Marketing Analytics for High-Dimensional Assortments

Marketing Analytics for High-Dimensional Assortments

Marketing analyse methodes voor hoog-dimensionale assortimenten

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

prof. dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on
Friday, December 22, 2017 at 13:30 hours

by

Bruno Julian David Jacobs
born in Rotterdam.

Erasmus University Rotterdam

Doctoral Committee

Promotors: Prof. dr. A.C.D. Donkers
Prof. dr. D. Fok

Other members: Prof. dr. T.H.A. Bijmolt
Prof. dr. B.G.C. Dellaert
Prof. dr. R. Paap

Erasmus Research Institute of Management – ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam.
Internet: https://www.erim.eur.nl

ERIM Electronic Series Portal: https://repub.eur.nl

ERIM PhD Series in Research in Management, 445
ERIM reference number: EPS-2017-445-MKT
ISBN 978-90-5892-503-9

Design: PanArt, www.panart.nl

This publication (cover and interior) is printed by Tuijtel on recycled paper, BalanceSilk®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Ecolabel, FSC®, ISO14001.
More info: www.tuijtel.com

FSC www.fsc.org MIX Paper from responsible sources FSC® C007225

EU Ecolabel www.ecolabel.eu

Dankwoord

Het is zover: het boekje (of werkstuk, zoals mijn vader het steevast blijft noemen) is eindelijk af. Nu het einde van mijn promotietraject nadert, besef ik mij pas goed hoe veel ik heb geleerd en gedaan de afgelopen jaren. Natuurlijk bestaat het promoveren uit het volgen van vakken, lezen van literatuur, schrijven van papers, presenteren van onderzoek, ontdekken van nieuwe programmeertalen en lesgeven voor volle collegezalen op de universiteit. Maar het heeft mij ook, letterlijk, een bredere kijk op de wereld gegeven door mij te leiden van Rotterdam naar Istanbul, Atlanta, Philadelphia, Baltimore, London, New York, Atlanta, Groningen, Durham, College Park, Alberta, Sydney, Tilburg en tot slot Fontainebleau naar uiteindelijk mijn nieuwe thuis in Maryland, USA. Zonder de steun van vrienden, familie en collega's had ik deze reis misschien wel nooit volbracht en was het zeker niet zo leuk geweest. Graag wil ik jullie hier voor bedanken in dit dankwoord.

Allereerst mijn promotoren: Bas en Dennis. Het was, zoals wel vaker het geval, best nog even spannend of het allemaal op tijd ging lukken maar de eindstreep is in zicht. Ik ben jullie zeer dankbaar dat jullie er in 2012 voor kozen om het avontuur met mij aan te gaan door mij een promotieplek aan te bieden. Bas, je was altijd bereid om mijn vragen te beantwoorden, vaak buiten de spreekwoordelijke kantooruren om. Ook wist je mij als geen ander scherp te houden en te motiveren. Dennis, jouw colleges in de Master Econometrie wakkerde mijn interesse aan voor een promotieplek onder jouw begeleiding. Dat heeft alle verwachtingen overtroffen, ik heb veel van je geleerd. Ook heb ik genoten van onze gesprekken die niet (direct) gerelateerd waren aan werk, waarbij de onderwerpen konden variëren van hardlopen tot de laatste technologische ontwikkelingen. Dat jullie het beste in mij naar boven hebben gehaald is ook tot uiting gekomen in onze publicatie in *Marketing Science*. Ik kijk met veel plezier terug op onze samenwerking en ik hoop dat wij hier in de toekomst een mooi vervolg aan kunnen geven.

Asim Ansari, who hosted me during my research visit at Columbia Business School. Asim, during my visit you were always cheerful and I thank you for your hospitality and the time you dedicated to me during my visit to New York. Now that the geographical distance between us has diminished considerably, I hope that we can continue collaborating in the future.

Tammo Bijmolt, Benedict Dellaert en Richard Paap, bedankt dat jullie plaats hebben genomen in mijn leescommissie en mijn proefschrift hebben beoordeeld. Daarnaast wil ik jullie graag bedanken voor de opmerkingen die het proefschrift hebben verbeterd.

Mijn twee paranimfen: Tom en Maarten. Tom, al vroeg in onze Master Econometrie had jij mijn carrièrepad al uitgestippeld, ik citeer: *"Bruno!! Gewoon gaan promoveren, vet mooi!!"*. Ik ben blij dat ik niet van regime ben geswitcht en heb genoten van onze promotietijd samen. Helaas is onze dubbele tenure track in Groningen op een haar na niet rond gekomen. En voordat ik het vergeet, nog gefeliciteerd met je verjaardag trouwens! Maarten, zoals Tom mij de promotie heeft ingetrokken, zo heb jij mij er doorheen gesleept. Op de juiste momenten wist jij mij even helemaal uit het onderzoek te trekken en daar heb ik heel veel aan gehad. Verder heb je zeker mijn vocabulaire verrijkt (of verkleind, ouwes?) en dat je zelfs de wereld bent overgevlogen om met mij de finale van Twin Peaks te bekijken zegt genoeg over zowel onze gedeelde interesses als onze vriendschap.

Tijdens mijn promotie heb ik met niemand zoveel tijd doorgebracht als met mijn kamergenoten in het Tinbergen gebouw in H7-33 en later H8-11. Aiste, I truly hope it was bearable for you to share an office with Tom and me. I enjoyed your company and I am sure that you are having a great time in Australia! Tom, gemakshalve ga ik er maar even van uit dat jouw plotselinge verhuizing naar Groningen, midden in onze promotie, niets met mij als kamergenoot te maken had. Bart! Bedankt dat jij mij wilde opvangen op H8-11. De mini-tafeltennis sessies waren legendarisch en worden gemist! Ongetwijfeld waren die potjes ook het fundament voor onze derde plaats op de Econometric Game, behaald onder jouw bezielende leiding als teamcaptain. Nu op UMD blijkt al jouw lacrosse kennis van onschatbare waarde te zijn. Go Terps!

Voor een groot gedeelte is mijn promotie ook gevormd door mijn (voormalige) collega's op de universiteit. Zonder jullie was mijn promotie niet hetzelfde geweest! Dennis, Lisa is een blijvertje gebleken maar Julia (vooralsnog) niet meer dan een escapade. Ik heb genoten van alle tijd die we binnen, maar misschien nog wel meer buiten, de universiteit hebben doorgebracht. En Esther, heb jij de prinses? Didier, misschien wel de snelste PhD ter wereld en zeker de beste coparanimf die ik mij kon wensen. Wie bestelt het volgende kaasplankje?! Gertjan, bedankt dat je op de meest onmogelijke tijdstippen bereid was de meest obscure Python zaken te bespreken. Ook heel veel dank voor het beschikbaar stellen van de LaTeX-template voor dit proefschrift! Sander, voor het delen van een aantal muzikale pareltjes (George Fitzgerald!) en het voorzien van de werkvloer van enige hoofdstedelijke branie. Francine, jij wist de academische wereld altijd in perspectief te plaatsen. Koen, voor het delen van een tent op Pukkelpop en dat je deadmau5 (on)vrijwillig hebt aangehoord. Myrthe, voor de gezellige koffiepauzes en al je hulp en advies, in het bijzonder met de job market. Arash, thanks for sharing my enthusiasm for graphical models and variational inference. Tommi, for sparking my enthusiasm in an academic career and sharing the following words of wisdom: "*I mean like .... it's an academic degree so maybe it couldn't hurt to learn a thing or two*". I would like to thank my other (former) colleagues at the Marketing department at the Erasmus School of Economics, the Econometric Institute, and ERIM. In particular Anne, Bert, Damir, Elio, Florian, Gert-Jan, Marcel, Martijn, Max, Michel, Tülay, Victor, Yuri.

Mijn ouders, Vincent en Dorien. Dorien, omdat jij altijd hebt gezegd dat ik er wel zou komen. Vincent, zonder jou was ik niet de persoon geworden die ik nu ben. Ik ben je voor altijd dankbaar. Voor alles. Milan, wat er ook gebeurt, je blijft mijn kleine broertje. Oma Riet, ik koester warme herinneringen aan de woensdagmiddagen in Pendrecht, samen met opa Wim heb je veel voor mij betekend. Oma Greet, u bent daadwerkelijk de krachtigste vrouw die ik ken en ik neem hier graag een voorbeeld aan. Conny, Peter, Erwin, Amy, Wouter, als ik bij jullie ben dan voelt het als een tweede thuis en is het altijd goed. Peter, Yvonne, Maud, Milou, bedankt voor alle Limburgse gezelligheid en gastvrijheid!

Lieve Rianne, natuurlijk is dit dankwoord niet compleet zonder jou te noemen. Tijdens mijn promotie ben jij de constante factor geweest waar ik altijd op kon vertrouwen. Jij haalt het beste in mij naar boven en laat mij mijn grenzen verleggen. Samen hebben wij de stap gemaakt van het vertrouwde Rotterdam naar het toen nog onbekende Maryland. Ongetwijfeld staan er ons nog veel avonturen te wachten en zolang dat samen met jou is, durf ik het aan.

Table of Contents

1

Introduction

In this dissertation I present solutions that address challenges in the context of marketing analytics in high-dimensional retail assortments. The common denominator of these solutions is that they all build on recent advances made in the machine learning literature, which are adapted and extended to make them suitable for marketing analytics. In turn, I try to contribute to the machine learing literature as well, with some of the more technical results I derived in this dissertation.

1.1 Marketing analytics in large assortments

The size of the product assortment at a typical retailer has exploded over the past two decades. This increase can largely be attributed to the rise of online shopping which has allowed retailers to retain assortments that are virtually unlimited in size, against relatively low costs. Another implication of online retailing is that much more information for each individual customer can be obtained. This information is no longer limited to purchases, which can be tracked at a brick-and-mortar store as well, but additionally search and click behavior are often readily available.

In turn, this customer-specific information can be used by a retailer to create personalized marketing efforts, for example, to determine the relevant products in the assortment for a specific customer. This opens the way for advanced (online) personalization that can improve a customer's shopping experience. Additionally, more high-level insight can be obtained from such purchase data. Examples are the identification of purchase patterns and how these patterns vary with customer characteristics and over time. Such insights can be used to improve a wide range of marketing actions.

It is only naturally to assume that the larger the assortment, the more relevant these insights are both for the retailer that holds a large product assortment, as well as for the customer who wants to browse and purchase products from such an assortment. However, as the number of products in the assortment increases, the dimensionality of the available purchase data will increase as well. As a result it becomes increasingly more difficult to use all available data to gain actionable marketing insights.

In practice, retailers circumvent this problem by resorting to simpler heuristics, e.g. a collaborative filter. Such a filter discovers patterns by counting the co-occurrence of products in the purchase history data (Jannach et al., 2010). Applications include the display of relevant alternative products on a product page ("*Customers who bought this item also bought*"), or to display a set of products based on the purchase behavior of a customer ("*Recommended for you*"). Typically, these results can be obtained in near real-time and the filter scales well to very large customer bases and product assortments (e.g. Amazon). A collaborative filter is not without downsides, however. For example, it is not able to display personalized content for a new customer automatically, or to vary offerings by time of day or season. In a stochastic model-based approach this concern could be alleviated by including additional available customer information, such as characteristics. Such information is difficult to include in a filter without partitioning the data into smaller subsets. But arguably the biggest limitation of a collaborative filter is that it hardly provides any insight beyond identifying co-occurring products. This makes it difficult to abstract actionable insights from the data.

Alternatively, there is a long stream of research in the marketing literature involving choice models (Guadagni and Little, 1983, McFadden, 1986, Wagner and Taudes, 1986, Wedel and Kamakura, 2000). These model-based methods are well suited to account for both observed and unobserved customer heterogeneity and can readily be extended to include additional information that is available beyond the purchase history data. Furthermore, a choice model is able to provide insight on both the customer and population level, enabling advanced marketing analytics for the retailer. A major drawback, however, lies in the poor scalability of a typical choice model with respect to the number of alternatives. Especially in the context of large product assortments, i.e. those that contain (at least) hundreds of products, a choice model tends to break down (Naik et al., 2008). This severely limits the application of these methods in practice.

In the marketing literature there are a few solutions commonly used to alleviate these concerns. However, they are inadequate in the setting of large product assortments. In Zanutto and Bradlow (2006), it is advocated that one should focus on just a subset of the data. For example, by considering only a selection of customers instead of the entire customer base. Such an approach may be suitable if one is strictly interested in effects at the population level, but it is inadequate if we want to gain insight in the preferences of individual customers. This is typically the case in marketing applications, as the notion of heterogeneity in customer behavior is widely accepted. In addition, it does not seem intuitive to exclude products a priori from our analysis. Some low-frequency products might be highly relevant for a specific set of customers and by ignoring such products we preclude ourselves from discovering such insights. Even more, the assortment that remains after excluding the low-frequency products may still be too large.

In this dissertation I try to combine the best of both worlds by developing methods that can abstract profound marketing insights from purchase history data, while at the same time ensuring that these methods scale well to the size of the product assortment. To enable this I turn to methods that originate from the machine learning literature. More specifically, the methods that I present in this dissertation are grounded in the class of so-called *topic models*, with as canonical example the latent Dirichlet allocation (Blei et al., 2003) model. Traditionally, these methods have been used to analyze collections of text documents and discover the topics underlying a collection of documents. I adapt and extend these topic models to enable advanced analyses in the context of large product assortments. However, even these scalable models need to be estimated and to make them of practical use this estimation should be fast. This holds especially in marketing applications where near real-time results are required. Therefore, for fast estimation I rely and extend on advanced techniques that originate from the machine learning literature as well. These techniques are the subject of the next section.

1.2 Estimation of complex models

If one cause has to be identified to explain the rise of Bayesian inference in the modern-day literature, the answer would most likely be the development of Markov Chain Monte Carlo (MCMC) samplers (Gelfand and Smith, 1990). The evolution of these methods at the end of the twentieth century, combined with an ever-increasing supply of computational power, have enabled researchers to estimate complex models that simply were not possible to infer in the pre-MCMC era. Today's world, however, is very different when compared to the end of the last century. As a society we produce and store data at an unprecedented rate and in all likelihood, this will only increase over the next decades. All this data can potentially serve as input for statistical models, allowing us to research and reason about increasingly complex phenomena. This raises the question how much longer we will be able to estimate such models using the traditional MCMC methods. Fortunately, several successors are ready to take their place.

The first alternative estimation method we discuss is variational inference (VI) (Jordan et al., 1999, Blei et al., 2017). In VI a new distribution is introduced, called the variational distribution, that is used to approximate the posterior of the model. Subsequently, Bayesian inference is cast into an optimization problem, where the distance from the variational distribution to the posterior distribution is minimized. This optimization can be solved using traditional techniques from the optimization literature. By imposing smart restrictions on the set of admissible variational distributions, this optimization can be simplified. The most common set of restrictions leads to so-called mean-field variational inference (MFVI), which imposes that the variational factorizes over the variables in the posterior. This enables statistical inference at a fraction of the time it would take using traditional inference methods, such as the MCMC samplers (Ansari et al., 2016, Kucukelbir et al., 2017). This approximation, however, affects the results: Typically the variance of the posterior density is underestimated (Blei et al., 2017), while the posterior means are accurately recovered. Hence, in MFVI we exchange asymptotic correctness for estimation speed, which for complex models will often be the difference between being able or unable to estimate a model in a realistic amount of time. Furthermore, there are two properties of VI that make it especially suited for inference in models involving large data sets: We can stochastically subsample sets of data points, which can speed up inference by an order of magnitude in hierarchical models (Hoffman et al., 2013). In addition, because VI is intrinsically a deterministic method, it is often trivial to parallelize its optimization routine across all computing nodes available. These two properties can be combined, if desired.

Alternatively, if one is concerned about retaining the asymptotic properties of the traditional MCMC sampling-based approaches, one could utilize the Hamiltonian Monte Carlo (HMC) sampler (Neal, 1996, Hoffman and Gelman, 2014). These samplers make use of gradient information to more effectively explore the posterior space. In the past, setting up such an HMC sampler was not very user friendly and required a lot of manual derivations and tuning of nuisance parameters. These problems have to a large extent been resolved by Stan (Carpenter et al., 2017), a generic programming language that relies on the No-U-Turn (NUTS) HMC-sampler (Hoffman and Gelman, 2014) for statistical inference. Other alternatives for automatic inference are available, each with certain advantages and disadvantages: Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2017), which uses an automatic mean-field variational inference algorithm to approximate the posterior, Black Box Variational Inference (BBVI) (Ranganath et al., 2013), which uses noisy gradients in MFVI; or Edward

(Tran et al., 2016), which is a generic programming language similar to Stan that allows for rapid prototyping of models, with a focus on model criticism such as posterior predictive checks.

The development of these fast and generic inference tools is invaluable for academia, as it enables a researcher to dedicate his attention to the research problem at hand, instead of getting distracted by implementation details. An additional advantage is that this allows for a critical evaluation of the model specification used. Naturally, the design of an optimal model is not clear a priori (at least not in the social sciences). Especially in the context of models that combine data from various sources, it becomes increasingly important to be able to rapidly explore different model specifications. Recently, this iterative process has been formalized and advocated in Blei (2014), which proposes an adaptation of Box's classical loop (Box and Hunter, 1962). It can be summarized as "*Build, compute, critique, repeat*": i) Build a model. ii) Estimate the model parameters. iii) Criticize the model. iv) If satisfied, use the model for further analysis, if not, return to step i). I share the belief that this is the way empirical models should be build. However, we must beware not to overfit the design of the model to a single application. Instead, our goal should be to find models that can provide insights which are generalizable across multiple settings.

1.3 Outline of the dissertation

In this dissertation I combine the developments in marketing with those in the machine learning literature. This dissertation contains three chapters that can be read independently. Below, I briefly discuss each of them.

In Chapter 2, based on Jacobs, Donkers, and Fok (2016), the primary contribution is the adaptation and extension of latent Dirichlet allocation (LDA), introduced in Blei et al. (2003), in order to model purchases in a retailing context. LDA is a so-called topic model, originally developed to model the occurrence of words in a collection of text documents by using a set of latent topics. I adapted this notion to the retail setting: Documents become customers; words from the vocabulary are replaced by purchases from an assortment; and just as the content of a document covers a set of topics, a customer's purchase history can be described by a set of purchase drivers, which I label motivations. Besides the similarities, there are also major distinctions between the two domains. Contrast the size of a typical document against that of an average purchase history. The adaptation of LDA presented in Chapter 2 accounts for these differences, by explicitly modeling the prior baseline relevance of each motivation and linking this to observed customer characteristics. The resulting method is applied to a data set from a medium-sized online retailer that manages an assortment of about 400 products. Note that this dimension differs by an order of magnitude with a standard choice model in marketing, which typically concerns just a few alternatives. The inferred motivations make intuitive sense. For example, they describe preferences for eco-friendly products, diet products, or products for the sensitive skin. These motivations also provide a high-level insight into the factors that drive the purchase behavior in the customer base. The method is used to predict future purchases of a customer. It obtains a high predictive accuracy, emphasizing the face validity of the method once more. Furthermore, its performance is similar or better than competing methods, while being far more scalable. This opens up opportunities for all sorts of retailing applications, of which the most apparent one is serving as input for a personalized shopping experience. This chapter is based on Bruno J.D. Jacobs, Bas Donkers, Dennis Fok (2016), “Model-Based Purchase Predictions for Large Assortments,” *Marketing Science*, 35 (3), 389–404. The author contributions for Chapter 2 are as follows: All three co-authors contributed significantly to this chapter.

Chapter 3 consists of two parts. In the first part, I start with an overview of the fundamental concepts associated with variational inference (VI), after which I zoom in on mean-field variational inference (MFVI). While introducing these concepts, I provide additional insights that can be valuable for the reader who is more familiar with traditional Bayesian statistical inference techniques, but wants to understand VI. In the second part, I build upon these concepts and present two new results in the context of MFVI applied to models that involve hierarchical Normals. A hierarchical Normal is defined as a Normal distribution where the mean parameter is specified as a function of one (or more) other random variables that each are Normally distributed as well. Such hierarchical models are often used to capture customer heterogeneity in marketing applications (Rossi et al., 2012). The first result revolves around the dependencies of parameters in a hierarchical Normal model. The intuition is simple, but elegant: Consider a set of 100 numbers and suppose one is tasked with the computation of the sum for every subset of 99 numbers (without repetition). A direct, but naive, approach is to construct 100 sets of 99 numbers and calculate the sum for each of these sets. An indirect, but smarter, approach is to first calculate the total sum over all 100 numbers. From this total sum, we can subtract each number individually to arrive at the answer. This is exactly the intuition underlying my result. I rewrite the dependencies of a hierarchical Normal model by considering a common error term. The important implication of this result is that it becomes easier to estimate generic hierarchical Normal models in MFVI, without requiring model-specific derivations. The second result is about the (computationally) efficient estimation of a common precision matrix for a set of multivariate Normal (MVN) variables. I show that in MFVI this common covariance structure can be inferred without specifying a separate multivariate Normal variational distribution for each of the MVN variables. Instead, the marginal posterior of each MVN variable is approximated using a set of univariate Normals, called an independent Multivariate Normal (iMVN). I provide closed-form solutions for the parameters of an iMVN that approximates an MVN. This decreases the number of variational parameters that have to be estimated, which significantly lowers the computational complexity of estimating a common covariance structure with MFVI in hierarchical Normal models. The author contributions for Chapter 3 are as follows: The chapter is written by the author of this dissertation. Both prof. dr. Bas Donkers and prof. dr. Dennis Fok provided valuable feedback that improved the structure of this chapter.

