & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

-
- Van Heerde, H.J., Mela, C.F. & Manchanda, P. (2004). The dynamic effect of innovation on market structure. *Journal of Marketing Research*, 41(2), 166–183.
- Wang, C. & Blei, D.M. (2009). Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. *Advances in Neural Information Processing Systems*, 1982–1989.
- Windle, J., Polson, N. & Scott, J. (2014). BayesLogit: Bayesian logistic regression. *R package version: 0.5, 1*.
- Windle, J., Polson, N.G. & Scott, J.G. (2014). Sampling Pólya-Gamma random variates: Alternate and approximate techniques. *arXiv preprint arXiv:1405.0506*.
- Wu, F., Han, Y., Liu, X., Shao, J., Zhuang, Y. & Zhang, Z. (2012). The heterogeneous feature selection with structural sparsity for multimedia annotation and hashing: A survey. *International Journal of Multimedia Information Retrieval*, 1(1), 3–15.
- Yang, M. (2012). Bayesian variable selection for logistic mixed model with non-parametric random effects. *Computational Statistics & Data Analysis*, 56(9), 2663–2674.
- Yang, S. & Ghose, A. (2010). Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence? *Marketing Science*, 29(4), 602–623.
- Yao, S. & Mela, C.F. (2011). A dynamic model of sponsored search advertising. *Marketing Science*, 30(3), 447–468.
- Yu, G., Huang, R. & Wang, Z. (2010). Document clustering via Dirichlet process mixture model with feature selection. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 763–772.
- Zhao, L., Hu, Q. & Wang, W. (2015). Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Transactions on Multimedia*, 17(11), 1936–1948.

Nederlandse samenvatting (Summary in Dutch)

In dit proefschrift ontwikkel ik methodes om individuele uitkomsten te verklaren. Deze methodes richten zich op het accuraat schatten en voorspellen van reacties van individuen: hoe reageren individuen (bijvoorbeeld met hun aankoopgedrag) op een verandering in verklarende variabelen (bijvoorbeeld prijs)? Wanneer de reacties van individuen bekend zijn, kunnen publieke en private organisaties deze informatie gebruiken om effectief beleid te maken. Zo kunnen bijvoorbeeld zorgverleners hun gezondheidsbehandelingen afstemmen op een individu of kunnen supermarkten persoonlijke aanbiedingen en aanbevelingen doen.

De methodes ontwikkelt in dit proefschrift dragen bij aan de literatuur door het toestaan van meer realistisch individueel gedrag, met name wanneer datasets weinig informatie per individu bevatten. Zo staan de methodes toe dat individuen sterk van elkaar kunnen verschillen, en dat sommige factoren geen invloed hebben op bepaalde individuen (hoofdstuk 2). Daarnaast staan de methodes toe dat het gedrag van individuen over tijd kan veranderen (hoofdstukken 3 en 4). In de toepassingen in dit proefschrift zien we dat de methodes leiden tot betere voorspellingen van individuele reacties. Deze verbeterde voorspellingen kunnen worden gebruikt om beter beleid te ontwerpen. De methodes zijn algemeen toepasbaar op verschillende soorten problemen, zoals binnen de gezondheidszorg en marketing.

About the author



As a researcher, Aniek (1992) focuses on developing state-of-the-art models and approaches to accurately capture individual responses. As such, she aims to provide useful methodological tools that can be used by practitioners in a broad range of research fields, including health and consumer choice-making. She has presented her work at several scientific meetings including the European Marketing Academy and the Netherlands Econometric Study Group.

Aniek obtained her MSc (2015) degree, cum laude, in Econometrics and Management Science at the Erasmus University Rotterdam. During her studies, she gained practical experience working as a trainee at the data analytics consultancy firm MIcompany (2012-2015). There, she was responsible for gaining insights from customer data for clients which came from the telecom, banking and retailing industries. In 2015, she started her PhD candidacy at the Econometric Institute at the Erasmus University Rotterdam under the supervision of Prof. Dr. Dennis Fok and Prof. Dr. Richard Paap.

Portfolio

Publications

Working papers

A. Castelein, D. Fok and R. Paap (2020). *Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach*. TI discussion paper TI 2020-061/III. Tinbergen Institute.