Chapter 4, based on work by Jacobs, Donkers, and Fok, extends Chapter 2 in several ways, in order to provide managerial insights that go beyond purchase prediction. The most notable extensions are the following: First, the separate shopping trips in a customer's purchase history are distinguished. Next, motivations are allowed to be correlated in the model. Finally, the model is extended using time-specific effects and explanatory variables at the basket and the customer level. This allows for a richer model structure that provides further insight into the purchase patterns underlying the data. These extensions, however, do not come without a computational cost, and the second direction in which I contribute is by inferring the model parameters using variational inference (VI). This allows us to estimate the model parameters in a fraction of the time that would have been required by status quo inference techniques. More details on VI are provided in Chapter 3. For the application in this chapter I use purchase history data that was made available to me by a retailer through the Wharton Customer Analytics Initiative (WCAI). The model is estimated on a subset of this data and contains over 4,000 distinct products, which is an increase in size by a factor of 10 when compared to the product assortment from the application in Chapter 2. The obtained results are intuitively plausible, have face validity, and show a high degree of internal consistency. The time effects and explanatory variables allow a manager to zoom in on granular levels of the data, providing managerial insights that cannot be obtained by just observing descriptive statistics of the data. Combined with the high predictive power of an LDA-based method, for which empirical evidence was provided in Chapter 2, this opens the way for advanced marketing analytics, which allows targeting with respect to whom to contact when, in high-dimensional product assortments. The author contributions for Chapter 4 are as follows: All three co-authors contributed significantly to this chapter.

To conclude this introduction, I believe that the research conducted in this dissertation is among the first steps on a long and exciting road ahead for large-scale marketing models inspired by the machine learning literature. This claim does not just apply to me personally, but also to the field of advanced marketing analytics in high-dimensional retail settings as a whole. Many open questions are waiting for an answer: What is the optimal way to personalize a customer's shopping experience? Can we think of novel ways to exploit the unique structure of a topic-like model to gain actionable insights? Can we test these in practice? How can we combine different sources of customer data, such as purchase, search, and click data, to better understand customer behavior? What are optimal product recommendations? These open research questions, tied to the recent rapid technical developments, foreshadow an exciting time ahead of us that paves the way for many ground-breaking developments. I feel both excited and privileged to be a part in this.

2

Model-Based Purchase Predictions for Large Assortments

This chapter is based on Bruno J.D. Jacobs, Bas Donkers, Dennis Fok (2016), “Model-Based Purchase Predictions for Large Assortments,” *Marketing Science*, 35 (3), 389–404.

2.1 Introduction

The ability to predict what a customer will purchase next is valuable in many marketing applications and this holds especially true for online retailing. Adequate predictions for the next products to be purchased enable online retailers to: implement a product recommendation system; determine the positions of products in the result of a customer's search query; optimize the collection of products to be displayed on a personalized landing page; or suggest products to complement the contents of a customer's shopping basket.

Examples of personalization in practice are Amazon's “Customers Who Bought This Item Also Bought” section, Apple's iTunes Genius and the Netflix recommendation system. There is also clear evidence that such personalized configurations influence behavior (Ghose et al., 2014, Pan et al., 2007, Salganik et al., 2006). All these applications have in common that they require a personalized selection of products out of a potentially large assortment. Ideally, the selection contains those products that are most likely to be of interest to the customer. Moreover, the selection should be relatively small as the available space to show products is often limited.

The effectiveness of personalization attempts crucially depends on the accuracy of the predictions. A complicating factor in purchase prediction is the fact that the typical online retailer sells items from a very broad assortment to an even larger customer base. Hence predictions should not only be accurate, but the prediction method should scale to large applications as well (Naik et al., 2008).

Additionally, in order to be useful in an online setting the predictions should be available in near real-time. Obtaining predictions, and updating them as new information comes in, should therefore be fast.

The typical data available at an online retailer for purchase prediction are the customer purchase histories. In some cases additional customer characteristics (e.g. demographics) are also available. However, on the product level characteristics are often absent and if such product descriptions are available, it is not obvious how to extract useful predictors from this information. In this chapter we therefore focus on predicting purchase behavior based on purchase history data, possibly complemented with customer characteristics.

Many online retailers predict a customer's next purchase using collaborative filtering algorithms, for example, by relying on counts of the co-occurrence of items in purchase history data (Jannach et al., 2010, Liu et al., 2009). In such a count-based approach a decision has to be made on how to measure the co-occurrence of items, as one can count pairs, triplets, or even higher-order product combinations. A choice for small sets of items results in information loss, i.e. purchase patterns that span many products might not be easily identified. On the other hand, for large combinations of products the matrix of co-occurrence counts becomes very sparse, resulting in predictions that are based on just a few matches in the customer base. Another challenge in collaborative filtering algorithms is incorporating customer characteristics. One possible approach is to partition the customer-base using such characteristics. However, this can only be done for a couple of variables with a limited number of levels, as otherwise sample sizes per subgroup become too small.

In contrast, model-based approaches to predict individuals' choices have a long history in marketing (Guadagni and Little, 1983, McFadden, 1986, Wagner and Taudes, 1986, Fader and Hardie, 1996) and such methods are well-suited to include customer characteristics. However, the usual implementations of these models tend to break down in the typical online retail setting, where a wide variety of products is sold to a large number of customers (Naik et al., 2008). One way to make methods more scalable is to consider only a subset of the data in terms of customers and/or products (Zanutto and Bradlow, 2006). Clearly, this is not a viable solution if the aim is to predict purchase behavior for each individual customer across the entire product assortment.

In this chapter we try to bridge the gap between retail practice and marketing academia by discussing model-based prediction methods that do work in the context of large assortments. By developing such methods we open up an avenue for future research on marketing interventions in large-scale assortments, for example on the effectiveness of product recommendations, extending the work of Bodapati (2008). Note that this would not be feasible with the heuristic, count-based approaches currently used in practice. We consider two model-based approaches. In addition, we present (an implementation of) a count-based collaborative filter and a scalable version of a discrete choice model that will serve as benchmarks. We compare the methods on their (i) heterogeneity assumptions, (ii) estimation complexity, (iii) memory requirements for real-time online predictions, and (iv) predictive performance.

The first method we present is a novel approach inspired by topic models as used in the text modeling literature. Traditionally, a topic model describes a document by relating the words in the text to latent topics. We adapt this class of models to the purchase prediction context: Words become product purchases, a document is a customer's purchase history, and a topic represents a certain preference for products in the assortment. Given that the word "topic" does not make much sense in a retailing context, we refer to topics as motivations.[1] Naturally, customers can have more than one motivation, just like a document can cover multiple topics. This idea leads to a class of models that can describe and predict customer purchase behavior in large assortments.

The most frequently used topic model is latent Dirichlet allocation (LDA) by Blei et al. (2003). This model has been used to analyze very large text corpora (Ramage et al., 2010, Mimno et al., 2012), showing that LDA provides the necessary scalability. In contrast to the text modeling literature, where documents tend to contain many words, customers often have only a couple of purchases, or they might even be entirely new to the retailer. Given such limited information per customer, we need to formally estimate the population-level a-priori probabilities of having particular motivations. This extends the text modeling implementation of LDA, where these probabilities are typically considered to be known, or at best calibrated using heuristics (Wallach et al., 2009, Asuncion et al., 2009).

To account for observed heterogeneity, we extend LDA by relating customer characteristics to the a-priori motivation probabilities. This can capture heterogeneity that is related to variables such as referrer site, demographics, or other customer characteristics. Most likely this increases the predictive power of the model, in particular for the customers with few or no observed purchases. We refer to this model as LDA-X.

The next method we consider is a mixture of Dirichlet-Multinomials (MDM) (Jain et al., 1990). MDM specifies individual-specific probability vectors that contain a customer's purchase probabilities over all products in the assortment. In

[1] While intuitively plausible, we do not claim that the actual decision process is driven by these motivations.

turn, these probability vectors follow a discrete mixture of Dirichlet distributions. MDM has previously been applied in marketing (Jain et al., 1990), but to the best of our knowledge never to a large product assortment. Although, in theory customer characteristics can also be included in MDM we will argue that the resulting model will no longer be feasible in terms of estimation complexity, given the setting of our application.

The predictive performance of LDA(-X) and MDM is compared to that of a count-based collaborative filter and a discrete choice model. We assess the predictive performance using data from an online retailer. For each method, we create customer-specific prediction sets that contain the products that are most likely to be purchased. These sets are next matched with hold-out purchase data. To gain more insight into the differences between the methods, we also consider the predictive performance for groups of customers that differ in the length of their observed purchase history. Furthermore, in a setting where repeat purchases are frequently made, e.g. fast moving consumer goods, performing well by correctly predicting frequently purchased products or repeat purchases might not be too difficult. Such recommendations might even be perceived as trivial or boring (Fleder and Hosanagar, 2009). We therefore also study the predictive performance for *unexpected* products, which we define as products that have not previously been purchased by the customer and are in the tail of the assortment.

The remainder of this article proceeds as follows: In Section 2.2 we present the methods used in this research and discuss their heterogeneity assumptions and scalability. Technical details are available in appendices. Subsequently, we apply the methods to data of an online retailer. An overview of this data is provided in Section 2.3 and the results are reported in Section 2.4. To conclude, we summarize our findings and provide directions for future research in Section 2.5.

2.2 Methods

In this section we present the prediction methods we consider in this chapter. First, we introduce two model-based prediction methods, LDA(-X) and MDM, that infer latent customer traits from purchase data. We compare these methods on their heterogeneity assumptions and estimation complexity. Next, the two benchmark methods are introduced: a set of collaborative filters (CF) and a model built on discrete choice methodology (DCM) that captures customer heterogeneity through constructed, but *observed* predictor variables.

Subsequently, all methods are compared on their suitability to update predictions in a real-time setting. Finally, we discuss how we assess the quality of predictions.

All methods share the following notation: The products from the J-dimensional assortment are numbered $j = 1,\ldots,J$. For each customer $i = 1,\ldots,I$ we observe n_i product purchases from this assortment. The purchase history of customer i is denoted by the vector $\mathbf{y}_i = \left[y_{i1},\ldots,y_{in_i}\right]$, where $y_{in} \in \{1,\ldots,J\}$ represents the n-th purchase of customer i. In addition we have customer-level characteristics coded in the K-dimensional vector $\mathbf{x}_i = [x_{i1},\ldots,x_{iK}]'$. We combine the purchase histories in $\mathbf{Y} = \{\mathbf{y}_1,\ldots,\mathbf{y}_I\}$ and the predictor variables in $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_I\}$.

2.2.1 *Latent Dirichlet allocation*

Our first model is inspired on topic models. The key idea underlying our application of these models to the context of purchase history data is that customer purchases are driven by a (small) set of latent motivations (the topics). Each motivation then drives preferences for a subset of products in the assortment, for example, a preference for eco-friendly products, for low-fat products, or for products for the sensitive skin.

In general, customers are likely to be driven by different motivations over time and even within a single purchase occasion. Additionally, the same product purchased by different customers may be driven by different underlying motivations: A movie can be purchased by a fan of the lead actor, or by a customer that is fond of the movie's genre. These features are embedded in topic models, in which customers may have multiple motivations and products may be associated with more than one motivation.

The basis for our method is latent Dirichlet allocation (LDA) introduced by Blei et al. (2003). LDA has been proven to scale to applications well beyond the dimensions of a typical online retailer. For example, it has been used to analyze over 8 million posts on Twitter that contain words from a vocabulary of more than 5 million entries (Ramage et al., 2010), or for the analysis of 1.2 million out-of-copyright books (Mimno et al., 2012). Below, we first present the details of our adaptation of LDA in the context of predicting customer purchase behavior. Next, we extend LDA by including customer-level predictor variables.

In LDA each latent motivation $m = 1,\ldots,M$ is represented by a probability vector $\boldsymbol{\phi}_m$ over the complete J-dimensional assortment. Given that a purchase is driven by motivation m, the probability of buying product j is simply ϕ_{mj}. The motivation-specific probability vectors are distributed as

$$\boldsymbol{\phi}_m|\boldsymbol{\beta} \sim \text{Dirichlet}_J(\boldsymbol{\beta}). \tag{2.1}$$

A priori there is no reason to favor one product over another in a motivation. This is reflected in the parameterization of $\boldsymbol{\beta}$, where we set each element equal to a common value β_0. This value determines whether the distribution in (2.1) tends to favor more narrow (β_0 close to zero) or more broad (large β_0) motivations (Wallach et al., 2009).

Even though each purchase is driven by a single motivation, a customer's entire purchase history may be driven by multiple motivations. This variation is described by an individual-specific discrete mixture $\boldsymbol{\theta}_i$ over the M motivations. The probability that a product purchase of customer i is driven by motivation m is then given by θ_{im}. These probabilities differ across customers and are modeled as

$$\boldsymbol{\theta}_i|\boldsymbol{\alpha} \sim \text{Dirichlet}_M(\boldsymbol{\alpha}). \tag{2.2}$$

Here, $\boldsymbol{\alpha}$ is an M-dimensional vector that captures the relevance of each motivation across the customer base. The expected value of the probability that motivation m drives a purchase equals

$$\text{E}[\theta_{im}|\boldsymbol{\alpha}] = \frac{\alpha_m}{\sum_{l=1}^{M}\alpha_l}. \tag{2.3}$$

Therefore, the larger the value of α_m, the more likely it is that a customer will make a purchase driven by motivation m.

The last step is to link motivations to actual purchases. We denote by $z_{in} \in \{1,\ldots,M\}$ the actual motivation that drives purchase y_{in}. As motivations are latent, we have to account for all possible motivations to obtain the marginal probability that customer i will purchase product j, resulting in

$$\begin{aligned}
&\Pr\left[y_{in} = j|\{\boldsymbol{\phi}_l\}_{l=1}^{M}, \boldsymbol{\theta}_i\right] \\
&= \sum_{m=1}^{M} \Pr\left[y_{in} = j|z_{in} = m, \{\boldsymbol{\phi}_l\}_{l=1}^{M}\right] \Pr[z_{in} = m|\boldsymbol{\theta}_i] \\
&= \sum_{m=1}^{M} \phi_{mj}\theta_{im}.
\end{aligned} \tag{2.4}$$

In the topic modeling literature it is common practice to determine the parameters of the Dirichlet distributions $\boldsymbol{\alpha}$ (for $\boldsymbol{\theta}_i$) and β_0 (for $\boldsymbol{\phi}_m$) by means of heuristics, rather than formally inferring their values from available data (Wallach et al., 2009), for example, imposing $\boldsymbol{\alpha} = 50/M$ (Griffiths and Steyvers, 2004) and $\beta_0 = 0.01$ (Steyvers and Griffiths, 2013), or by applying a grid search for $\boldsymbol{\alpha}$ and β_0 (Asuncion et al., 2009). These heuristics are not directly applicable in our setting as they have been designed for text modeling. Given that purchase histories tend to be much shorter than documents, we expect the LDA predictions to be more sensitive to the values of $\boldsymbol{\alpha}$ and (to a lesser degree) of β_0. We therefore extend the common LDA model and place proper prior distributions on both parameters and formally estimate $\boldsymbol{\alpha}$ and β_0 in a Bayesian setting.

We specify a log-normal distribution for α_m, that is, we define

$$\log(\alpha_m) = \gamma_m, \tag{2.5}$$

and set a normal prior for γ_m. We set the mode of the log-normal distribution equal to M^{-1}, which is within the range of values frequently used in the literature on text modeling, and place 10% of its probability mass above 1.[2] This prior specification favors $\boldsymbol{\theta}_i$-vectors that allocate the majority of the probability mass to a small number of motivations, while it still allows for more uniformly distributed $\boldsymbol{\theta}_i$-vectors. Similarly, we place a log-normal distribution on β_0 with its mode equal to 0.01 and 10% of its probability mass above 1. This specification supports $\boldsymbol{\phi}_m$-vectors where only a few products from the assortment receive significant probability mass, representing fairly specific motivations. Still, this prior is rather uninformative and broader motivations that spread the probability mass more equally over the assortment remain quite likely.

These prior specifications also allow us to easily extend LDA by including customer characteristics, coded in $\mathbf{x}_i$. Such variables are likely to improve the predictive performance of the model. We extend the log-linear specification for α_m in (2.5) to α_{im} as follows:

$$\log(\alpha_{im}) = \gamma_m + \mathbf{x}_i'\boldsymbol{\delta}_m. \tag{2.6}$$

This links customer preferences – represented by the likelihood of each of the motivations – to the additional customer-level information, resulting in LDA-X. To illustrate the effect of this specification on the distribution of $\boldsymbol{\theta}_i$ consider the expected value of θ_{im}, which gives the probability that a typical customer with characteristics $\mathbf{x}_i$ makes a purchase driven by motivation m:

$$\mathrm{E}[\theta_{im}|\boldsymbol{\alpha}_i] = \frac{\alpha_{im}}{\sum_{l=1}^{M}\alpha_{il}} = \frac{\exp(\gamma_m + \mathbf{x}_i'\boldsymbol{\delta}_m)}{\sum_{l=1}^{M}\exp(\gamma_l + \mathbf{x}_i'\boldsymbol{\delta}_l)}. \tag{2.7}$$

The $\boldsymbol{\delta}_m$ parameters capture the dependence of the probability that motivation m is used, on the customer-specific variables $\mathbf{x}_i$. The prior distribution of γ_m and $\boldsymbol{\delta}_m$ can only be sensibly determined if the level and scale of the $\mathbf{x}_i$ variables are known. We therefore standardize the customer-level variables such that they have mean zero and unit variance. Given this scale, we assume that all elements in $\boldsymbol{\delta}_m$ are normally distributed with zero mean and variance equal to 0.04. This corresponds to a prior 95% confidence interval that is approximately equal to $[-0.4, +0.4]$. Note that this prior distribution is chosen to be relatively narrow on

[2] These two conditions implicitly identify the two parameters of the log-normal distribution.

purpose, as the effect of $\boldsymbol{\delta}_m$ on α_{im} is exponential. As $\mathbf{x}_i$ is mean-centered, we use the same prior for γ_m as in LDA.

To obtain customer-specific predictive distributions, we condition on the model structure of LDA. In particular, given the model parameters $\boldsymbol{\alpha}$, β_0 and the latent purchase assignments $\mathbf{Z}$, the predictive distribution for a new purchase $\tilde{y}_{in}$ can be shown to equal (Griffiths and Steyvers, 2004):

$$\begin{aligned}
&\Pr\left[\tilde{y}_{in} = j|\mathbf{Z}, \boldsymbol{\alpha}, \beta_0, \mathbf{Y}\right] \\
&\quad= \sum_{m=1}^{M} \Pr\left[\tilde{y}_{in} = j|\tilde{z}_{in} = m, \mathbf{Z}, \beta_0, \mathbf{Y}\right] \Pr[\tilde{z}_{in} = m|\mathbf{z}_i, \boldsymbol{\alpha}] \\
&\quad= \sum_{m=1}^{M} \mathbb{E}\left[\phi_{mj}|\mathbf{Z}, \beta_0, \mathbf{Y}\right] \mathbb{E}[\theta_{im}|\mathbf{z}_i, \boldsymbol{\alpha}] \\
&\quad= \sum_{m=1}^{M} \left(\frac{\beta_0 + c_{mj}^{\mathrm{MJ}}}{J\beta_0 + \sum_{p=1}^{J} c_{mp}^{\mathrm{MJ}}}\right)\left(\frac{\alpha_m + c_{im}^{\mathrm{IM}}}{\sum_{l=1}^{M} \alpha_l + c_{il}^{\mathrm{IM}}}\right),
\end{aligned} \tag{2.8}$$

where c_{mj}^{MJ} is the number of times a purchase of product j is driven by motivation m and c_{im}^{IM} is the number of purchases made by customer i that are driven by motivation m. To obtain the predictive distribution for the LDA-X model one simply replaces $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}_i$ in (2.8).

2.2.2 *Dirichlet-Multinomial models*

The Dirichlet-Multinomial (DM) model (Jeuland et al., 1980, Goodhardt et al., 1984) is a known model-based approach to capture heterogeneity in purchase behavior. Applications of this model can be found in Grover and Srinivasan (1987), Fader (1993) and Fader and Schmittlein (1993). In this model, each customer is endowed with an individual-specific vector $\boldsymbol{\varphi}_i$ containing the purchase probabilities for each product in the J-dimensional assortment, where $\sum_{p=1}^{J} \varphi_{ip} = 1$. The probability that customer i purchases product j at a specific purchase occasion n is given by:

$$\Pr\left[y_{in} = j|\boldsymbol{\varphi}_i\right] = \varphi_{ij}. \tag{2.9}$$

Large values for the purchase probability φ_{ij} imply that customer i is likely to buy product j. In the DM model the customer-specific $\boldsymbol{\varphi}_i$-vectors are assumed to arise from a single Dirichlet distribution:

$$\boldsymbol{\varphi}_i|\boldsymbol{\beta} \sim \mathrm{Dirichlet}_J(\boldsymbol{\beta}). \tag{2.10}$$

The $\boldsymbol{\beta}$-vector describes the overall purchase behavior in the customer base: If product j is frequently purchased, β_j will have a large value relative to the other values in $\boldsymbol{\beta}$ and vice versa.

The original DM model has been extended such that the probability vectors $\boldsymbol{\varphi}_i$ originate from a finite mixture of Dirichlet distributions (Jain et al., 1990), not from a single Dirichlet distribution. This extension is known as a mixture of Dirichlet-Multinomials (MDM).

In MDM, each customer is assigned to one of M segments and each segment is characterized by its own Dirichlet distribution. Given that customer i is allocated to segment m, denoted by $s_i = m$, the customer's purchase probabilities $\boldsymbol{\varphi}_i$ are distributed as

$$\boldsymbol{\varphi}_i | s_i = m, \boldsymbol{\beta}_m \sim \text{Dirichlet}_J(\boldsymbol{\beta}_m). \tag{2.11}$$

The $\boldsymbol{\beta}_m$-vectors are segment specific, describing the distribution of the purchase probability vectors for customers in segment m. Customers are hence expected to be similar, although not identical, within a segment, but rather different across segments.

Segment membership in MDM is described by an M-dimensional Categorical distribution with probability vector $\boldsymbol{\pi}$. The element π_m gives the a-priori probability that a customer is a member of segment m, that is,

$$\Pr[s_i = m | \boldsymbol{\pi}] = \pi_m. \tag{2.12}$$

As we consider MDM within the Bayesian paradigm we also specify prior distributions over $\boldsymbol{\pi}$ and the $\boldsymbol{\beta}_m$-vectors. For $\boldsymbol{\pi}$ it is natural to favor no segment over any other a priori, therefore we use a uniform distribution over the $(M-1)$-dimensional simplex. This corresponds to an M-dimensional Dirichlet distribution, parameterized by a vector of ones. For each β_{mj} we use a log-normal prior distribution with its mode located at 0.01 and 10% of the probability mass located above 1. This specification allows for $\boldsymbol{\varphi}_i$-vectors that allow many products to be purchased with a large probability, but it favors segments of customers who purchase from a more limited subset of the assortment.

Similar to the approach in LDA, we obtain customer-specific predictive distributions of a new purchase $\tilde{y}_{in}$ conditional on the data, parameters, and segment allocations. In MDM this requires a prediction of segment membership of the customer, combined with the purchase probabilities, conditional on segment membership:

$$
\begin{aligned}
&\Pr\Big[\,\tilde{y}_{in} = j|\mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{y}_i\,\Big] \\
&\quad = \sum_{m=1}^{M} \Pr\big[\,\tilde{y}_{in} = j|s_i = m, \boldsymbol{\beta}_m, \mathbf{y}_i\,\big] \Pr\Big[\,s_i = m|\mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{y}_i\,\Big] \\
&\quad = \sum_{m=1}^{M} \mathbb{E}\big[\,\varphi_{ij}|s_i = m, \boldsymbol{\beta}_m, \mathbf{y}_i\,\big] \Pr\Big[\,s_i = m|\mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{y}_i\,\Big] \\
&\quad = \sum_{m=1}^{M} \left(\frac{\beta_{mj} + c_{ij}^{\mathrm{IJ}}}{\sum_{p=1}^{J} \beta_{mp} + c_{ip}^{\mathrm{IJ}}} \right) \Pr\Big[\,s_i = m|\mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{y}_i\,\Big],
\end{aligned}
\tag{2.13}
$$

where $\Pr\Big[\,s_i = m|\mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{y}_i\,\Big]$ is specified in Appendix 2.A (see equation (2.32)) and c_{ij}^{IJ} equals the number of times customer i has purchased product j. If i is a new customer $c_{ij}^{\mathrm{IJ}} = 0$ for all j by definition. Note that both components in (2.13) depend on the customer's purchase history, unlike LDA where only the motivation probabilities are customer specific.

2.2.3 *Model inference*

The predictive distributions specified above are conditional on the number of segments/motivations M, the model parameters, and segment/motivation allocations to customers/purchases. For a given number of M, we rely on Bayesian methodology to infer the model parameters and latent variables of the models. Direct inference on the posterior distribution is not tractable and therefore we derive Markov Chain Monte Carlo (MCMC) methods to generate samples from the posterior distribution. To be specific, we use a random walk Metropolis-Hastings within Gibbs sampler to draw samples from the target posterior distribution. The predictive distributions can then be obtained by averaging over these draws.

The full posterior of LDA(-X) is given by:

$$
p(\mathbf{Z}, \{\boldsymbol{\phi}_l\}_{l=1}^{M}, \beta_0, \{\boldsymbol{\theta}_i\}_{i=1}^{I}, \boldsymbol{\gamma}, \{\boldsymbol{\delta}_l\}_{l=1}^{M} | \mathbf{Y}, \mathbf{X}), \tag{2.14}
$$

where $\{\boldsymbol{\delta}_l\}_{l=1}^{M}$ is only relevant when customer characteristics $\mathbf{X}$ are included. Straightforward use of a Gibbs sampler for this posterior distribution is very inefficient. This is the result of a strong dependence between the latent motivation assignments $\mathbf{Z}$ on the one hand and the parameters $\boldsymbol{\phi}_m$ and $\boldsymbol{\theta}_i$ on the other hand. A Gibbs sampler would therefore require an excessive number of draws to properly explore this posterior. Instead, we take advantage of the fact that the Dirichlet distribution is the conjugate prior for a Categorical random variable. This allows us to marginalize over the $\boldsymbol{\phi}_m$ and $\boldsymbol{\theta}_i$ parameters, while retaining closed-form expressions for the conditional distributions of the other parameters in LDA. By doing so we substantially improve the mixing properties of the Gibbs sampler (Griffiths and Steyvers, 2004). Hence, we examine the so-called *collapsed* posterior distribution of LDA(-X), defined as:

$$p(\mathbf{Z}, \beta_0, \boldsymbol{\gamma}, \{\boldsymbol{\delta}_l\}_{l=1}^{M} | \mathbf{Y}, \mathbf{X}). \tag{2.15}$$

The elements of **Z** are sampled using a Gibbs sampler, while for the other parameters we implement a random walk Metropolis-Hastings sampler.

The set-up for inference in MDM is very similar to LDA(-X). The complete posterior distribution is given by:

$$p(\mathbf{s}, \{\boldsymbol{\varphi}_i\}_{i=1}^{I}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \boldsymbol{\pi} | \mathbf{Y}). \tag{2.16}$$

Again, we marginalize over the discrete distributions $\boldsymbol{\varphi}_i$ and $\boldsymbol{\pi}$, resulting in a collapsed posterior distribution of MDM:

$$p(\mathbf{s}, \{\boldsymbol{\beta}_l\}_{l=1}^{M} | \mathbf{Y}). \tag{2.17}$$

Here the segment allocations **s** can be sampled in a Gibbs step, while the $\boldsymbol{\beta}_l$ parameters require a random walk Metropolis-Hastings sampler.

LDA(-X) and MDM are both members of the general class of mixture models. This class of models is well known to be susceptible to end up in an area around a local maximum of the posterior distribution. As is common in this literature, this risk is reduced by using multiple random starts (Wedel and Kamakura, 2000, Train, 2009). For each value of M, we consider 250 different random starts. We reduce the computational burden of this approach by evaluating each random start at several intermediate steps of the estimation routine. At each step, we continue only with the best performing candidates. The performance is measured by the likelihood that results from the model's predictive distributions, averaged over purchases in a model-selection data set. This measure is closely related to the goal of predicting a new purchase as accurately as possible.

The same performance measure is also used to determine the number of motivations (for LDA(-X)) or segments (for MDM). In particular, for each model we increase the value of M until we find a decrease in the average predictive likelihood of the model-selection data.[3] More details on the estimation routines are provided in Appendix 2.A.

[3] In order to validate this approach we also consider the models for larger values of M. The predictive performance stabilized at the values obtained with the selected value of M.

2.2.4 *Comparison of LDA(-X) and MDM*

Although the structures of LDA(-X) and MDM might appear quite similar at first sight, these models differ fundamentally on various grounds. In this subsection we first discuss this difference in terms of customer heterogeneity. Next, we consider the estimation complexity of the LDA(-X) and MDM models.

Heterogeneity assumption

MDM assumes that heterogeneity in purchase behavior can be described by segmenting the customer base in groups of customers. Customers across segments are expected to be dissimilar, while customers within a segment are expected to be rather similar. Hence, similarity between customers is mainly driven by segment membership. In LDA(-X) purchase behavior is described by motivations, where each motivation represents a preference for certain products in the assortment. Heterogeneity in purchase behavior is described by customer-specific probabilities for these motivations. This leads to a model where the purchases of a single customer are driven by *multiple* motivations. Here similarity between customers is motivation specific. Customers can have very similar purchase behavior for one set of products – corresponding to a shared motivation – and be very different for a set of products that belong to another motivation.

Which heterogeneity structure fits best depends on the specific situation. If customers typically have one or very few motivations, grouping customers in segments might be beneficial. If many combinations of motivations are present, the continuous mixture of motivations in LDA(-X) will be more parsimonious. Therefore, if a retailer has many different (latent) subcategories in its product assortment, and preferences across these subcategories vary rather independently across individuals, it is likely that the heterogeneity can be specified more parsimoniously by LDA(-X).

Although MDM assumes a hard clustering of customers into segments, one will use posterior segment probabilities to eventually make predictions. This will typically lead to a form of soft clustering, where a weighted combination of different segments is used. This brings the heterogeneity structure of MDM closer to that of LDA(-X). As we observe more purchases, the posterior segment probabilities in MDM will of course become more and more extreme, and in the end this converges to strictly assigning a customer to a single segment.

Estimation complexity

The different heterogeneity assumptions underlying LDA(-X) and MDM have a large impact on estimation complexity through the number of customer-specific parameters. In MDM each customer is endowed with a distribution over the J-dimensional assortment, while in LDA(-X) a customer is described by a probability distribution over the M motivations, where M is generally much smaller than J. Even though we marginalize over these customer-specific distributions, this still affects the scalability of the models. Table 2.1 summarizes for each model the parameters that need to be sampled to infer the model structure after marginalization. We differentiate between the sampling technique required, as Gibbs steps tend to be much faster and have better mixing properties than Metropolis-Hastings steps (Damien et al., 1999).

TABLE 2.1 – *Parameters to sample in the MCMC estimation procedures across different models.*

	Gibbs sampler		Metropolis-Hastings sampler	
Model	Parameters	No. parameters	Parameters	No. parameters
MDM	$\mathbf{s}$	I	$\{\boldsymbol{\beta}_l\}_{l=1}^{M}$	$M \times J$
LDA	$\mathbf{Z}$	N	$\beta_0, \boldsymbol{\gamma}$	$1+M$
LDA-X	$\mathbf{Z}$	N	$\beta_0, \boldsymbol{\gamma}, \{\boldsymbol{\delta}_l\}_{l=1}^{M}$	$1+M\times(1+K)$

where

I: number of customers
J: assortment size
N: total number of purchases
M: number of segments/motivations
K: number of predictor variables in $\mathbf{x}_i$

In LDA(-X) we need as many motivation allocations as purchases (N in total), whereas for MDM we only need to sample one segment allocation per customer (I in total). Although the number of allocations is larger in LDA(-X), this does not imply that the total allocation in LDA(-X) is computationally more demanding. The sampling step for each motivation assignment in LDA(-X) involves only elementary arithmetic operations, while for each segment allocation in MDM we have to evaluate complex Gamma functions. It is difficult to exactly quantify the difference in computational complexity as it also depends on the (latent) structure in the data, but we anticipate that MDM will be slightly more complex for these Gibbs sampling steps.[4]

The remaining model parameters are sampled using Metropolis-Hastings steps and each of these steps is computationally demanding. For LDA we sample $1+M$ parameters and for LDA-X this increases to $1+M\times(1+K)$ parameters. These numbers are in sharp contrast to MDM in which $M\times J$ parameters are sampled. This renders MDM much more demanding in terms of estimation time,

[4] More details on the required sampling steps can be found in Appendix 2.A.

as the assortment size J is large. This is the price that has to be paid for the many degrees of freedom per customer. The number of Metropolis-Hastings steps in LDA(-X) is largely insensitive to the size of the assortment, the number of customers, and the number of purchases. In MDM, on the other hand, the number of Metropolis-Hastings steps linearly increases with the assortment size. This limits the scalability of MDM, which is why we can only extend LDA by including observed heterogeneity through $\mathbf{x}_i$.

2.2.5 *Benchmark methods*

In this section we present the two benchmark methods to which we will compare the predictive performance of LDA(-X) and MDM. The first benchmark is a collaborative filter while the second is built on standard discrete choice modeling.

Collaborative filtering

A collaborative filter is a deterministic algorithm that predicts purchases by matching customers to each other based on purchase histories. There are many possible ways to implement a collaborative filter. Details of the actual implementations used in industry are not common knowledge. Therefore, below we develop our own implementation of a collaborative filter.

Ideally, a focal customer is matched to customers who purchased the focal customer's previously purchased products and at least one additional item. However, such a matching on the complete purchase history is in general not feasible due to the curse of dimensionality; the larger the purchase history, the less likely it matches with other customers' histories.

We alleviate this curse of dimensionality by instead matching on parts of the purchase history. First, for each customer i we replace the complete purchase history vector $\mathbf{y}_i$ by the set of unique sorted subvectors of length k that can be created from $\mathbf{y}_i$. We denote this set of vectors by H_i^k. For example, for $k = 2$ a customer's purchase history is replaced by all the unique sorted pairs that can be formed using the purchase history, so $\mathbf{y}_i = [1,1,1,2,3]$ would be reduced to the set H_i^2 containing the pairs $(1,1)$, $(1,2)$, $(1,3)$, and $(2,3)$.[5] Next, for each subvector in this set we match the focal customer against all customers. If k is relatively small, this will result in many more matches compared to a matching on the complete purchase history. This solves the curse of dimensionality problem at the cost of a loss of information.

[5] The use of unique sorted pairs implies that $(1,1)$ occurs in H_i^2 only once and that H_i^2 contains the pair $(1,2)$ and not $(2,1)$.

We refer to a subvector of a customer's purchase history as a product combination, denoted by $\mathbf{h}$, and $c(\mathbf{h})$ gives the number of customers who purchased product combination $\mathbf{h}$, that is,

$$c(\mathbf{h}) = \sum_{i=1}^{I} \mathrm{I}\left[\mathbf{h} \in H_i^{\dim(\mathbf{h})}\right], \tag{2.18}$$

where $\dim(\mathbf{h})$ denotes the dimension of $\mathbf{h}$ and $\mathrm{I}[A]$ equals 1 if condition A is true and 0 otherwise. To obtain purchase predictions for customer i, using product combinations of size k, we score all products in the assortment based on their co-occurrence with each of the product combinations in H_i^k. For product combination $\mathbf{h} \in H_i^k$ the prediction score for product j equals the number of customers who purchased j and the products in $\mathbf{h}$, normalized by the sum of the score for $\mathbf{h}$ and *any* product $p = 1, \ldots, J$. This normalization ensures that each product combination $\mathbf{h} \in H_i^k$ receives the same weight, independent of the prevalence of $\mathbf{h}$ in other customers' purchase histories. The final product score is the sum of the normalized scores across all $\mathbf{h} \in H_i^k$. Formally, for combination size k, the overall score of product j for customer i equals

$$s_{ij}^k = \sum_{\mathbf{h} \in H_i^k} \frac{c(\langle \mathbf{h}, j \rangle)}{\sum_{p=1}^{J} c(\langle \mathbf{h}, p \rangle)}, \tag{2.19}$$

where the arguments between angle brackets represent a single product combination of size $k+1$.[6] Hence, to obtain product scores s_{ij}^k, by matching customers based on purchase histories that are reduced to combinations of size k, we need the summary of all purchase histories reduced to product combinations of size $k+1$. So, matching customers on pairs of products requires counts over triplets of products as input for the purchase predictions.

The product ranking for each customer is constructed by sorting the products on the product score defined above.[7] This ranking obviously depends on k. In our application we consider collaborative filters with two combination sizes, $k = 1$ and $k = 2$, denoted by CF-1, and CF-2 respectively. Using $k = 1$, customers are matched on the presence of single products in their purchase history. For $k = 2$ customers are matched on the presence of pairs of products in their purchase history. Larger product combinations are not desirable in our application, both in terms of computational feasibility and the degree of sparseness in these larger

[6] For $k > 0$ it is possible that a product combination $\mathbf{h}$ is never purchased with another product, i.e. for all p we have $\sum_{p=1}^{J} c(\langle \mathbf{h}, p \rangle) = 0$ in (2.19). If a customer's purchase history contains such a combination, we regress to a lower value of k for this customer.

[7] In the rare case that two or more products receive the same score, they are ranked according to their order in the data set, which is alphabetic.

combinations.

Discrete choice models

Random utility based multinomial choice models (Maddala, 1983, McFadden, 1986) have been extensively used in marketing to model discrete choices from a set of given alternatives. Implementing a traditional discrete choice model that directly uses purchase history data from a large assortment, however, is not feasible. Such a model would have to predict purchases for J products based on J predictor variables, where each predictor variable describes whether a product was purchased by the customer in the past, or not. This model specification would require the simultaneous estimation of $J(J-1)$ parameters, which is infeasible from a computational perspective and will also likely result in identification issues due to sparse data. Hence, traditional discrete choice models do not scale well when the number of products J becomes large.

The benchmark discrete choice model that we propose, resolves these problems by constructing the predictor variables in a smart way, enabling a huge reduction in the number of parameters to be estimated. To get there, we first review the structure of the regular logit model.

In the binary logit model, the probability that customer i purchases product j is specified by:

$$\Pr[y_{in} = j] = \frac{\exp(\theta_{ij})}{1+\exp(\theta_{ij})}.$$

Here, θ_{ij} represents the log odds of having purchased product j. Ignoring heterogeneity among customers for the moment, these odds will largely be driven by the log of the number of (unique) products purchased by customer i, denoted by u_i, and the relative attractiveness of product j. We capture the relative attractiveness of product j using the log odds of this product based on the *observed* product-purchase frequencies in the purchase data at the customer-base level. This leads to the following expression for the log odds of customer i buying product j:

$$\theta_{ij} = \alpha + \beta \log(u_i) + \gamma \log(\text{odds}_j). \tag{2.20}$$

The product ranking resulting from this specification will be the same for all customers as the product attractiveness is defined at the customer-base level, not the customer level. To obtain predictions that do differ across customers, we need to introduce heterogeneity in the model. To achieve this without resorting to a model with unobserved heterogeneity, as in LDA(-X) or in MDM, or requiring an excessive number of parameters, as in a regular choice model implementation, we construct variables at the customer-product level that characterize the attractiveness of product j for customer i using the available purchase history data.

The first step is to characterize customers based on their purchase history. We describe each customer's purchases by $\mathbf{v}_i$, a J-dimensional vector containing the proportions of each product in the customer's purchase history, with $\sum_{p=1}^{J} v_{ip} = 1$.[8] We then perform k-means clustering on these proportion vectors using M clusters. Customer heterogeneity can now be characterized by a customer's similarity with respect to each of the cluster means. We define the similarity of customer i with cluster m as:

$$w_{im} = \frac{1}{1 + ||\mathbf{v}_i - \bar{\mathbf{v}}^{(m)}||},$$

where $||\mathbf{v}_i - \bar{\mathbf{v}}^{(m)}||$ measures the Euclidean distance between customer i's proportion vector and the m-th cluster mean $\bar{\mathbf{v}}^{(m)}$.

We can now introduce customer-level heterogeneity in a parsimonious way by combining the cluster-level product attractiveness and the similarity measures w_{im}, that capture the relevance of each cluster for each customer. In particular, we can specify the log odds of customer i purchasing product j as:

$$\theta_{ij} = \alpha + \beta \log(u_i) + \sum_{m=1}^{M} \log(o_{mj})(\gamma_{1m} + \gamma_{2m} w_{im}), \tag{2.21}$$

where o_{mj} denotes the odds for product j that corresponds to the purchase proportions in cluster mean $\bar{\mathbf{v}}^{(m)}$. Note that in this model specification, the parameters are not product specific, as the relative attractiveness of each product is captured through the summary of the purchase behavior of the various clusters.[9]

Maximum Likelihood estimation of this parsimonious discrete choice model (DCM) is relatively straightforward and including the other available predictor variables is therefore feasible. To do so, we extend the specification in (2.21) to include interactions with the customer-specific predictor variables in $\mathbf{x}_i$, resulting in:

$$\begin{aligned}\theta_{ij} = {} & \alpha + \beta \log(u_i) \\ & + \sum_{m=1}^{M} \log(o_{mj}) \left(\gamma_{1m} + \gamma_{2m} w_{im} + \sum_{k=1}^{K} x_{ik} (\delta_{1km} + \delta_{2km} w_{im}) \right).\end{aligned} \tag{2.22}$$

[8] For smoothing purposes we add one pseudo observation to each customer's purchase history that is equal to the relative market shares of each product.

[9] Model specifications where the coefficients were allowed to be product specific suffered from severe identification problems in our application as the number of parameters is increased by a factor J.

2.2.6 *Real-time online predictions*

For each of the prediction methods, it is straightforward to construct a product ranking over the assortment for each individual customer. In the context of online retailing it is important to continuously update this ranking based on the customer's new purchases. Re-estimating the (population-level) parameters can be done offline after a substantial amount of new data has been collected. However, updating the predictions for a specific customer should be feasible online. This allows the retailer to update predictions while customers select products during a shopping trip. For all methods, the real-time update step itself consists of simple arithmetic operations with the details provided in Appendix 2.B. A possible bottleneck could be the amount of data that has to be available, retrieved and processed to enable the updates. In the top half of Table 2.2 we display the number of elements needed in order to update a single customer's product ranking in real-time, for each new product purchase that is observed. The bottom half of the table provides information on the amount of data that needs to be stored for the entire customer base to enable the aforementioned real-time update step.