A. Castelein, D. Fok and R. Paap (2020). *A multinomial and rank-ordered logit model with inter- and intra-individual heteroscedasticity*. TI discussion paper 2020-069/III. Tinbergen Institute.

A. Castelein, D. Fok and R. Paap (2019). *Dynamics in clickthrough and conversion probabilities of paid search advertisements*. TI discussion paper TI 2019-056/III. Tinbergen Institute.

Teaching

Bachelor Thesis, Erasmus School of Economics, Econometrics and Operations Research. Supervision. 2015-2020.

Bachelor Case Studies, Erasmus School of Economics, Econometrics and Operations Research Ba3. Supervision. 2015-2020.

Time Series Analysis, Erasmus School of Economics, Econometrics and Operations Research Ba2. Programming exercise lectures. 2015-2017.

Master Thesis, Erasmus School of Economics, Econometrics and Management Science. Supervision. 2016-2020.

Data Analyse - Grip krijgen op Big Data, Erasmus Academie, post-academic course. Providing assistance with programming questions. 2016-2017.

Seminar Case Studies, Erasmus School of Economics, Econometrics and Management Science. Supervision. 2017-2018.

Statistics, Erasmus School of Economics, Econometrics and Operations Research Ba1. Design, organization and supervision of programming assignments. 2018-2020.

PhD Courses

Microeconomics I (Individual Decision Making and General Equilibrium)

Microeconomics II (Game Theory)

Microeconomics III (Information and Contract Theory)

Microeconomics IV (Behavioral Economics)

Advanced Time Series Econometrics

Bayesian Econometrics

Computational Economics with Python

Statistical Learning and Data Science (TI Econometrics Lectures 2017)

Advanced Econometrics II (Instrumental variables, GMM, Likelihood-based techniques)

Advanced Econometrics III (State Space Methods)

Health Economics

Labor Economics

Scientific Integrity

Conferences Prestented at

International Econometrics PhD Conference 2019, Rotterdam, The Netherlands.

Netherlands Econometrics Study Group 2019, Amsterdam, The Netherlands.

Econometric Institute PhD Conference 2019, Rotterdam, The Netherlands.

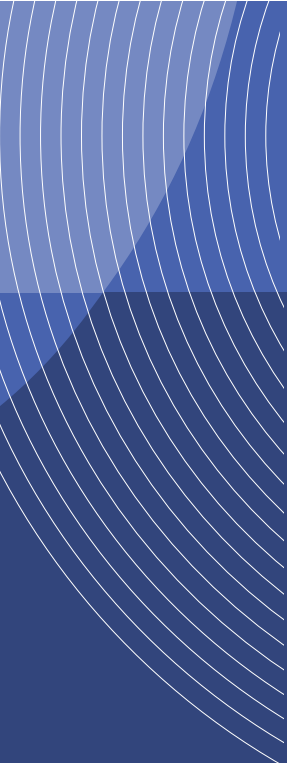
European Marketing Academy 2017, Groningen, The Netherlands.

Econometric Institute PhD Conference 2017, Rotterdam, The Netherlands.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

- 725 S. ALBRECHT, *Empirical Studies in Labour and Migration Economics*
726 Y. ZHU, *On the Effects of CEO Compensation*
727 S. XIA, *Essays on Markets for CEOs and Financial Analysts*
728 I. SAKALAUŠKAITE, *Essays on Malpractice in Finance*
729 M.M. GARDBERG, *Financial Integration and Global Imbalances.*
730 U. THÜMMEL, *Of Machines and Men: Optimal Redistributive Policies under Technological Change*
731 B.J.L. KEIJSERS, *Essays in Applied Time Series Analysis*
732 G. CIMINELLI, *Essays on Macroeconomic Policies after the Crisis*
733 Z.M. LI, *Econometric Analysis of High-frequency Market Microstructure*
734 C.M. OOSTERVEEN, *Education Design Matters*
735 S.C. BARENDSE, *In and Outside the Tails: Making and Evaluating Forecasts*
736 S. SÓVÁGÓ, *Where to Go Next? Essays on the Economics of School Choice*
737 M. HENNEQUIN, *Expectations and Bubbles in Asset Market Experiments*
738 M.W. ADLER, *The Economics of Roads: Congestion, Public Transit and Accident Management*
739 R.J. DÖTTLING, *Essays in Financial Economics*
740 E.S. ZWIERS, *About Family and Fate: Childhood Circumstances and Human Capital Formation*
741 Y.M. KUTLUAY, *The Value of (Avoiding) Malaria*
742 A. BOROWSKA, *Methods for Accurate and Efficient Bayesian Analysis of Time Series*
743 B. HU, *The Amazon Business Model, the Platform Economy and Executive Compensation: Three Essays in Search Theory*
744 R.C. SPERNA WEILAND, *Essays on Macro-Financial Risks*
745 P.M. GOLEC, *Essays in Financial Economics*
746 M.N. SOUVERIJN, *Incentives at work*
747 M.H. COVENEY, *Modern Imperatives: Essays on Education and Health Policy*
748 P. VAN BRUGGEN, *On Measuring Preferences*
749 M.H.C. NIENTKER, *On the Stability of Stochastic Dynamic Systems and their use in Econometrics*
750 S. GARCIA MANDICÓ, *Social Insurance, Labor Supply and Intra-Household Spillovers*
751 Y. SUN, *Consumer Search and Quality*