TABLE 2.2 – *Comparison of memory requirements for real-time updating.*

No. selected data elements for each real-time update step							
Retailer context							
I	J	n_i	M	LDA(-X)	MDM	CF-2	DCM
10,000	500	10	20	$1.00 \cdot 10^4$	$1.00 \cdot 10^4$	$5.51 \cdot 10^3$	$1.01 \cdot 10^4$
100,000	5,000	20	40	$2.00 \cdot 10^5$	$2.00 \cdot 10^5$	$1.05 \cdot 10^5$	$2.00 \cdot 10^5$
1,000,000	50,000	40	80	$4.00 \cdot 10^6$	$4.00 \cdot 10^6$	$2.05 \cdot 10^6$	$4.00 \cdot 10^6$
No. stored data elements for the real-time update step							
Retailer context							
I	J	N/I	M	LDA(-X)	MDM	CF-2	DCM
10,000	500	10	20	$2.10 \cdot 10^5$	$3.10 \cdot 10^5$	$6.77 \cdot 10^7$	$1.10 \cdot 10^5$
100,000	5,000	20	40	$4.20 \cdot 10^6$	$6.20 \cdot 10^6$	$6.30 \cdot 10^{10}$	$2.20 \cdot 10^6$
1,000,000	50,000	40	80	$8.40 \cdot 10^7$	$1.24 \cdot 10^8$	$6.26 \cdot 10^{13}$	$4.40 \cdot 10^7$

where
I: number of customers
J: assortment size
N: total number of purchases
M: number of segments/motivations/clusters
n_i: number of purchases made by customer i

The first row in Table 2.2 mimics the context of our application: A medium-sized online retailer with an assortment of 500 products, 10,000 customers, and on average 10 purchases per customer. The number of segments/motivations/clusters (M) is set to 20, which is slightly larger than our empirical findings in this chapter, and we consider our implementation of a collaborative filter with combination size $k = 2$. In this context, the number of elements that have to be selected for the real-time update step is of the same order of magnitude across the prediction methods. The storage requirements, on the other hand, are of a different order of magnitude, i.e. millions for the collaborative filter versus thousands for the model-based approaches. However, for these settings all methods can easily be used in practice.

To illustrate the scalability of the various methods we increase the size of the assortment and customer base by a factor of ten and we double both the average purchase history size and M. Naturally, all memory requirements increase in this setting, but the rate of growth differs significantly. For the collaborative filter the storage requirements increase approximately by a factor of thousand, while the model-based approaches only increase by a factor of twenty. The same holds for the third context, in which we again increase the dimensions. This illustrates that the dimension reduction achieved by the model-based approaches ensures that they are suitable for real-time predictions in large scale applications, even if the number of underlying dimensions grows with the amount of available data. In addition, it is not feasible to use a combination size larger than $k = 2$ in our implementation of a collaborative filter, as in that case the storage requirements would increase even faster. For very large applications, one might even need to rely on the simpler CF-1, which only matches purchase histories on the presence of single products.[10]

2.2.7 *Performance measures*

To evaluate the methods for a range of different customization applications, we consider *prediction sets* of different sizes. A prediction set of size S contains the S highest ranked products for a customer. In case one is interested in recommending a single product, the prediction set of size 1 is most relevant. However, when customizing a page with search results the prediction set of size 10 may be more relevant. We assess the quality of a prediction set by matching its contents against hold-out purchase data. These purchases are denoted by $\mathbf{y}'_i$ for customer i and the number of unique purchased products in $\mathbf{y}'_i$ is given by u'_i.

We denote a complete ranking of all J products for customer i by the vector $\mathbf{r}_i$. The first element, r_{i1}, is the product that has the highest predicted purchase probability for the model-based rankings, the highest product score for the col-

[10] In our application, this simpler collaborative filter performs systematically worse than CF-2.

laborative filters, and the highest odds for DCM. The quality of a prediction set of size S can be measured by the number of products in the prediction set that overlap with the hold-out purchases: $\sum_{s=1}^{S} \mathrm{I}\left[r_{is} \in \mathbf{y}_i'\right]$. This number should be seen relative to the maximum number of hits possible in order to obtain a hit rate that may be compared across prediction sets of different sizes. This maximum is bounded by S, the size of the prediction set, and the number of unique hold-out purchases u_i'. Hence, the hit rate for customer i could be defined as: $\sum_{s=1}^{S} \mathrm{I}\left[r_{is} \in \mathbf{y}_i'\right] / \min\left(S, u_i'\right)$.

If a prediction set is presented to a customer in an application, such as a recommendation list, the positions within the set are also of importance (Xu and Kim, 2008). We incorporate this notion in our hit rate by weighing the hits according to their ranks. For the s-th ranked product in a prediction set of size S this weight is specified as: $w(s,S) = 1 - \frac{s-1}{S}$. Combining the above, we obtain our final performance measure, the weighted hit rate:

$$h_i(\mathbf{r}_i, S) = \frac{\sum_{s=1}^{S} \mathrm{I}\left[r_{is} \in \mathbf{y}_i'\right] w(s,S)}{\sum_{s=1}^{\min\left(S, u_i'\right)} w(s,S)}. \tag{2.23}$$

2.3 Data

We apply the prediction methods to purchase data from a medium-sized online retailer in the Netherlands.[11] The data starts at the launch of the retailing platform and it covers a period of approximately 67 weeks. The product assortment primarily consists of non-food fast-moving consumer goods, such as detergents, deodorants and shampoo. The assortment is complemented with a small selection of high turnover products for infants and toddlers, such as diapers and baby food. As a consequence, the data contains many repeat purchases.

Initially, the data contains 3,226 unique products IDs. These IDs correspond to a very fine-grained classification, e.g. different package sizes of the same product each receive a unique ID. We opt for a more coarse-grained classification and combine products on the category-brand level. For example, different fragrances of the same deodorant brand are aggregated to one category-brand combination. This approach results in a total of 440 unique category-brand combinations. Additionally, this aggregation step is applied to the customer orders: if an order contains multiple products from the same category-brand, we consider this as a single purchase from this category-brand. Finally, the category-brands that are purchased five times or fewer across all purchases are removed from the data.

[11] The authors wish to thank Christian van Someren, former Managing Director of Truus.nl, for kindly providing us this data.

Below we will simply refer to the category-brand combinations as "products". After the aggregation steps the data contains 95,208 product purchases of 394 products made by 11,783 distinct customers.

We chronologically split the data in two parts: The first 80% of the purchases are used as in-sample data, while the hold-out data comprises the last 20% of the purchases. The hold-out data is used to assess the predictive performance of the methods. This division mimics the setting of predicting future purchase behavior. Subsequently, we split the in-sample data into an estimation and a model-selection subset. We randomly select half of the customers from the in-sample data and for each of these customers, a single product purchase is randomly selected as model-selection data. The remaining in-sample data is used to estimate LDA(-X), MDM, DCM, and to create the collaborative filters. The model-selection data is used to determine the number of motivations/segments/clusters (M) in LDA(-X), MDM and DCM respectively. Table 2.3 summarizes the three subsets of the data, in terms of number of customers, unique products, and number of product purchases.

TABLE 2.3 – *Characteristics of the three subsets of the purchase data.*

Subset	Customers	Unique products	Purchases
Full data	11,783	394	95,208
Estimation data	8,831	393	71,346
Model-selection data	4,820	323	4,820
Hold-out data	3,745	369	19,042

It is quite likely that the type of customer acquired by the retailer changes over time, for example due to (a shift in) brand awareness or the mix of advertising channels that are used. Therefore, we investigate whether the customer's time of adoption at the retailer systematically shifts customer preferences. Model-free evidence for such a shift is provided in Table 2.4, which shows the purchase frequencies of the 10 most frequently purchased products in the estimation data for the first 25% of the estimation customer base, the early adopters, and similarly for the late adopters, the last 25% of the estimation customer base. The ordering of the 10 products for early versus late adopters is not only different, but the relative difference in purchase frequencies is quite substantial as well. For example, the product 'Baby/toddler nutrition – Olvarit' is purchased more than twice as often by early adopters relative to late adopters. In the tail of the assortment such relative shifts may even be larger.

TABLE 2.4 – *Purchase frequencies of the 10 products that are most frequently purchased in the estimation data by the early and late adopters, respectively.*

	Early adopters		Late adopters	
Rank	Products	%	Products	%
1	Diapers – Pampers	9.40	Diapers – Pampers	8.95
2	Baby/toddler nutrition – Nutrilon	5.20	Laundry – Ariel	4.73
3	Baby/toddler nutrition – Olvarit	4.65	Dishwashing – Dreft	3.70
4	Baby care – Zwitsal	3.65	Dental care – Oral-B	3.33
5	Laundry – Ariel	3.41	Baby care – Zwitsal	3.13
6	Paper towels – Page	3.01	Baby care – Pampers	3.00
7	Baby/toddler nutrition – Bambix	3.01	Baby/toddler nutrition – Nutrilon	2.79
8	Baby care – Pampers	2.14	Cleaning – Ambi Pur	2.04
9	Dishwashing – Dreft	2.07	Baby/toddler nutrition – Olvarit	2.02
10	Shaving – Gillette	1.95	Laundry – Lenor	1.98

This model-free evidence suggests that the predictive performance could be improved by including customers' time of adoption. We define the time of adoption as the number of days between a customer's first order, and the starting date of the retailing platform. We take the natural logarithm of this variable to allow for larger shifts in the preferences of customers acquired during the early stages of the retailing platform. Finally, this variable is standardized using the mean and variance in the in-sample data.

2.4 RESULTS

In this section we present the results of the prediction methods considered in this chapter. First, for LDA(-X), MDM, and DCM we determine M, the number of motivations, segments, and clusters respectively. Next, we focus on some details of the model results to highlight the concepts that underlie LDA(-X) and MDM. In this part we also illustrate how predictions are updated when a new purchase is observed for a customer. Finally, we compare the prediction methods by evaluating their predictive performance on the hold-out data, using the weighted hit rate.

2.4.1 *Model selection*

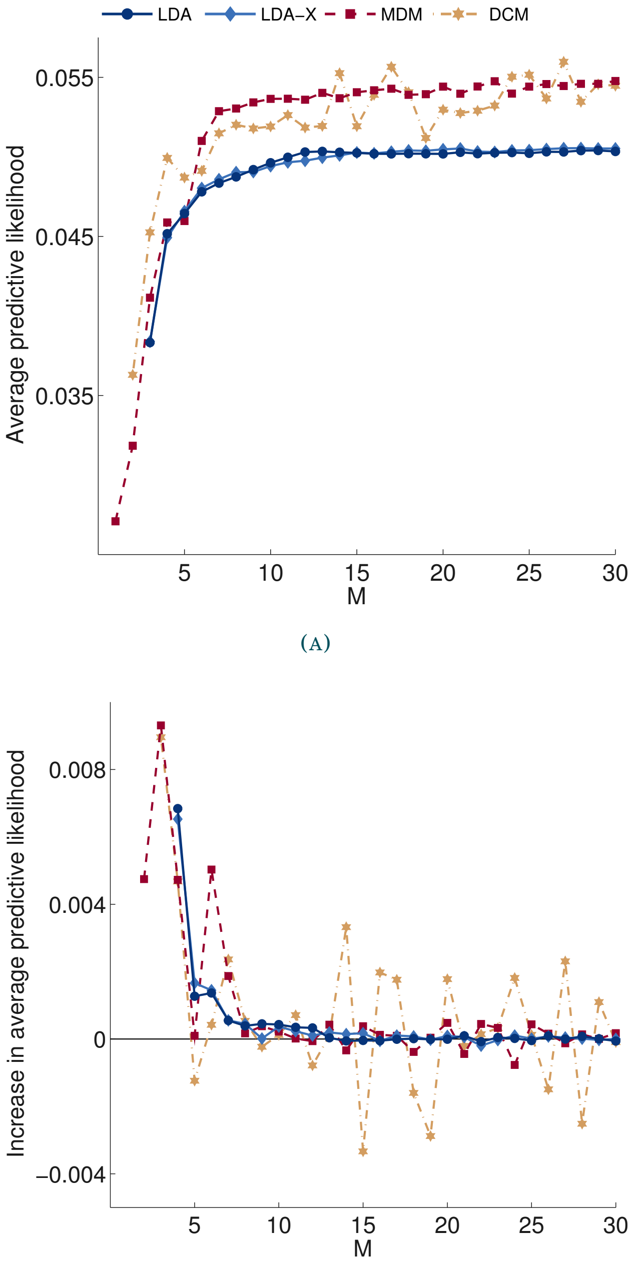
In all model-based approaches we have to determine M: the number of motivations, segments, and clusters. We evaluate LDA(-X) for $M = 3,\dots,30$ and MDM for $M = 1,\dots,30$, where MDM with $M = 1$ corresponds to the DM model. For each of these model configurations (choice of model plus a value of M) we use 250 different random starts to avoid local maxima. Throughout the estimation procedure the performance of each random start is measured by the average predictive likelihood for the model-selection data and, as discussed in Section 2.2.3, at several points during the procedure we drop the worst-performing starting values (see Appendix 2.A). At the end of the estimation routine we use the parameter estimates that result from the random start that has the highest average predictive likelihood. We evaluate DCM for $M = 2, \ldots, 30$. To avoid local maxima in the k-means algorithm used in DCM, we use 1000 different random cluster initializations. For each value of M, the clustering that obtains the lowest within cluster sum-of-squares is selected.

The average predictive likelihoods for the model-based approaches are displayed in Figure 2.1a. We find that for each method the average predictive likelihood steeply increases for the first few values of M and then levels off for larger values of M. This result indicates that choosing M too small likely impedes performance more than choosing M too large. The average predictive likelihood of LDA and LDA-X is similar, reaching a value of approximately 0.05 for the larger values of M. MDM performs slightly better, reaching a value close to 0.055. DCM performs similar and in between LDA(-X) and MDM, although its performance fluctuates across values of M. Note that the average predictive likelihood is merely an indicator for the actual predictive performance in our application, as we will consider the rank assigned to purchased products to evaluate the predictive performance and not the actual purchase likelihoods.

To determine the number of motivations and segments in LDA(-X) and MDM, we select the first value of M for which the average predictive likelihood decreases when M is increased by 1, i.e. we select the first local maximum. As the graphs in Figure 2.1a stabilize after their first local maximum, this approach results in a parsimonious, yet high performing model specification. Figure 2.1b shows the differences in performance between subsequent values of M. The first negative value – corresponding to a decrease in performance – is obtained at $M = 14$ for LDA, $M = 16$ for LDA-X, and $M = 12$ for MDM. Hence, we select $M = 13$ for LDA, $M = 15$ for LDA-X, and $M = 11$ for MDM. The average predictive likelihood is more volatile across values of M for DCM, resulting in the first local maximum for $M = 4$. In the spirit of our M selection criterion for LDA(-X) and MDM, we instead select the smallest value of M that corresponds to a local maximum in the range of the values of M where the predictive likelihood has leveled off. For DCM, this happens at $M = 14$.

2.4.2 *Model results for LDA(-X) and MDM*

Both LDA(-X) and MDM require quite a large number of parameters to capture the heterogeneity in preferences across the full assortment. For example, to characterize purchase behavior across the segments/motivations the models use $M \times J$ parameters. Clearly it does not make sense to display all these parameters.



(A)

(B)

FIGURE 2.1 – *Average predictive likelihood for the model-selection data as a function of M.*

However, as LDA(-X) and MDM approach heterogeneity in a very different way, it is interesting to consider some differences in the estimation results across the models. In MDM heterogeneity is defined at the customer-segment level, while LDA(-X) models heterogeneity through motivations, i.e. preferences for a set of coherent products, with customers differing in the strength of these motivations. To illustrate this difference, we display the 10 most likely products in the two most likely segments/motivations for each model in Table 2.5.

TABLE 2.5 – *The 10 most likely products in the two most likely motivations (LDA and LDA-X) or segments (MDM).*

LDA	Motivation 1 (Probability 0.21)		Motivation 2 (Probability 0.12)	
M=13	Product	%	Product	%
1	Diapers – Pampers	20.25	Shampoo – Andrelon	3.86
2	Baby/toddler nutrition – Nutrilon	19.13	Paper towels – Page	3.47
3	Baby/toddler nutrition – Olvarit	15.63	Laundry – Ariel	2.94
4	Baby/toddler nutrition – Bambix	9.87	Cleaning – Glorix	2.82
5	Baby care – Zwitsal	7.77	Laundry – Robijn	2.79
6	Baby care – Pampers	4.49	Conditioner – Andrelon	2.40
7	Pacifiers – Bibi	2.17	Shaving – Gillette	2.32
8	Bottle appliances – Philips AVENT	2.03	Deodorant – Dove	2.27
9	Diapers – Huggies	1.56	Baby care – Zwitsal	2.09
10	Bottle appliances – Nuby	1.21	Dishwashing – Dreft	2.05
LDA-X	Motivation 1 (Probability 0.21)		Motivation 2 (Probability 0.13)	
M=15	Product	%	Product	%
1	Diapers – Pampers	20.11	Cleaning – Glorix	5.79
2	Baby/toddler nutrition – Nutrilon	19.26	Paper towels – Page	5.37
3	Baby/toddler nutrition – Olvarit	16.04	Dishwashing – Dreft	3.78
4	Baby/toddler nutrition – Bambix	10.13	Laundry – Robijn	3.54
5	Baby care – Zwitsal	7.94	Cleaning – Ajax	3.50
6	Baby care – Pampers	4.10	Laundry – Ariel	3.27
7	Pacifiers – Bibi	2.13	Disposables – Komo	3.08
8	Bottle appliances – Philips AVENT	2.05	Paper towels – Edet	3.03
9	Diapers – Huggies	1.70	Cleaning – Sorbo	2.97
10	Bottle appliances – Nuby	1.25	Cleaning – Cif	2.29
MDM	Segment 1 (Probability 0.32)		Segment 2 (Probability 0.23)	
M=11	Product	%	Product	%
1	Diapers – Pampers	11.35	Diapers – Pampers	16.23
2	Laundry – Ariel	4.05	Baby/toddler nutrition – Nutrilon	14.34
3	Baby care – Zwitsal	4.03	Baby/toddler nutrition – Olvarit	11.76
4	Baby/toddler nutrition – Nutrilon	3.50	Baby/toddler nutrition – Bambix	7.08
5	Baby/toddler nutrition – Olvarit	3.37	Baby care – Zwitsal	6.50
6	Baby care – Pampers	3.14	Baby care – Pampers	4.05
7	Dishwashing – Dreft	3.12	Bottle appliances – Philips AVENT	1.82
8	Paper towels – Page	3.01	Laundry – Ariel	1.70
9	Dental care – Oral-B	2.60	Pacifiers – Bibi	1.64
10	Baby/toddler nutrition – Bambix	2.20	Diapers – Huggies	1.44

The top 10 most likely products are primarily baby related for the largest as well as the second largest segment in MDM. Additionally, there is much overlap at the product level: 7 products appear in both top 10 lists. For LDA and LDA-X the largest motivation relates to baby products and the order of the top 10 is the same, with only minor differences between the purchase probabilities. The second motivation for LDA-X is driven by cleaning products, while in LDA it is a mix of cleaning and personal care products (and one baby related product). So, for both LDA and LDA-X the second motivation is very different from the first, which contrasts with the results for MDM.

This difference can be explained by the distinction between a motivation and a segment. Motivations represent coherent sets of products, where customers can be interested in multiple of these sets. Segments capture the purchase behavior of groups of customers, and purchase behavior across groups likely overlaps. In other words, the motivations in LDA(-X) correspond to a clustering on the product level, whereas the segments in MDM represent a clustering on the customer level.

As the models differ substantially in terms of the underlying data structures that are captured, their predictions are also likely to be different. We investigate these differences in a hypothetical scenario. First, let us consider a customer with average customer characteristics who is new to the store, i.e. without previous observed purchases. Each model approximately yields the marginal distribution as predictive distribution for this customer. The top 5 products in the marginal distribution of the estimation data are displayed in Table 2.6.

Next suppose that the customer purchases 'Shampoo – Herbal Essences'. For each model the updated top 5, conditional on this purchase, is displayed in Table 2.7. Indeed, each model now provides a different ranking. It is interesting to focus on the new rank of the shampoo itself and the complementary conditioner of the 'Herbal Essences' brand. In the marginal distribution the shampoo and conditioner are ranked 113 and 119, respectively. Conditional on the purchase of the shampoo, these two products reach the top 5 in LDA (they get rank 3 and 2). For LDA-X the products do not occur in the top 5 but receive rank 17 and 16. Finally, in MDM the rank of the shampoo shifts to 26, while the rank of the conditioner barely changes and reaches only 117. This indicates that MDM fits the observed purchase well, but is hardly able to discover that the conditioner is a complement to the shampoo.

TABLE 2.6 – *Purchase frequencies of the 5 products that are most frequently purchased in the estimation data.*

Rank	Product	Frequency
1	Diapers – Pampers	9.20 %
2	Baby/toddler nutrition – Nutrilon	4.09 %
3	Laundry – Ariel	4.07 %
4	Dishwashing – Dreft	3.65 %
5	Baby/toddler nutrition – Olvarit	3.47 %

TABLE 2.7 – *Purchase probabilities of the 5 most likely product for each model, conditioned on the purchase of a single product.*

	Purchased product: **Shampoo – Herbal Essences**	
LDA	Product	Probability
1	Diapers – Pampers	0.05
2	Conditioner – Herbal Essences	0.05
3	Shampoo – Herbal Essences	0.05
4	Baby/toddler nutrition – Nutrilon	0.02
5	Paper towels – Page	0.02
LDA-X	Product	Probability
1	Diapers – Pampers	0.05
2	Paper towels – Page	0.04
3	Laundry – Ariel	0.03
4	Dishwashing – Dreft	0.03
5	Baby care – Zwitsal	0.03
MDM	Product	Probability
1	Diapers – Pampers	0.09
2	Baby/toddler nutrition – Nutrilon	0.04
3	Baby care – Zwitsal	0.03
4	Laundry – Ariel	0.03
5	Paper towels – Page	0.03

2.4.3 *Predictive performance*

To assess a method's predictive performance we evaluate its weighted hit rate for the hold-out data, see (2.23). In the weighted hit rate, each hit receives a weight that depends on the rank assigned to the prediction. A better (numerical lower) rank receives a larger weight than a worse (numerical higher) rank. Figure 2.2 presents the predictive performance on the complete hold-out data for the model-based approaches, LDA(-X), MDM, DCM, and the two count-based collaborative filters, CF-1, and CF-2.

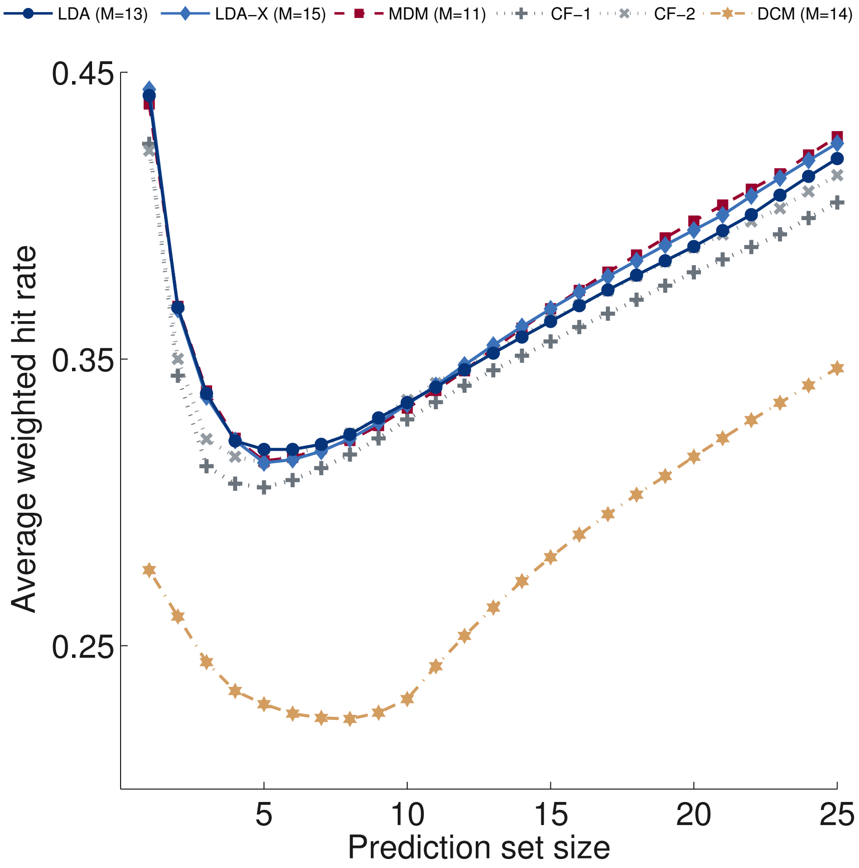


FIGURE 2.2 – *Predictive performance for the complete hold-out data, as a function of prediction set size.*

In case we predict only a single product for each customer, i.e. a prediction set of size one, LDA-X has the best performance with a hit rate close to 0.45. For most prediction set sizes, LDA(-X) and MDM outperform the collaborative filters. The best performing collaborative filter is CF-2, which matches customers on the presence of pairs of products in their purchase history. Given the decent predictive likelihoods generated by DCM (see Figure 2.1), it has an unexpected poor performance in terms of ranking the purchased products.

Note that the average hit rate declines for the first few prediction set sizes. This is a direct consequence of the denominator in the definition of the hit rate in (2.23), which divides the total number of hits by the maximum number of hits possible for a given customer and prediction set size. This number increases with the size of the prediction set until it reaches the number of unique products purchased by the customer. As the average number of unique purchases per customer in the hold-out data is almost 5, we indeed see the hit rates increase beyond that value for most methods.

We study the difference in performance for the prediction methods in more detail by separately considering specific groups of customers and products. In particular, we first divide the customers in the hold-out data into three groups based on the number of purchases in the estimation data: (i) 2185 customers with no prior observed purchases (Figure 2.3); (ii) 809 customers with a moderate amount (1-9) of purchases (Figure 2.4); and (iii) 751 customers with many (10 or more) purchases (Figure 2.5).

The most apparent difference in performance between these groups is visible in the range of the y-axis. If we observe many purchases for a customer the average hit rates are twice as large for the smaller prediction sets, compared to those for customers with no purchases in the estimation data. This is exactly according to our expectations, and provides empirical evidence that purchase history data is indeed very informative about a customer's future purchases.

By examining Figure 2.3 we see that for customers without previous purchases the collaborative filters perform very well (particularly for moderate-sized prediction sets). Note that for this specific group of customers the collaborative filters rank the products according to their market penetration in the customer base. Also for LDA and MDM there is no information that can be used to make a personalized prediction. LDA-X uses the time of adoption, although this does not seem to shift the baseline predictions a lot. Hence, the performance differences between LDA(-X) and MDM are small.

In the absence of a purchase history, the similarity of a customer to each of the M clusters, used to create predictor variables in DCM, is rather meaningless. As a result, the DCM's predictive power is low for these customers. In fact, a

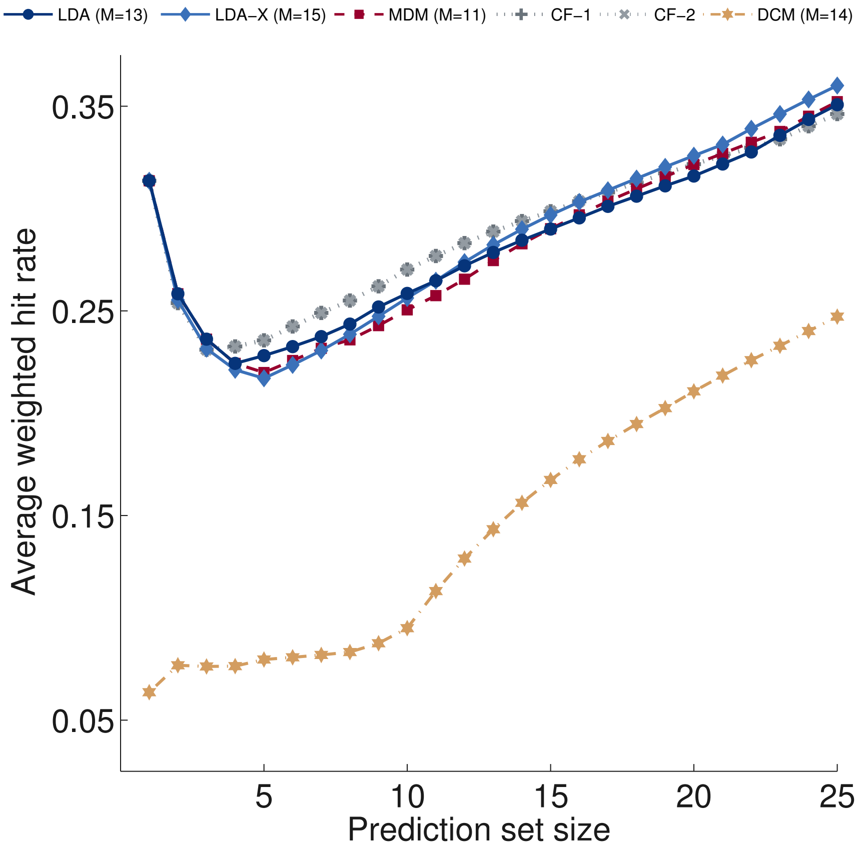


FIGURE 2.3 – *Predictive performance for the customers with no purchases in the estimation data.*

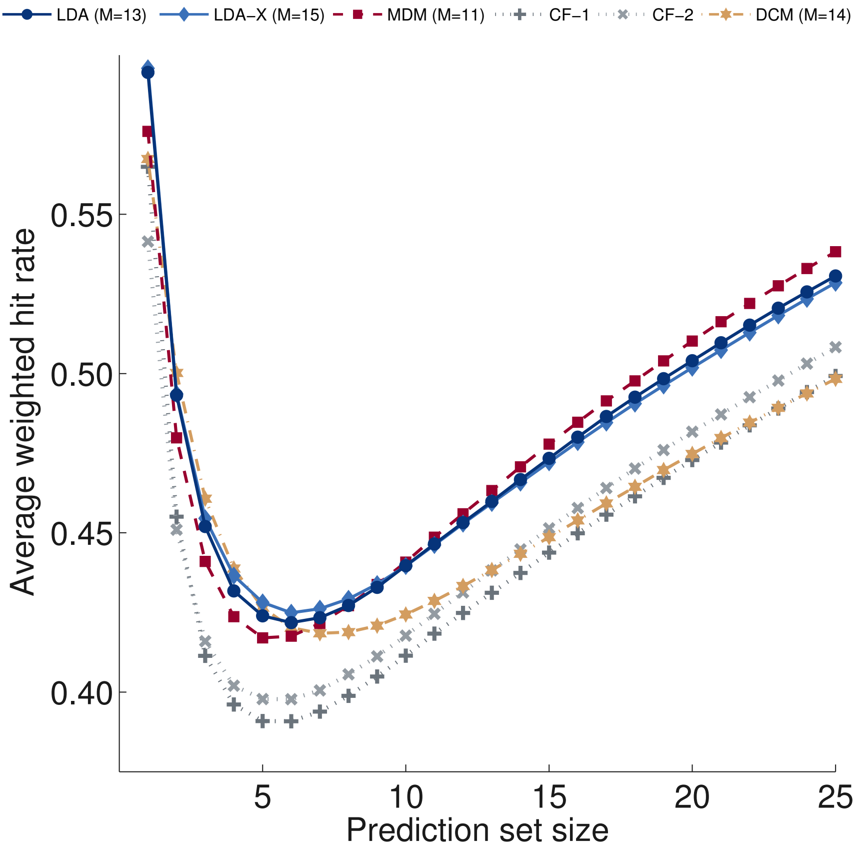


FIGURE 2.4 – *Predictive performance for the customers with a few purchases (1-9) in the estimation data.*

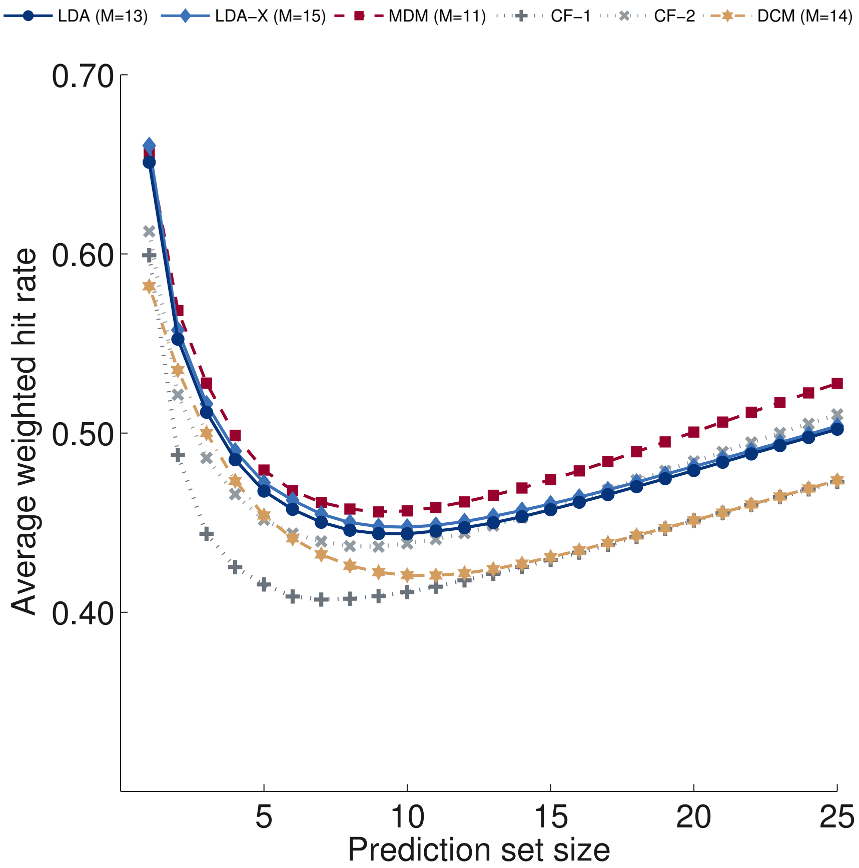


FIGURE 2.5 – *Predictive performance for the customers with many purchases (10 or more) in the estimation data.*

large part of the performance gap on the complete hold-out data between DCM and the other prediction methods is driven by the poor performance for the group of customers without a purchase history.

We observe a different pattern for customers with a moderate number of past purchases in Figure 2.4, where LDA(-X) and MDM consistently outperform the collaborative filters. This indicates that these model-based methods are better able to learn from a customer's previous purchases than the collaborative filters. Comparing the methods, LDA(-X) attains the highest overall performance and performs best when we predict only a single product, while MDM performs better for larger prediction sets. The performance of DCM is competitive for the smaller prediction sets, although its relative performance drops substantially for larger prediction set sizes.

The final group of customers that we consider consists of those who made many purchases, displayed in Figure 2.5. The general conclusion is similar to that of the customers with a moderate number of purchases. However, in this case MDM obtains the highest performance for prediction sets that contain more than one product. This result, combined with the previous findings, may be explained by the flexibility of the customer-level heterogeneity structure. In MDM preferences are modeled by a customer-specific probability vector over the product assortment. On the other hand, in LDA(-X) a customer's individual preferences are described by a lower-dimensional probability vector over the M motivations. Both models learn from previous purchases, but in MDM this learning is directly incorporated in the preferences over the assortment, while in LDA(-X) it is done indirectly through the probabilities for the motivations. As a consequence, MDM has more degrees of freedom at the level of the individual customer as the assortment size J is much larger than the number of motivations M. This additional flexibility turns out to pay off when many purchases are observed for a customer.

The results above highlight the performance of the methods for the complete assortment. However, many of the highly-ranked products are products that are frequently purchased, or products that have been previously purchased by the focal customer. Customers can easily anticipate such recommendations and might even be bored by them (Fleder and Hosanagar, 2009). It is therefore interesting to evaluate the performance of the methods when predicting products that may be more *unexpected*.

To assess the performance of the methods for predicting such unexpected products, we evaluate the predictive performance for a restricted subset of the product assortment. This subset is constructed as follows: First, we remove 20% of the products in the assortment that are most frequently purchased in the estimation data. Second, we create a customer-specific restriction by removing the products that have previously been purchased by this customer. Subsequently, for each individual customer, we only consider the predictions and hold-out purchases for products that are contained in this restricted subset of the assortment. As customers are less likely to be aware of these products, performing well on this aspect could potentially increase the *cross-selling* performance of marketing actions that are based on such predictions.

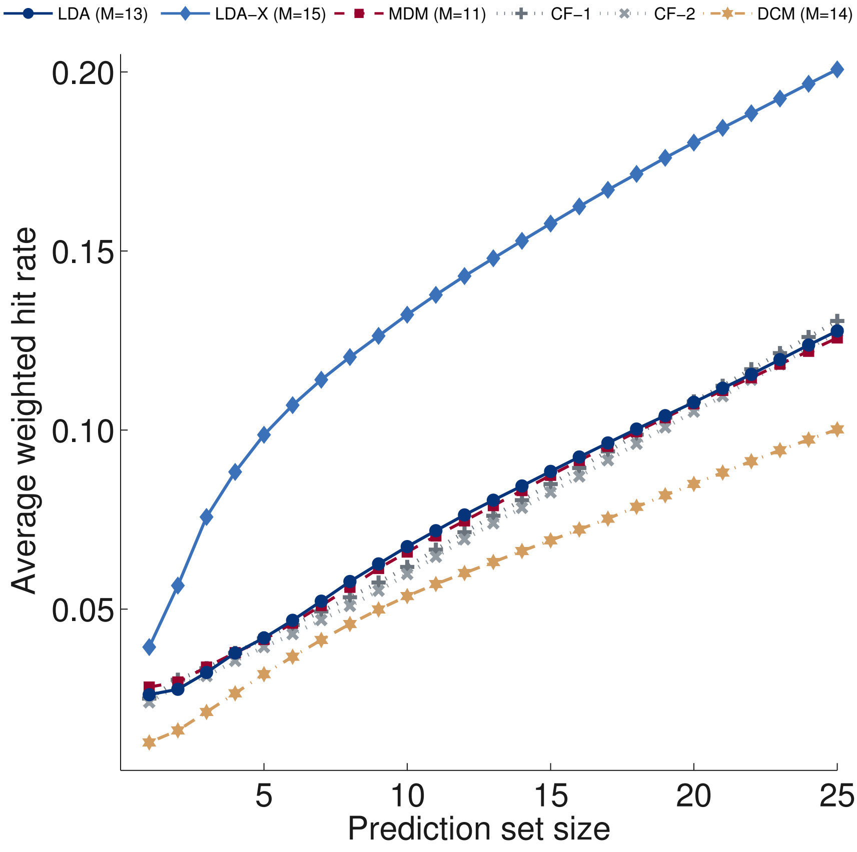


FIGURE 2.6 – *Predictive performance for the restricted subset of the hold-out data.*

The predictive performance for the restricted set of products is displayed in Figure 2.6. LDA and MDM perform better than the collaborative filters and DCM, but LDA-X clearly outperforms all the other prediction methods. This remarkable performance difference primarily arises for the highly-ranked products. By examining these products, we find that the product ‘Slimming nutrition - Weight Care’ appears in the top of many of the LDA-X customer-level prediction sets. The prediction sets resulting from the other methods, however, do not contain this product. In fact, it turns out that ‘Slimming nutrition - Weight Care’ is the most frequently purchased product in the hold-out data. Its purchase frequency has shifted from 0.04% in the estimation data to 4.88% in the hold-out data. LDA-X is able to capture this shift through the time of adoption variable.[12] This shows that the inclusion of predictor variables has merit in the context of purchase prediction, even though the time of adoption variable in general does not add much explanatory power. The reason why we do not see a similar shift for DCM can be explained by the way the predictor variables enter the model. In LDA-X, it directly influences the likelihood of a certain motivation, in effect being able to boost a motivation that is relevant for customers who adopted later in time. In this case, it boosts the motivation that contains products that are purchased more frequently later in the observation period, including the period of the hold-out predictions. In contrast, in DCM the clusters are determined ‘outside’ the model, using the k-means algorithm. The performance of the clustering algorithm does not benefit from selecting a cluster that is linked to the other prediction variables, as the predictor variables are not included when constructing the clusters. In the absence of such clusters of customers, inclusion of the predictor variables cannot shift the importance of these products, as they are not contained in a separate cluster.

2.5 Conclusion

In this chapter we have evaluated several methods for purchase prediction in large assortments using purchase history data. Inspired by the text modeling literature, we have introduced a novel model-based approach that uses latent Dirichlet allocation (LDA(-X)) to predict purchases. In addition, we have considered mixtures of Dirichlet-Multinomials (MDM), a framework well known in the brand-choice modeling literature. The performance of these model-based approaches has been contrasted against two benchmarks: a set of count-based collaborative filters, in which customers are matched on the contents of their purchase history, and a scalable implementation of a discrete choice model (DCM), that does not break down when used with a large product assortment. All methods are able to construct customer-specific product rankings over the assortment that can be used for purchase prediction.

Naturally, the prediction methods differ in their heterogeneity assumptions,

[12] We acknowledge that there can be many external influences that drive this shift in purchase behavior. Our predictor variable (time of adoption) most likely serves as a proxy for the actual causes.

estimation complexity, and memory requirements. In MDM purchase heterogeneity is specified at the customer level by segmenting the customer base. In LDA(-X), on the other hand, this heterogeneity is specified at the motivation level, which groups products, not customers. These heterogeneity assumptions also affect the estimation complexity of the models. MDM has more flexibility to model an individual customer's purchase behavior than LDA(-X), but this comes at the price of increased estimation complexity as more parameters have to be estimated. The estimation complexity of the logit part of the DCM is relatively low, but it does depend on customer clusters from an external method (i.e. the k-means algorithm). The collaborative filter has as advantage that no (latent) model structure has to be estimated, but its storage requirements for generating real-time online predictions rapidly increase for large applications. In contrast, the model-based approaches require less storage and additionally this grows much slower with the size of the application.

The performance of the methods was assessed based on purchase prediction sets derived from the product rankings, and comparing these sets to actual hold-out purchases. In general, LDA(-X) and MDM perform best and, even though these two models are conceptually rather different, their predictive performance is comparable. In addition, we have considered the setting where we focus on the predictive performance for products in the tail of the assortment that have not been purchased yet by the customer. In this case LDA-X clearly outperforms the other methods, which can be attributed to the time of adoption variable that is included in LDA-X. Although DCM also includes this predictor variable, its dependence on the k-means algorithm prevents it from effectively using the additional information to generate better predictions.