- 752 I. KERKEMEZOS, *On the Dynamics of (Anti) Competitive Behaviour in the Airline Industry*
- 753 G.W. GOY, *Modern Challenges to Monetary Policy*
- 754 A.C. VAN VLODRUP, *Essays on Modeling Time-Varying Parameters*
- 755 J. SUN, *Tell Me How To Vote, Understanding the Role of Media in Modern Elections*
- 756 J.H. THIEL, *Competition, Dynamic Pricing and Advice in Frictional Markets: Theory and Evidence from the Dutch Market for Mortgages*
- 757 A. NEGRIU, *On the Economics of Institutions and Technology: a Computational Approach*
- 758 F. GRESNIGT, *Identifying and Predicting Financial Earth Quakes using Hawkes Processes*
- 759 A. EMIRMAHMUTOGLU, *Misperceptions of Uncertainty and Their Applications to Prevention*
- 760 A. RUSU, *Essays in Public Economics*
- 761 M.A. COTOFAN, *Essays in Applied Microeconomics: Non-Monetary Incentives, Skill Formation, and Work Preferences*
- 762 B.P.J. ANDRÉE, *Theory and Application of Dynamic Spatial Time Series Models*
- 763, P. PELZL, *Macro Questions, Micro Data: The Effects of External Shocks on Firms*
- 764 D.M. KUNST, *Essays on Technological Change, Skill Premia and Development*
- 765 A.J. HUMMEL, *Tax Policy in Imperfect Labor Markets*
- 766 T. KLEIN, *Essays in Competition Economics*
- 767 M. VIGH, *Climbing the Socioeconomic Ladder: Essays on Sanitation and Schooling*
- 768 YAN XU, *Eliciting Preferences and Private Information: Tell Me What You Like and What You Think*
- 769 S. RELLSTAB, *Balancing Paid Work and Unpaid Care over the Life-Cycle*
- 770 Z. DENG, *Empirical Studies in Health and Development Economics*
- 771 L. KONG, *Identification Robust Testing in Linear Factor Models*
- 772 I. NEAMȚU, *Unintended Consequences of Post-Crisis Banking Reforms*
- 773 B. KLEIN TEESELINK, *From Mice to Men: Field Studies in Behavioral Economics*
- 774 B. TEREICK, *Making Crowds Wiser: The Role of Incentives, Individual Biases, and Improved Aggregation*



In this thesis, I develop approaches to explain individual outcomes. These approaches focus on accurately estimating and predicting individual responses: how do individuals react (e.g. with their purchase behavior) to changes in explanatory variables (e.g. price)? When the responses of individuals are known, public and private organizations can use the information to develop effective policies. For example, health care providers can personalize their health treatments, or supermarkets can create personalized recommendations.

The approaches developed in this thesis contribute to the literature by allowing for more realistic individual behavior, especially when the dataset contains little information per individual. The approaches allow for individuals to have widely different responses, and for some individuals to be unaffected by certain variables (chapter 2). Also, the approaches allow for the responses of individuals to change over time (chapters 3 and 4). In the applications in this thesis, I find that the proposed approaches lead to improved predictions of individual outcomes. These improvements can lead to the design of more effective policies. The approaches are generally applicable to many real-life problems, including problems in health and consumer choice-making.