In summary the LDA(-X) prediction method that we have introduced in this chapter is the most promising approach to purchase prediction, particularly in the context of large online retailers. Its predictive performance is very competitive compared to the other methods and it scales well with the size of the application. Finally, it is a self-contained prediction method that can readily accommodate additional information available to the retailer. In our application we only had access to a fairly weak predictor, but the potential benefits of including stronger predictors of customer preferences into the model could be large.

To conclude, LDA(-X) can be readily used as a stepping stone for further model-based research that quantifies and optimizes the impact of marketing interventions in large-scale retailing environments. For example, one could optimize a recommendation system that differentiates between the likelihood of purchasing a product and the added benefits from recommending that product (Bodapati, 2008, Wagner and Taudes, 1986); something that is difficult to implement in a count-based method such as a collaborative filter. We obviously consider such extensions an interesting avenue for further research.

Appendices

2.A Estimation details for LDA(-X) and MDM

In this appendix we present the estimation details for LDA(-X) and MDM. First, we discuss our random start routine, aimed at minimizing the risk of ending up in locally optimal solutions. Second, we present the conditional posterior distributions that are used in the MCMC samplers. Finally, at the end of this appendix we provide a high-level description of the inference algorithm for LDA(-X) in pseudocode.

2.A.1 *Random start routine*

LDA(-X) and MDM are both members of the general class of mixture models. This class of models is well known to be susceptible to end up in an area around a local maximum of the posterior distribution. We reduce this risk by considering multiple random starts. For MDM a random start is an initialization of the segment assignments **s**, while in LDA(-X) it is an initialization of the motivation assignments **Z**.

For each model, we initially consider 250 different random starts. For each of these starts we draw 1,000 samples using our MCMC methodology. These samples are used to infer each customer's posterior predictive distribution and to calculate the average predictive likelihood of the model-selection data. The 50 starts that obtain the highest average predictive likelihood are selected. For these starts, we repeat the above procedure and next select the 15 best performing starts. Again, we repeat the procedure but this time draw 20,000 samples. Finally, we continue with the random start that obtains the highest average predictive likelihood.

The 22,000 draws that are generated within the random start routine for the single remaining model are considered as the burn-in period of the chain. For this selected random start we finally draw another 10,000 samples. We thin this chain by selecting every tenth draw, resulting in 1,000 posterior samples.

2.A.2 *Conditional posterior distributions*

In this section we present the details of our MCMC sampler. For each sampling step in each model, we present the corresponding conditional posterior distribution. In the presentation below we use the notation superscript $^{\setminus n}$ to indicate that the n-th element is excluded from a vector, matrix, or set. A general density function is denoted by $p()$, while we use $\pi()$ in case the density corresponds to a prior distribution in which the parameters are fixed. The probability density function of the standard normal distribution is denoted by $\phi()$ and the Gamma function is denoted by $\Gamma()$. Finally, in LDA-X we replace $\boldsymbol{\gamma}$ and $\{\boldsymbol{\delta}_l\}_{l=1}^{M}$ by $\{\boldsymbol{\alpha}_i\}_{i=1}^{I}$, whenever this simplifies notation.

As the derivations in this appendix rely on the Dirichlet-Multinomial distribution, we first provide its density in terms of Gamma functions. The Dirichlet-Multinomial distribution corresponds to a data generating process where first a probability vector $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha})$ is generated and subsequently, this vector is used to generate a set of Categorical random variables, denoted by $\mathbf{z}$. The marginal density of $\mathbf{z}$ in terms of $\boldsymbol{\alpha}$ is called the Dirichlet-Multinomial distribution. This density is given by:

$$\begin{aligned} p(\mathbf{z}|\boldsymbol{\alpha}) &= \int_{\boldsymbol{\theta}} p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta} \\ &= \frac{\Gamma\left(\sum_{l=1}^{M}\alpha_l\right)}{\Gamma\left(\sum_{l=1}^{M}\alpha_l + c_l^{\mathrm{M}}\right)} \prod_{m=1}^{M} \frac{\Gamma\left(\alpha_m + c_m^{\mathrm{M}}\right)}{\Gamma(\alpha_m)}, \end{aligned} \tag{2.24}$$

where c_m^{M} is the number of elements in $\mathbf{z}$ that are equal to m and M gives the number of categories.

LDA

The joint density for the collapsed LDA model can be written as

$$p(\mathbf{Y},\mathbf{Z},\beta_0,\boldsymbol{\alpha}) \propto p(\mathbf{Y}|\mathbf{Z},\beta_0)p(\mathbf{Z}|\boldsymbol{\alpha})\pi(\beta_0,\boldsymbol{\alpha}). \tag{2.25}$$

In our implementation of LDA we impose $\beta_0 \sim \text{logN}(\mu_{\beta_0},\sigma_{\beta_0}^2)$ and $\alpha_m \sim \text{logN}(\mu_\alpha,\sigma_\alpha^2)$. The prior distributions, combined with the LDA model specification, define the complete joint distribution in (2.25). The MCMC sampler for this model contains

Gibbs steps for all the separate elements of **Z** and Metropolis-Hastings steps for β_0 and the elements of $\boldsymbol{\alpha}$.

The conditional posterior probability that $z_{in} = m$, i.e. that the n-th purchase of customer i is driven by motivation m, is proportional to:

$$\begin{aligned}
&\Pr\left[z_{in} = m | y_{in} = j, \mathbf{Z}^{\setminus in}, \beta_0, \boldsymbol{\alpha}, \mathbf{Y}^{\setminus in}\right] \\
&\quad \propto \Pr\left[z_{in} = m | \mathbf{Z}^{\setminus in}, \boldsymbol{\alpha}\right] \Pr\left[y_{in} = j | z_{in} = m, \mathbf{Z}^{\setminus in}, \beta_0, \mathbf{Y}^{\setminus in}\right] \\
&\quad \propto \left(\alpha_m + c_{im}^{\mathrm{IM}\setminus in}\right) \frac{\beta_0 + c_{mj}^{\mathrm{MJ}\setminus in}}{J\beta_0 + \sum_{p=1}^{J} c_{mp}^{\mathrm{MJ}\setminus in}},
\end{aligned} \tag{2.26}$$

where $c_{mj}^{\mathrm{MJ}\setminus in}$ is the number of times a purchase of product j is driven by motivation m and $c_{im}^{\mathrm{IM}\setminus in}$ is the number of purchases made by customer i that are driven by motivation m, excluding z_{in} and y_{in}. This result can straightforwardly be used to obtain samples for **Z**.

The conditional posterior density of β_0 is given by

$$\begin{aligned}
p(\beta_0 | \mathbf{Z}, \boldsymbol{\alpha}, \mathbf{Y}) &\propto \pi(\beta_0) p(\mathbf{Y} | \mathbf{Z}, \beta_0) \\
&\propto \pi(\beta_0) \prod_{l=1}^{M} \frac{\Gamma\left(J\beta_0\right)}{\Gamma\left(J\beta_0 + \sum_{p=1}^{J} c_{lp}^{\mathrm{MJ}}\right)} \prod_{p=1}^{J} \frac{\Gamma\left(\beta_0 + c_{lp}^{\mathrm{MJ}}\right)}{\Gamma\left(\beta_0\right)}.
\end{aligned} \tag{2.27}$$

As (2.27) results in a non-standard density, we use a random walk Metropolis-Hastings step to obtain samples for β_0. Candidate values are generated from $\log\mathrm{N}(\beta_0, s_{\beta_0}^2)$, where β_0 denotes the current value of the parameter and the variance $s_{\beta_0}^2$ is calibrated during the start value selection procedure such that we obtain an acceptance rate of about 50%.

The conditional posterior density of α_m is

$$\begin{aligned}
p(\alpha_m | \boldsymbol{\alpha}^{\setminus m}, \mathbf{Z}, \beta_0, \mathbf{Y}) &\propto \pi(\alpha_m) p(\mathbf{Z} | \boldsymbol{\alpha}) \\
&\propto \pi(\alpha_m) \prod_{i=1}^{I} \frac{\Gamma\left(\sum_{l=1}^{M} \alpha_l\right)}{\Gamma\left(\sum_{l=1}^{M} \alpha_l + c_{il}^{\mathrm{IM}}\right)} \left(\frac{\Gamma\left(\alpha_m + c_{im}^{\mathrm{IM}}\right)}{\Gamma\left(\alpha_m\right)}\right).
\end{aligned} \tag{2.28}$$

Again this is a non-standard density and the same type of random walk Metropolis-Hastings step as before is used to obtain samples for α_m.

LDA-X

LDA-X extends LDA by allowing customer-specific predictor variables **X** to affect the motivation probabilities. The collapsed joint density for the LDA-X model can be rewritten as

$$p(\mathbf{Y},\mathbf{Z},\beta_0,\boldsymbol{\gamma},\{\boldsymbol{\delta}_l\}_{l=1}^M) \propto p(\mathbf{Y}|\mathbf{Z},\beta_0)p(\mathbf{Z}|\{\boldsymbol{\alpha}_i\}_{i=1}^I)\pi(\beta_0,\boldsymbol{\gamma},\{\boldsymbol{\delta}_l\}_{l=1}^M), \tag{2.29}$$

where $\alpha_{im} = \exp(\gamma_m + \mathbf{x}_i'\boldsymbol{\delta}_m)$, $\gamma_m \sim \mathrm{N}(\mu_\gamma,\sigma_\gamma^2)$, and $\delta_{mk} \sim \mathrm{N}(\mu_\delta,\sigma_\delta^2)$.

The MCMC sampler for LDA-X includes a Gibbs step for every element of $\mathbf{Z}$ and random walk Metropolis-Hastings steps for β_0 and all elements of $\boldsymbol{\gamma}$ and $\{\boldsymbol{\delta}_l\}_{l=1}^M$. Considering the relation $\alpha_{im} = \exp(\gamma_m + \mathbf{x}_i'\boldsymbol{\delta}_m)$, it is easy to see that we obtain the conditional posterior distributions for the elements of $\mathbf{Z}$ by writing α_{im} instead of α_m in (2.26). The conditional posterior for β_0 is exactly the same as in (2.27).

The conditional posterior density of δ_{mk} equals

$$\begin{aligned} p(\delta_{mk}|\boldsymbol{\delta}_m^{\setminus k},\mathbf{Z},\beta_0,\boldsymbol{\gamma},\{\boldsymbol{\delta}_l\}_{l\neq m},\mathbf{Y},\mathbf{X}) &\propto \pi(\delta_{mk})\prod_{i=1}^I p(\mathbf{z}_i|\boldsymbol{\alpha}_i) \\ \propto \pi(\delta_{mk})\prod_{i=1}^I \frac{\Gamma\left(\sum_{l=1}^M \alpha_{il}\right)}{\Gamma\left(\sum_{l=1}^M \alpha_{il} + c_{il}^{\mathrm{IM}}\right)} &\frac{\Gamma\left(\alpha_{im} + c_{im}^{\mathrm{IM}}\right)}{\Gamma(\alpha_{im})}, \end{aligned} \tag{2.30}$$

where δ_{mk} influences the likelihood through α_{im}. A random walk Metropolis-Hastings step in the MCMC sampler is used to obtain samples for $\{\boldsymbol{\delta}_l\}_{l=1}^M$. Candidate values are obtained from $\mathrm{N}(\delta_{mk},s_{\delta_{mk}}^2)$, where δ_{mk} denotes the current value of the parameter and the variance $s_{\delta_{mk}}^2$ is calibrated during the start value selection procedure such that we obtain an acceptance rate of about 50%. The Metropolis-Hastings sampler for γ_m can be derived in an analogous way.

MDM

The joint collapsed density for the MDM model may be rewritten as

$$p(\mathbf{Y},\mathbf{s},\{\boldsymbol{\beta}_l\}_{l=1}^M) \propto p(\mathbf{Y}|\mathbf{s},\{\boldsymbol{\beta}_l\}_{l=1}^M)\pi(\{\boldsymbol{\beta}_l\}_{l=1}^M)\int_{\boldsymbol{\pi}} p(\mathbf{s}|\boldsymbol{\pi})\pi(\boldsymbol{\pi})d\boldsymbol{\pi}, \tag{2.31}$$

where the priors are given by $\boldsymbol{\pi} \sim \mathrm{Dirichlet}(1,\ldots,1)$ and $\beta_{mj} \sim \log\mathrm{N}(\mu_\beta,\sigma_\beta^2)$. As is clear from the notation we integrate over the prior of distribution $\boldsymbol{\pi}$. The prior distributions, combined with the MDM model specification, define the complete joint distribution in (2.31). In our MCMC sampler we use separate Gibbs sampling steps for all segment assignments in $\mathbf{s}$ and Metropolis-Hastings sampling steps for the elements of $\{\boldsymbol{\beta}_l\}_{l=1}^M$.

The conditional posterior probability that $s_i = m$, i.e. that customer i is allocated to segment m, is

$$\begin{aligned}
&\Pr\left[s_i = m|\mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{Y}\right] \\
&\quad\propto \Pr\left[s_i = m|\mathbf{s}^{\setminus i}\right] p(\mathbf{y}_i|s_i = m, \boldsymbol{\beta}_m) \\
&\quad\propto \left(1 + c_m^{\mathrm{M}\setminus i}\right) \frac{\Gamma\left(\sum_{p=1}^{J} \beta_{mp}\right)}{\Gamma\left(\sum_{p=1}^{J} \beta_{mp} + c_{ip}^{\mathrm{IJ}}\right)} \prod_{p=1}^{J} \frac{\Gamma\left(\beta_{mp} + c_{ip}^{\mathrm{IJ}}\right)}{\Gamma\left(\beta_{mp}\right)},
\end{aligned} \tag{2.32}$$

where c_{ip}^{IJ} is the number of times customer i purchased product p and $c_m^{\mathrm{M}\setminus i}$ denotes the number of customers allocated to segment m, excluding customer i. Equation (2.32) implies probabilities that can straightforwardly be used to obtain samples for s_i.

The conditional posterior density of β_{mj} is given by

$$\begin{aligned}
&p(\beta_{mj}|\boldsymbol{\beta}_m^{\setminus j}, \mathbf{s}, \{\boldsymbol{\beta}_l\}_{l \neq m}, \mathbf{Y}) \\
&\quad\propto \pi(\beta_{mj}) \prod_{i=1}^{I} p(\mathbf{y}_i|s_i = m, \boldsymbol{\beta}_m)^{\mathbb{I}[s_i = m]} \\
&\quad\propto \pi(\beta_{mj}) \prod_{i=1}^{I} \left(\frac{\Gamma\left(\sum_{p=1}^{J} \beta_{mp}\right)}{\Gamma\left(\sum_{p=1}^{J} \beta_{mp} + c_{ip}^{\mathrm{IJ}}\right)} \frac{\Gamma\left(\beta_{mj} + c_{ij}^{\mathrm{IJ}}\right)}{\Gamma\left(\beta_{mj}\right)} \right)^{\mathbb{I}[s_i = m]}.
\end{aligned} \tag{2.33}$$

As (2.33) clearly results in a non-standard density we use a random walk Metropolis-Hastings step in the MCMC sampler to obtain samples for $\{\boldsymbol{\beta}_l\}_{l=1}^{M}$. Candidate values are obtained from logN($\beta_{mj}, s_{\beta_{mj}}^2$), where β_{mj} denotes the current value of the parameter and the variance $s_{\beta_{mj}}^2$ is calibrated during the start value selection procedure such that we obtain an acceptance rate of about 50%.

2.A.3 *Pseudocode for LDA(-X)*

In this section we provide pseudocode for the inference algorithm for LDA(-X). Algorithm 2.1 contains a high-level description of the inference algorithm for LDA(-X). More detailed pseudocode for our initialization procedure, sampling, and calibration of the Metropolis-Hastings proposal variances can respectively be found in Algorithms 2.2, 2.3, and 2.4. The pseudocode depends on the implementation details of our random start routine, discussed in Appendix 2.A.1, as well as the conditional posterior distributions for LDA(-X) presented in Appendix 2.A.2. The target acceptance rate for all univariate Metropolis-Hastings samplers is set to 50%.

Algorithm 2.1. Pseudocode for LDA(-X)

```
$Q$: number of estimation rounds
$K(q)$: set of random starts in round $q$
$T(q)$: number of samples to be drawn in round $q$
for each estimation round $q = 1, \dots, Q$ do
    for each random start $k$ in $K(q)$ do
        if $q$ is the first round then initialize the $k$-th random start
            INITIALIZATION (Algorithm 2.2)
        end if
        for $t = 1, \dots, T(q)$ do
            // Sample a new state from the MCMC chain
            SAMPLING (Algorithm 2.3)
            for each parameter sampled with a Metropolis-Hastings step do
                CALIBRATION (Algorithm 2.4)
            end for
        end for
        Calculate the average predictive likelihood of the model-selection
        data over the last $T(q)$ states
    end for
    if $q$ is not the last round then
        Select the random starts with the highest average predictive likelihood
        in this round for $K(q+1)$
    end if
end for
```

Algorithm 2.2. Initialization of a random start in LDA(-X)

```
$N$: number of purchases in the estimation data
procedure INITIALIZATION
    // Set the initial model parameters
    $\beta_0 = 0.01$
    if the model is LDA then
        $\alpha_m = \frac{1}{M}$ for $m = 1, \ldots, M$
    else if the model is LDA-X then
        $\gamma_m = \log \frac{1}{M}$ for $m = 1, \ldots, M$
        $\delta_{mk} = 0$ for $m = 1, \ldots, M$, $k = 1, \ldots, K$
    end if
    Set all counts $c_{im}^{\mathrm{IM}}$ and $c_{mj}^{\mathrm{MJ}}$ to zero
    // Initialize the motivation assignments $\mathbf{Z}$ in random order
    for each $n$ in random permutation of 1 to $N$ do
        Sample $z_{in}$ with a Gibbs step, using the distribution in Equation (2.26)
        Increase the corresponding elements $c_{im}^{\mathrm{IM}}$ and $c_{mj}^{\mathrm{MJ}}$ using sampled $z_{in}$ and $y_{in}$
    end for
    // Set the initial Metropolis-Hasting variances and calibration window sizes
    $s_{\beta_0}^2 = 0.1, w_{\beta_0} = 10$
    if the model is LDA then
        $s_{\alpha_m}^2 = 0.01, w_{\alpha_m} = 10$, for $m = 1, \ldots, M$
    else if the model is LDA-X then
        $s_{\gamma_m}^2 = 0.01, w_{\gamma_m} = 10$ for $m = 1, \ldots, M$
        $s_{\delta_{mk}}^2 = 0.01, w_{\delta_{mk}} = 10$ for $m = 1, \ldots, M$, $k = 1, \ldots, K$
    end if
end procedure
```

Algorithm 2.3. Sampling a new state for LDA(-X)

```
$I$: number of customers
$n_i$: number of purchases by the $i$-th customer
$M$: number of motivations
$K$: number of predictor variables in $\mathbf{x}_i$
procedure SAMPLING
    for each customer $i = 1, \ldots, I$ do
        for each datapoint $n = 1, \ldots, n_i$ do
            Decrease the corresponding elements $c_{im}^{\mathrm{IM}}$ and $c_{mj}^{\mathrm{MJ}}$ using current $z_{in}$ and $y_{in}$
            Sample $z_{in}$ with a Gibbs step, using the FCD in Equation (2.26)
            Increase the corresponding elements $c_{im}^{\mathrm{IM}}$ and $c_{mj}^{\mathrm{MJ}}$ using new $z_{in}$ and $y_{in}$
        end for
    end for
    Sample $\beta_0$ with a Metropolis-Hastings step, using the distribution in Equation (2.27)
    if the model is LDA then
        for each motivation $m = 1, \ldots, M$ do
            Sample $\alpha_m$ with a Metropolis-Hastings step, using
            the distribution in Equation (2.28)
        end for
    else if the model is LDA(-X) then
        for each motivation $m = 1, \ldots, M$ do
            Sample $\gamma_m$ with a Metropolis-Hastings step, using
            the distribution similar to Equation (2.30)
            for each predictor variable $k = 1, \ldots, K$ do
                Sample $\delta_{mk}$ with a Metropolis-Hastings step, using
                the distribution in Equation (2.30)
            end for
        end for
    end if
end procedure
```

Algorithm 2.4. Calibration of the MH-sampler proposal variance s^2

```
$n$: number of samples drawn in this calibration window
$n_A$: number of accepted samples in this calibration window
$w$: size of the calibration window
$s^2$: current proposal variance
$AR$: target acceptance rate
procedure CALIBRATION
    if $n$ is equal to $w$ then
        // Calculate the 95% confidence bounds of the Binomial($n$, $w \times AR$) distribution
        bounds = quantile function for Binomial($n$, $w \times AR$) evaluated at 0.025 and 0.975
        if $n_A$ is outside these bounds then calibrate the proposal variance
            if $n_A > w \times AR$ then the variance is increased
                $s = s \times \min\left(\sqrt{\frac{n_A}{w \times AR}}, 4\right)$
            else if $n_A < w \times AR$ then the variance is decreased
                $s = s \times \max\left(\sqrt{\frac{n_A}{w \times AR}}, \frac{1}{4}\right)$
            end if
        end if
        // Reset $n$ and $n_A$ for new calibration window and increase $w$
        $n = 0$, $n_A = 0$
        if $w < 500$ then
            $w = w + 10$
        end if
    end if
end procedure
```

2.B Real-time online predictions

Once we observe a new purchase for a customer we naturally want to update the predictions based on this new information. However, in an online setting it is not feasible to re-estimate a complete model in real-time. Instead, we update the customer-specific elements in real-time based on the new information, and fix the parameters that are specified at the customer-base level to their posterior means. Naturally, after observing new purchases for many customers, it makes sense to re-estimate the model structure including this new data. In this appendix we discuss for each of the prediction methods how the predictions may be updated in real-time and what the corresponding memory requirements are.

2.B.1 *LDA(-X)*

The predictive distribution for customer i in LDA(-X) is given in (2.8). To calculate the predictive distribution we need to evaluate two expectations: $\mathbb{E}[\,\phi_{mj}|\mathbf{Z},\beta_0,\mathbf{Y}\,]$ and $\mathbb{E}[\theta_{im}|\mathbf{z}_i,\boldsymbol{\alpha}]$. The first expectation is the expected value of the purchase probability for product j under motivation m. As this expectation is specified at the customer-base level, we fix it to its posterior mean. The second expectation is the expected value of the individual-specific discrete mixture over the M motivations. This expectation is customer-specific and hence we update it after observing a new purchase.

For this update of $\mathbb{E}[\theta_{im}|\mathbf{z}_i,\boldsymbol{\alpha}]$ we use an approximation step. First we define $\eta_{im}=\alpha_m+c^{\mathrm{IM}}_{im}$ and use the property of the Dirichlet distribution that $\mathbb{E}[\theta_{im}|\mathbf{z}_i,\boldsymbol{\alpha}]$ is proportional to η_{im}. To update η_{im}, we add the expected value of the motivation allocation of the new purchase (denoted by $\tilde{y}_{in}$) to its previous value. To be more precise, after each new purchase $\tilde{y}_{in}$ we increase η_{im} by:

$$\begin{aligned}\Delta\eta_{im} &= \Pr\Big[\,\tilde{z}_{in}=m|\tilde{y}_{in}=j,\{\boldsymbol{\phi}_l\}_{l=1}^{M},\boldsymbol{\eta}_i\,\Big]\\ &= \frac{\Pr[\,\tilde{y}_{in}=j|\tilde{z}_{in}=m,\boldsymbol{\phi}_m\,]\Pr[\,\tilde{z}_{in}=m|\boldsymbol{\eta}_i\,]}{\sum_{l=1}^{M}\Pr[\,\tilde{y}_{in}=j|\tilde{z}_{in}=l,\boldsymbol{\phi}_l\,]\Pr[\,\tilde{z}_{in}=l|\boldsymbol{\eta}_i\,]}\\ &= \frac{\phi_{mj}\eta_{im}}{\sum_{l=1}^{M}\phi_{lj}\eta_{il}},\end{aligned}\tag{2.34}$$

for $m=1,\dots,M$. Subsequent updates of the posterior mean of $\boldsymbol{\theta}_i$ can be obtained by sequentially updating the value of $\boldsymbol{\eta}_i$. This approximating update procedure provides an effective and efficient way to incorporate new information from purchases in LDA(-X).

The number of elements that have to be retrieved for an individual update step is equal to $(M\times J)+M$, namely the $\{\boldsymbol{\phi}_l\}_{l=1}^{M}$ vectors and the individual-specific

$\boldsymbol{\eta}_i$ vector. To be able to perform this step for each customer, $(M \times J) + (I \times M)$ elements have to be stored in total.

2.B.2 *MDM*

The predictive distribution for customer i in MDM is given in (2.13). The $\{\boldsymbol{\beta}_l\}_{l=1}^{M}$ vectors describe the probability distributions that correspond to the purchase behavior of the customer segments and hence, are not individual-specific. As a consequence we fix them to their posterior mean. The customer-specific purchase counts c_{ij}^{IJ} are updated straightforwardly according to the new purchase, while the current segment probabilities $\Pr\left[s_i = m | \mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{y}_i \right]$ can be updated using the recursive property of the Gamma function, i.e. $\Gamma(n+1) = \Gamma(n)n$, see equation (2.32) in Appendix 2.A.

The number of elements that have to be retrieved for an individual update step is equal to $(M \times J) + M + n_i$, namely the $\{\boldsymbol{\phi}_l\}_{l=1}^{M}$ vectors, the customer-specific segment probabilities $\Pr\left[s_i = m | \mathbf{s}^{\setminus i}, \{\boldsymbol{\beta}_l\}_{l=1}^{M}, \mathbf{y}_i \right]$, and $\mathbf{y}_i$ the purchase history of customer i. To be able to perform this step for each customer, $(M \times J) + (I \times M) + N$ elements have to be stored in total.

2.B.3 *Collaborative filters*

Suppose that a customer has n_i previously observed purchases. A new purchase made by this customer adds a maximum of $\binom{n_i}{k-1}$ product combinations to H_i^k. In order to incorporate this new information in the product ranking of customer i, we need to add for every new product combination the corresponding normalized score to s_{ij}^k (see (2.19)). Hence, this update step requires the retrieval of $\binom{n_i}{k-1}$ rows with J counts, the J current scores s_{ij}^k, and the purchase history $\mathbf{y}_i$ containing n_i purchases. This results in $\left(\binom{n_i}{k-1} \times J\right) + J + n_i$ elements to be retrieved when making an individual update.

To enable real-time updates for all customers, we have to combine each of the J products with each of the $\binom{J+k-1}{k}$ possible product combinations of size k and store the count for this combination of size $k+1$. In addition, we have to store the current scores and purchase history of each customer. In total this requires $\binom{J+k-1}{k} \times J + (I \times J) + N$ elements to be stored. Dependent on the combination of k and the dimensions of the application, storage of this information and real-time updating of predictions may or may not be feasible.

2.B.4 *Discrete choice model*

The predictive distribution for customer i in the DCM is obtained by calculating the log odds for all J products, as specified in (2.22). To calculate these log odds we need the model parameters and the cluster centroids, which are both specified at the customer-base level. Updating the customer-specific purchase history $\mathbf{y}_i$ according to the new purchase and calculating the new weights for the M customer clusters is straightforward.

The number of elements that have to be retrieved for an individual update step is equal to $(M \times J) + (2 + 2M(1+K)) + n_i$, namely the cluster means $\{\bar{\mathbf{v}}^{(l)}\}_{l=1}^{M}$, the logit parameters, and $\mathbf{y}_i$ the purchase history of customer i. To be able to perform this step for each customer, $(M \times J) + (2 + 2M(1+K)) + N$ elements have to be stored in total.

3

Introduction to Variational Inference and Extensions for Hierarchical Normals

3.1 Introduction

With the dawn of *Big Data* we are not merely confronted with data sets that are large in volume and velocity, the bigger challenge perhaps lies in the variety of the available data. We often have data on various characteristics of individual behavior. For instance, besides the purchases a customer has made at an online retailer we can receive data on search and click behavior as well. All this information can be combined to predict a customer's next purchase. Potentially such data can be supplemented by additional user-input such as product reviews. Connecting such diverse data sets in order to answer complex research questions calls for advanced statistical methods that can adequately explain and predict the behavior we observe.

A popular class of models that facilitates working with layered data are hierarchical Bayesian models. A layer in the model can reflect a (latent) layer in the data, leading to a model structure that can be interpreted in an intuitive way. The levels in the model are connected in the sense that one can serve as input to the other, hence they are *hierarchical*. Returning to the previous example, several layers can be identified: From an individual purchase, to a shopping trip, to a customer, to a population.

Over the last two decades the application of such hierarchical Bayesian models has increased significantly, especially in marketing. This rise can be attributed to three factors: i) More complex data sets have become available, ii) paired to an increase in computational power, iii) combined with the development of estimation techniques. This combination has allowed researchers to statistically infer the parameters of complex models. Previously, this was not feasible within a reasonable time frame. In particular, the development of Markov Chain Monte Carlo (MCMC) samplers (Gelfand and Smith, 1990) to infer the posterior distribution of model parameters played a key role in the advent of hierarchical Bayesian modeling. For a general review of the application of hierarchical Bayesian models we refer the reader to Gelman et al. (2013), and for a review from a marketing perspective to Rossi et al. (2012).

However, these MCMC methods have been initially developed almost two decades ago and we are currently approaching their practical limits: For example, consider that we are dealing with a model that augments data with latent variables, that is typically the case in mixture or mixed-membership models, as well as models for limited dependent variables. Such a model, tied to a vast data set with many data points, rapidly becomes unwieldy to infer using traditional inference techniques, at least not in a reasonable time frame (Kucukelbir et al., 2017). This problem is even further exacerbated in case we need to use and update the model in real-time, for example to make online predictions. In addition, the draws from an MCMC chain can be highly autocorrelated, especially if one employs a Metropolis-Hastings sampler (Hastings, 1970).

In addition, if there are multiple sources of data that can potentially be connected, it may be difficult to tell how the different components in the model should interact. From the example in the opening paragraph: Should we include the available product reviews in our model? If so, in what way? What about product descriptions? A priori, these questions are difficult to answer and will primarily be based on a researcher's domain knowledge and intuition. In practice one will resort to an iterative approach where different model specifications are tested before one arrives at a specification that is most suitable to answer the research question at hand (Blei, 2014). It can significantly impede the research process if a researcher has to wait days, or even weeks, for each iteration of this approach to be completed.

In these new large-scale data environments the traditional MCMC methods are slowly but surely being replaced by faster alternatives. Most notable are Hamiltonian Monte Carlo (HMC) samplers (Neal, 1996, Hoffman and Gelman, 2014), or inference via optimization through variational inference (VI) (Jordan et al., 1999, Blei et al., 2017).

HMC uses gradient information to construct a sampler that can more effectively explore the parameter space compared to traditional MCMC that does not use any gradient information. Similar to MCMC, HMC enjoys the asymptotic property that the stationary distribution of its Markov Chain is the posterior of interest. The emergence of HMC is in part driven by the development of Stan (Carpenter et al., 2017), a programming language specifically designed to infer parameters of probabilistic models. The workhorse underlying it is the No-U-Turn (NUTS) Sampler (Hoffman and Gelman, 2014), an efficient implementation of

Hamiltonian Monte Carlo (HMC). One of NUTS's most effective features is that it takes away the need for users to manually specify nuisance parameters that are involved with running an HMC sampler. A disadvantage of the NUTS sampler is that it cannot be directly applied to models that contain discrete variables, as a gradient for such variables is undefined (Hoffman and Gelman, 2014).

VI on the other hand provides a deterministic alternative to these sampling-based approaches as it casts probabilistic inference into a non-stochastic optimization problem (Blei et al., 2017), that can be solved using proved and honed optimization techniques. In contrast to conventional optimization problems, e.g. where we optimize over scalars or vectors, we optimize in VI over functions or te be more precise, over probability distributions. Such optimizations are rooted in a branch of mathematics called the *calculus of variations*, the namesake of VI. The promise of VI is that we can solve this optimization problem much faster than we can do inference using sampling based methods (Kucukelbir et al., 2017). However, this speed-up does come at a trade-off as we are only able to solve the optimization problem in a short amount of time if some restrictions are imposed on the optimization. The most commonly used restriction in practice is the *mean-field assumption*, which will be discussed in more detail in Section 3.2. This turns VI into an approximate inference method.

In this chapter an overview of variational inference is provided and its usefulness as an alternative to the sampling-based inference methods is discussed. The usage of variational inference comes with an initial cost: It enables fast inference in large models, but also requires additional work as the optimization problem needs to be solved. Typically, this requires deriving problem-specific gradients. These derivations can be both time consuming and error prone. This could hamper the exploration of different model specifications which is especially relevant in case one is dealing with varied data sets that need to be connected to each other in a way that is not known a priori. A potential solution is to rely on a technique called *automatic differentiation* (Baydin et al., 2015). It enables the automatic computation of gradients for arbitrary user-specified functions. The results are accurate to machine precision and have succeeded numerical differentiation techniques such as finite differences. A small disadvantage of automatic differentiation is that its execution time can be somewhat slower than manually derived and optimized code.

The focus will be directed towards hierarchical Bayesian linear models that contain so-called hierarchical Normals. In the literature this is also known as a hierarchical linear model (Gelman and Hill, 2007), however we want to place emphasis on the assumption of Normally distributed variables as it plays a key role in the derivations later in this chapter. A hierarchical Normal is defined as a

Normal distribution where the mean parameter is specified as a function of (one or more) other random variables that each follow a Normal distribution. In fact, each of these variables can even be a hierarchical Normal. Let us illustrate this by hand of an example: The preferences of a customer at a certain moment in time may be a combination of the customer's baseline preferences combined with a time-dependent effect (e.g. the season). If we let the time-specific preferences be Normally distributed and additionally place Normal distributions on the baseline customer-specific preferences and time-dependent effects, we have created a hierarchical Normal model.

Note that these hierarchical Normals can be part of a larger Bayesian model, for example if the output of the hierarchical Normals serves as a prior parameter for another model component. These model components do not necessarily need to follow a Normal distribution. This effectively creates additional layers in the hierarchical model, enabling the design of complex model structures. To return to our example: The customer's time-specific preferences may be used in a model to explain the purchases made by the customer in that specific time period.

For such hierarchical Normals a new result that funnels all the dependencies of the variables in the mean-specification through a common error term, will be presented. This result allows us to be flexible with the exact specification of the hierarchical Normal and therefore does not require application-specific derivations. Effectively, if we change the mean-specification of a variable that is part of a hierarchical Normal, we only need to update the common error definition to reflect this change in the other parameters. As the error specification of a Normal is of a standard and well-known form, this practically enables us to derive a variational inference algorithm that is generic and does not require manual derivations when the mean-specification changes. This is an important result, especially for researchers who want to explore different model specifications involving hierarchical Normals in variational inference.

Additionally, we derive a result for a set of variables that follow a multivariate Normal distribution with a common precision matrix that we want to infer with variational inference under the mean-field assumption. We show that by approximating each of the multivariate Normal variables with a set of independent Normals as variational distribution, i.e. effectively a multivariate Normal with a diagonal covariance matrix, the common precision matrix can still be approximated. In addition, we illustrate that the optimal solution for these set of independent normals has a simple analytical closed-form solution with an intuitive interpretation. This enables us to estimate the common covariance structure of the multivariate Normally distributed parameters at a fraction of the computational cost we would have incurred, had we specified a full covariance matrix in the variational distribution for each multivariate Normal variable.

The remainder of this chapter is set-up as follows: In Section 3.2 we introduce the reader to variational inference and its basic principles. In addition, we briefly review the exponential family and its convenient properties when applied in conjunction with variational inference. We derive our results for the hierarchical Normals in Section 3.3, while the results for the efficient estimation of a common precision matrix for a set of multivariate Normals are discussed in Section 3.4. Finally, we conclude in Section 3.5 with recommendations and reflections on how our results fit in the existing variational inference literature. Throughout the chapter we provide illustrations using small examples.

3.2 Introduction to variational inference

In this section the key concepts of variational inference (VI), a deterministic probabilistic inference technique that can be used to estimate model parameters, is reviewed. This is not merely a list of definitions of the concepts of VI as throughout the section we will provide additional intuition for the concepts introduced. This will further enhance the understanding of VI, especially for researchers with experience in estimating hierarchical Bayesian models via MCMC methods.

We start with introducing the goal of variational inference and show how it can be used to cast the problem of Bayesian inference in to a formal optimization problem. Next we discuss the mean-field assumption, which is the most commonly used assumption in the literature to restrict the variational distribution so that it is feasible to work with. After that we briefly review Markov Blankets and some properties of the exponential family, and show their intuitive connection to mean-field variational inference. This section is concluded with a short discussion of conditionally conjugate models in the context of mean-field VI, building on features of the exponential family and Markov Blankets.

3.2.1 *The problem of probabilistic inference*

Let y be the set of variables that we observe and consider that a generative model for y is specified, also known as the data generating process (DGP). Typically, this DGP contains unknown quantities that may range from latent variables to parameters of a probability distribution. We denote the collection of all these unknowns by $z = \{z_j\}_{j=1}^{J}$, where each z_j can refer to a univariate or multivariate variable.

The goal of Bayesian inference is to make statements about the distribution of the unknown variables z given y, the observed data. That is, we need to infer the *posterior* distribution of our model defined as:

$$\begin{aligned} p(z|y) &\triangleq \frac{p(y|z)p(z)}{p(y)} \\ &\overset{z}{\propto} p(y|z)p(z) = p(y,z). \end{aligned} \tag{3.1}$$

Let us clarify some of the used notation: $\triangleq$ means "defined as" and $\overset{z}{\propto}$ indicates proportionality relative to the variable z. Hence, the posterior $p(z|y)$ is proportional to the joint $p(y,z)$ where the proportionality is taken with respect to z.

Directly evaluating the posterior is intractable because of the denominator $p(y)$. This is the marginal likelihood of the data, and in principle it can be obtained by marginalizing over the entire latent model structure:

$$p(y) = \int_z p(y,z)dz. \tag{3.2}$$

Two challenges arise when one wants to evaluate this integral: i) An analytical closed-form solution for the integral may not be available; ii) In case an analytical solution exists, the elements in z are often coupled under the joint distribution. As a result, if the number of elements in z increases, the complexity of evaluating this integral grows exponentially. Hence, in practice it is intractable to evaluate this integral over z for all but the simplest models. That is, to examine the posterior one cannot rely on exact inference methods and instead we need to turn to approximate inference methods. We remark that even though MCMC enjoys the property of asymptotic exactness, in practice only a limited number of samples can be drawn from the chain and hence, its results are an approximation as well. In the next section the reader is introduced to an alternative approximate technique, called variational inference (VI).

3.2.2 *The objective of VI*

The goal of VI is to find a distribution that best approximates $p(z|y)$, the posterior of interest. This approximation is called the variational (distribution) and will be denoted by $q(z)$.[1] Note that just as $p(z|y)$, $q(z)$ is a proper joint density function over all elements in z. The quality of the fit of the variational to the posterior is

[1] A small remark on the notation used: q can either refer to the actual probability distribution or the corresponding probability density function. From the context it should be clear which of the two we are referring to.

measured by the Kullback-Leibler (KL) divergence, defined as:

$$\begin{aligned} KL[q(z)||p(z|y)] &\triangleq E_q[\log q(z) - \log p(z|y)] \\ &= \int_z q(z)[\log q(z) - \log p(z|y)]dz. \end{aligned} \tag{3.3}$$

In words, the KL-divergence from $q(z)$ to $p(z|y)$ is defined as the expectation of the difference between the log variational density and the log posterior density, where the expected value is calculated under the variational distribution. From here on the aforementioned expectation will be called the *variational expectation* and denoted by E_q.

By definition, the KL-divergence is always non-negative and zero if and only if the two distributions are equal, i.e. if $q(z) = p(z|y)$ (Kullback and Leibler, 1951), where y is fixed. Hence, a value for $KL[q(z)||p(z|y)]$ that is closer to zero indicates a better fit, while larger values that are further away from zero suggest a worse fit. Note that the KL-divergence is not a proper distance measure in the sense that it is asymmetric: The KL-divergence from $q(z)$ to $p(z|y)$ as defined above in (3.3), is generally not the same as the KL-divergence from $p(z|y)$ to $q(z)$. An alternative inference technique that focuses on the minimization of this latter KL-divergence is Expectation Propagation (Minka, 2001).

For more intuition behind the mechanics of the variational KL-divergence defined in (3.3), we note that the integrand $\log q(z) - \log p(z|y)$ is equivalent to $\log \frac{q(z)}{p(z|y)}$. Consider a given value of z: This ratio will be large if the density of $q(z)$ is relatively high compared to the posterior density $p(z|y)$. Moreover as $q(z)$ is (relatively) high in this setting, this ratio will have a large weight in the computation of the integral. This results in higher values of the KL-divergence, which indicates a worse fit. The contribution equals zero if both densities are equal. To summarize, in order to obtain a low KL-divergence, and hence a good fit, the variational distribution $q(z)$ should refrain from placing significant amounts of mass on configurations for z that are unlikely under the posterior $p(z|y)$.

More formally, the objective of our optimization is defined as finding $q^\star(z)$, the variational distribution that minimizes the KL-divergence in (3.3):

$$q^\star(z) = \underset{q(z)}{\arg\min} KL[q(z)||p(z|y)]. \tag{3.4}$$

Explicitly, this casts our inference problem in to an optimization by considering the KL-divergence from $q(z)$ to $p(z|y)$, which we want to minimize with respect to the variational distribution $q(z)$. In principle this optimization includes the type of distribution as well as the variational parameter corresponding to that distribution.

3.2.3 *The ELBO: log evidence lower bound*

A measure closely related to the KL-divergence is the ELBO. It is defined as:

$$\begin{aligned} ELBO &\triangleq E_q[\log p(y,z) - \log q(z)] \\ &= \int_z q(z)[\log p(y,z) - \log q(z)]dz. \end{aligned} \tag{3.5}$$

An important property of the ELBO is that it is a lower bound for $\log p(y)$, the log marginal likelihood. In fact, ELBO is the abbreviation of log Evidence Lower BOund. This lower bound property can be verified by rewriting the ELBO as follows:

$$\begin{aligned} ELBO &= E_q[\log p(y,z) - \log q(z)] \\ &= \log p(y) + E_q[\log p(z|y) - \log q(z)] \\ &= \log p(y) - KL[q(z)||p(z|y)], \end{aligned} \tag{3.6}$$

where we can take $\log p(y)$ out of the expectation operator as y is constant with respect to the variational distribution. Using this result it is straightforward to verify that the ELBO forms a lower bound for $\log p(y)$: The KL-divergence between two proper distributions is non-negative by definition, so it holds that $ELBO \leq \log p(y)$ for all proper $q(z)$.

In the special case that the variational is equal to the posterior, the KL-divergence between them will be zero and the corresponding ELBO is equal to the log marginal likelihood of y. However, typically the posterior is unknown and is the distribution that we actually want to infer. Hence, in practice the KL-divergence is positive and the ELBO will form a true lower bound for the log marginal likelihood of y.

Additionally, (3.6) shows that the ELBO and the KL-divergence are closely related:

$$ELBO + KL[q(z)||p(z|y)] = \log p(y). \tag{3.7}$$

By combining this result with the fact that the ELBO is a lower bound for $\log p(y)$ it follows that finding the $q(z)$ that minimizes $KL[q(z)||p(z|y)]$ is equivalent to finding the $q(z)$ that maximizes the ELBO. The convention in the (machine learning) literature is to consider this maximization of the ELBO as the objective in variational inference. In this chapter we will follow this convention as well, but it is important to reiterate that the maximization of the ELBO is equivalent to the minimization of the KL-divergence between the variational and the posterior of interest.

Besides the definition of the ELBO provided in (3.5) we will consider some alternative representations that can further aid in the understanding of the goal of variational inference, especially if the reader is trained in (Bayesian) statistics but less familiar with the machine learning literature.

The first alternative ELBO representation that is considered is:

$$\begin{aligned} ELBO &\triangleq E_q[\log p(y,z) - \log q(z)] \\ &= E_q[\log p(y,z)] + E_q[-\log q(z)] \\ &= E_q[\log p(y,z)] + \text{Entropy}[q(z)]. \end{aligned} \tag{3.8}$$

This illustrates that large (better) values for the ELBO are obtained if two (conflicting) criteria are satisfied. The variational expectation of the log joint density of the probability model should be high. In words, the variational $q(z)$ should allocate its density to configurations of z where the log joint density of the model is high. This is an intuitive result as the variational is an approximation of the posterior and in turn, the posterior is proportional to the joint density. The second term is the entropy of $q(z)$. This shows that larger values for the ELBO are obtained for a variational with a high entropy, i.e. a distribution that spreads it density over different configurations for z. Alternatively, very narrow variational distributions that resemble point masses will be penalized. This is in line with the Bayesian paradigm, where one is often not only interested in a point estimate of an unknown parameter, but also seeks a way to quantify its uncertainty.

Further insight can be obtained if one ignores the entropy in (3.8) for a moment. In this case, the variational distribution that maximizes the ELBO will maximize the variational expectation of the log joint density. The optimal $q(z)$ is in this case a point mass on the mode of $p(y,z)$, which is equivalent to the mode of the posterior $p(z|y)$ because y is fixed, cf. (3.1). This insight connects variational inference to maximum a posteriori estimation.

The second alternative representation of the ELBO we consider is given by:

$$\begin{aligned} ELBO &\triangleq E_q[\log p(y,z) - \log q(z)] \\ &= E_q[\log p(y|z) + \log p(z) - \log q(z)] \\ &= E_q[\log p(y|z)] - KL[q(z)||p(z)]. \end{aligned} \tag{3.9}$$

The first term is the variational expectation of the log-likelihood and will be high if the variational distribution places much density on areas where the log-likelihood is high. Intuitively, to obtain large values for the ELBO this first term “pulls” the variational towards the data (likelihood) of the model, favoring configurations for z for which the data is likely. The second term is the negative

KL-divergence between the variational distribution and the prior density of z in our model. This divergence reduces the value of the ELBO if the variational distribution diverges too far from the prior, unless this is compensated by an increase of the variational expectation of the log-likelihood. These two insights combined reveal that variational inference is closely related to the fundamental idea of Bayesian inference, where the posterior is defined as a combination of the likelihood and the prior.

If one ignores the KL-divergence in (3.9), the variational distribution that maximizes the ELBO will maximize the variational expectation of the log-likelihood. This corresponds to the setting of an improper uniform prior. It will result in the optimal $q(z)$ being a point mass on the maximum likelihood estimator for $p(y|z)$, connecting variational inference to maximum likelihood estimation.

In this brief introduction on variational inference we have reformulated Bayesian inference as an optimization problem. One of the promised advantages of variational inference is that it is fast and that it scales well to large data sets and complex model structures. However, the observant reader would point out that thus far we have not made our problem easier, instead we have made it more difficult: The introduced optimization problem seems no less tractable than our initial inference problem, in particular because the optimal variational distribution is equal to our posterior of interest, which is precisely the object we want to infer.

As we are now in the domain of optimization we can consider solving a closely related but less complicated optimization problem as a proxy. A natural way to reduce the complexity is by imposing restrictions on the variational distribution, although this might exclude the true posterior distribution from the set of allowed distributions. Naturally, this will lead to an approximation of the probabilistic quantities of interest. However, this is offset by much faster inference. A generic set of restrictions that facilitates faster inference will be the topic of the next section.

3.2.4 *The mean-field assumption*

In the previous section we introduced the ideas underlying variational inference and we illustrated how probabilistic inference can be viewed as an optimization problem. In case no restrictions are imposed on this optimization problem, it follows that the variational objective function is optimized if the variational distribution is equal to the posterior distribution. Clearly, this is not a feasible solution if we are not able to evaluate the posterior. That is, in practice restrictions need to be imposed on the variational, so that it remains tractable while at the same time providing a good approximation to the posterior. The most commonly used restriction in the literature is to impose the variational factorizes over the unknown variables z. This is called the mean-field assumption which leads to mean-field variational inference (MFVI):

$$q(z) = \prod_j q(z_j). \tag{3.10}$$

The mean-field assumption specifies that the variational $q(z)$ factorizes over each of the unknown elements z. Each element z_j is endowed with its own marginal density $q(z_j)$ that is independent of all other elements, denoted by $z_{\neg j}$. Recall that z_j can refer to either a univariate or multivariate variable and if it is multivariate, it is not split by the mean-field assumption.

Put differently, the unknown (sets of) parameters in our model are assumed to be uncorrelated under the variational distribution. Note that this does not imply that correlations between parameters in the model are ignored. The posterior of interest is unaltered, as the original data generating process is not affected by the mean-field assumption. Furthermore, the goal of variational inference remains the same: To find the variational that is as close as possible to our posterior of interest. The only thing that has changed in MFVI is that the form of the variational distribution is now restricted. As a result of this restriction, the variational distribution is not informative on posterior correlations between z_j and $z_{\neg j}$.

We apply the ideas behind mean-field variational inference to a toy example to make these concepts more concrete. This example is separated from the main text using gray boxes and we will continue with this example throughout the next sections.

Toy example

Let $y_1,\ldots,y_N$ be a set of N independent and identically distributed univariate random variables. Each y_i is Normally distributed with mean μ and precision τ. The prior of μ is again a Normal distribution with known parameters m_μ and t_μ. The prior of τ is a Gamma distribution with known parameters a_τ and b_τ.

We want to infer μ and τ using mean-field variational inference. The joint posterior of μ and τ is denoted by $p(\mu,\tau|y_{1:N})$, where we omit the explicit conditioning on the prior parameters which are assumed to be known. This posterior is approximated by a variational distribution that is restricted using the mean-field assumption:

$$q(\mu,\tau) = q(\mu)q(\tau).$$

Throughout the section the focus will be on the inference of μ, but the results that are derived can be applied for the estimation of τ as well.

Markov Blanket

Before we continue with the implications of the mean-field assumption we make a side-step to introduce the concept of a *Markov Blanket*. The fundamental idea here is that a model's joint probability distribution can be represented by a directed a-cyclical graph (DAG) (Pearl, 1988). In such a graph, variables in the model are represented by nodes and the interactions between variables are represented by vertices between the nodes. This facilitates reasoning about the (in)dependencies among variables contained in the model. A comprehensive overview of DAG's applied to probabilistic models and its properties can be found in Pearl (1988) and Pearl (2009). From this literature we direct our attention to the Markov Blanket. For parameter z_j we denote it by MB_j and it is comprised of three sets of nodes:

1. The parents of z_j, denoted by the set PA_j. It is defined as the parameters that affect z_j in the DGP. In the DAG this relation is displayed by arrows pointing towards z_j.

2. The children of z_j, denoted by the set CH_j. It is defined as the parameters that are affected by z_j in the DGP. In the DAG this relation is displayed by arrows pointing away from z_j.

3. The co-parents of all children of z_j, denoted by the set CP_j. It is defined as the parameters that affect the children of z_j in the DGP. In the DAG this relation is displayed by arrows pointing towards the children of z_j.

We visualize the Markov Blanket for a general case in Figure 3.1.

A key property is that the nodes in a Markov Blanket contain all the information in the model about the focal variable. That is, if we condition on the parameters in MB_j, the Markov blanket for z_j, we have all the information about z_j. We do not gain any extra knowledge of z_j by conditioning on additional nodes outside of z_j's Markov Blanket. More formally, this property can be stated as follows: The density of z_j, conditioned on all other variables in the model, is equivalent to the density of z_j conditioned on only the nodes in its Markov Blanket:

$$p(z_j|z_{\neg j}) = p(z_j|MB_j). \tag{3.11}$$

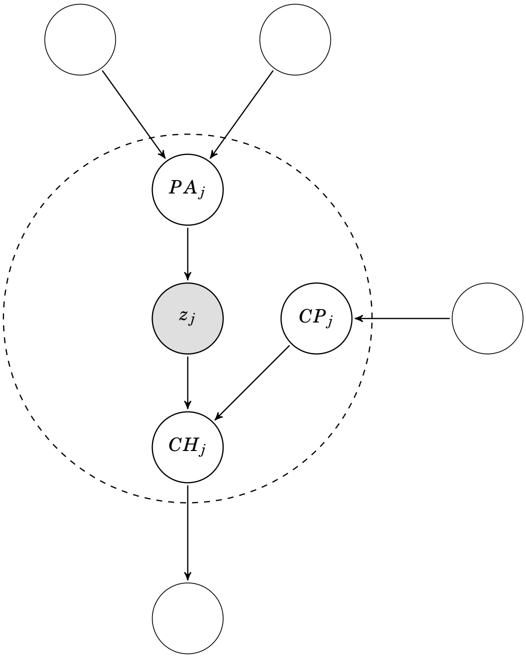
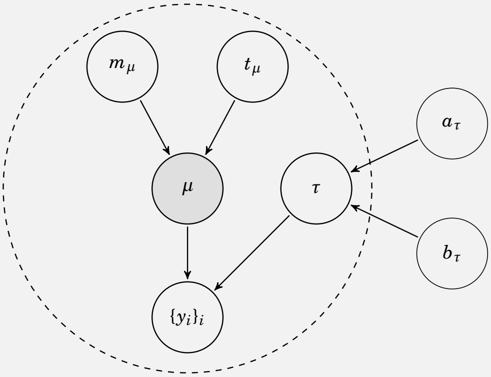


FIGURE 3.1 – *A graphical representation of a Markov Blanket. The focal parameter z_j is shaded gray. The nodes that belong to its Markov Blanket are contained in the dashed circle which represents the Markov Blanket. Conditioned on the nodes in the Markov Blanket, the nodes outside the dashed circle do not contain any additional information about z_j.*

Toy example - Markov Blanket

The Markov Blanket for μ in our example is given by:



The parents of μ are m_μ and t_τ, its children are given by $y_1,\ldots,y_N$ and τ is the co-parent of $y_1,\ldots,y_N$. Note that the prior parameters for τ are not included in the Markov Blanket for μ.

Markov blanket and MFVI

With the Markov Blanket in place we return to mean-field variational inference. The goal is to derive the optimal $q(z_j)$, the marginal variational distribution for z_j. We start by isolating the part of the ELBO that depends on $q(z_j)$:

$$\begin{aligned}
ELBO &= E_q[\log p(y,z) - \sum_i \log q(z_i)] \\
&= E_q[\log p(z_j|z_{\neg j},y) + \log p(z_{\neg j},y) - \sum_i \log q(z_i)] \\
&\overset{q(z_j)}{\propto} E_q[\log p(z_j|z_{\neg j},y) - \log q(z_j)] \\
&\overset{q(z_j)}{\propto} E_q[\log p(z_j|MB_j) - \log q(z_j)] \\
&\overset{q(z_j)}{\propto} E_q[\log p(z_j,MB_j) - \log q(z_j)].
\end{aligned} \tag{3.12}$$

where we have overloaded the proportionality operator $\propto$ to drop any additive terms that are constant with respect to $q(z_j)$. In addition, we have used the property of the Markov Blanket to rewrite the conditional density of z_j.

Next, the mean-field assumption allows us to split the variational expectation in two separate expectations, one for $q(z_j)$ and another for $q(MB_j)$:

$$\begin{aligned} ELBO &\overset{q(z_j)}{\propto} E_{q(z_j,MB_j)}[\log p(z_j,MB_j) - \log q(z_j)] \\ &= E_{q(z_j)}[E_{q(MB_j)}[\log p(z_j,MB_j)] - \log q(z_j)]. \end{aligned} \tag{3.13}$$

The $E_{q(MB_j)}[\log p(z_j,MB_j)]$ term can be interpreted as the log of an unnormalized density for z_j which we denote by $\tilde{p}_j(z_j)$. It can be normalized to a proper density by reinstalling the appropriate normalizing constant:

$$\begin{aligned} \log \tilde{p}_j(z_j) &\triangleq E_{q(MB_j)}[\log p(z_j,MB_j)], \\ p_j(z_j) &\triangleq \frac{\tilde{p}_j(z_j)}{Z_{\tilde{p}_j}} = \frac{\tilde{p}_j(z_j)}{\int_{z_j} \tilde{p}_j(z_j) dz_j} \\ &= \frac{\exp(E_{q(MB_j)}[\log p(z_j,MB_j)])}{\int_{z_j} \exp(E_{q(MB_j)}[\log p(z_j,MB_j)]) dz_j}. \end{aligned} \tag{3.14}$$

The normalizing constant $Z_{\tilde{p}_j}$ in (3.14) is not a function of $q(z_j)$ as z_j is marginalized over. Thus, adding this normalizing constant to (3.13) will not affect the optimal variational distribution for z_j:

$$\begin{aligned} ELBO &\overset{q(z_j)}{\propto} E_{q(z_j)}[E_{q(MB_j)}[\log p(z_j,MB_j)] - \log q(z_j)] \\ &\overset{q(z_j)}{\propto} -KL[q(z_j)||p_j(z_j)] + Z_{\tilde{p}_j}, \end{aligned} \tag{3.15}$$

This is an important result as it shows that the part of the ELBO that depends on $q(z_j)$ can be isolated and equals the negative KL-divergence between $q(z_j)$ and $p_j(z_j)$ as defined in (3.14). Remember that the goal in variational inference is to maximize the ELBO. It is clear that the maximization of the ELBO in (3.15) with respect to $q(z_j)$ is equivalent to minimizing this KL-divergence. By definition, this KL-divergence is minimized if $q(z_j) = p_j(z_j)$. This result can be summarized as follows:

$$\begin{aligned} q_j^\star(z_j) &\triangleq \underset{q(z_j)}{\arg\min} ELBO, \\ q_j^\star(z_j) &= p_j(z_j), \\ \log q_j^\star(z_j) &\overset{z_j}{\propto} E_{q(MB_j)}[\log p(z_j,MB_j)]. \end{aligned} \tag{3.16}$$

This result is rephrased for the reader who is more familiar with Gibbs sampling: The log density of the optimal variational distribution for z_j is (up to a constant) equal to the log full-conditional density for z_j where all moments of the variables in MB_j are replaced with their variational expectations. Note that $q(z_j)$ is our approximation of the marginal posterior for z_j. This shows a close connection between variational inference and a Gibbs sampler (Gelfand and Smith, 1990), where one would sample z_j from its full-conditional distribution to approximate the marginal posterior of z_j.

Toy example - MFVI

To derive $q^\star(\mu)$, the optimal marginal variational distribution for μ, first derive the joint density of μ and its Markov Blanket:

$$\begin{aligned} p(\mu, MB_\mu) &= p(\mu, m_\mu, t_\mu, y_1, \ldots, y_N, \tau) \\ &\overset{\mu}{\propto} p(\mu|m_\mu, t_\mu) \prod_{i=1}^{N} p(y_i|\mu, \tau) \\ &\overset{\mu}{\propto} \exp\left(-\frac{1}{2} t_\mu (\mu - m_\mu)^2 - \frac{1}{2}\tau \sum_{i=1}^{N} (y_i - \mu)^2\right) \end{aligned}$$

This can be recognized as the kernel of a Normal distribution with mean $(t_\mu + \tau N)^{-1}(t_\mu m_\mu + \tau \sum_{i=1}^{N} y_i)$ and precision $t_\mu + \tau N$. Using the result from (3.16), we know that $\log q^\star(\mu)$ is (up to a constant) equal to $E_{q(\tau)}[\log p(\mu, MB_\mu)]$, where we used that $q(MB_\mu) = q(\tau)$ because the other elements of MB_μ are fixed values. By distributing the variational expectation for τ, we find that the optimal variational distribution for μ is a Normal distribution with mean $(t_\mu + E_q[\tau]N)^{-1}(t_\mu m_\mu + E_q[\tau] \sum_{i=1}^{N} y_i)$ and precision $t_\mu + E_q[\tau]N$.

The solution in (3.16) for $q(z_j)$ contains a variational expectation over MB_j and hence, it is a function of the variational distributions $q(z_i)$ with $i \in MB_j$. This means that one cannot simultaneously solve for all $q(z_j)$ at once. Instead, a solution can be found by initializing each of the $q(z_j)$ distributions, and then iteratively updating each $q(z_j)$ conditional on the current variational distributions of the variables in MB_j. In this approach, convergence to a (local) optimum is guaranteed as the optimization problem for each $q(z_j)$ is convex. Each step is guaranteed to lead to an improvement. This optimization routine is known as coordinate ascent (Bishop, 2006).

Note that the results obtained in this section are only driven by the design of the model and the mean-field assumption. We have not made any distributional assumptions about our probabilistic model or the variational distribution. Working with distributions from the exponential family, which is a versatile class of parameterized distributions, will be the topic of our next section. The main advantage here is that these exponential family distributions simplify notation and derivations.

3.2.5 *Exponential family*

The exponential family covers a broad class of distributions that contains many well-known distributions such as the (multivariate) Normal, the Gamma, and the Dirichlet. It has been extensively studied in the literature and enjoys convenient properties, such as conjugate priors. Besides these properties it is often straightforward in VI to derive update steps for a coordinate ascent optimization that involves distributions from the exponential family. We will first briefly review the basic properties of the exponential family and subsequently discuss the properties that are convenient in the context of mean-field variational inference.

Let x be a random variable that is distributed according to a distribution from the exponential family. Each distribution in the exponential family has a log kernel of the following form:

$$\log \tilde{p}(x|\theta) = \log h(x) + f(\theta)^\top t(x). \tag{3.17}$$

Here h, f, and t are distribution-specific functions. Both f and t can be vector functions, but $h(x)$ is a scalar function by definition. $t(x)$ is the sufficient statistic of x and it is "sufficient" in the sense that it is the only interaction between the parameter θ and the random variable x in the density. $h(x)$ is the base measure and we emphasize that this function contains all the terms of x in the density that do not interact with the parameter θ.

An alternative parameterization of our distribution is to consider $\eta = f(\theta)$. In the literature, η is known as the natural parameter and considering η as the parameter of the distribution has many advantages that we will discuss shortly. The resulting (log) kernel as a function of η is:

$$\begin{aligned} \log \tilde{p}(x|\eta) &= \log h(x) + \eta^\top t(x), \\ \tilde{p}(x|\eta) &= h(x) \exp(\eta^\top t(x)). \end{aligned} \tag{3.18}$$

The constant that normalizes this kernel into a proper distribution is defined as:

$$Z(\eta) \triangleq \int_x \tilde{p}(x|\eta) dx = \int_x h(x) \exp(\eta^\top t(x)) dx, \tag{3.19}$$

and the log of this normalizer is defined as:

$$a(\eta) \triangleq \log Z(\eta). \tag{3.20}$$

The log kernel in (3.18) is normalized by subtracting the log normalizer $a(\eta)$ from it, which results in the following specification for a log density in the exponential family:

$$\log p(x|\eta) = \log h(x) + \eta^\top t(x) - a(\eta). \tag{3.21}$$

This is known as the canonical form of the exponential family and it holds that for every member of the exponential family, the log density can be written in this form.

Exponential family example

Let x be a univariate random variable that is Normally distributed with mean μ and precision τ. We will rewrite the log density of x in the canonical form.

First, we rewrite the log density:

$$\begin{aligned}\log p(x|\mu,\tau) &= -\frac{1}{2}\log 2\pi + \frac{1}{2}\log\tau - \frac{1}{2}\tau(x-\mu)^2 \\ &= \begin{bmatrix}\tau\mu \\ -0.5\tau\end{bmatrix}^\top \begin{bmatrix}x \\ x^2\end{bmatrix} - \frac{1}{2}\left(\log 2\pi - \log\tau + \tau\mu^2\right).\end{aligned}$$

In this form it is straightforward to identify the different components of the canonical form: The natural parameter $\eta = [\eta_0, \eta_1] = f(\mu,\tau) = [\tau\mu, -0.5\tau]$ and the sufficient statistic $t(x) = [x, x^2]$. The log base-measure is absent in the log density for x, which shows that for the Normal distribution it is equal to zero, or alternatively $h(x) = 1$. The natural parameter mapping can also be reversed, resulting in: $[\mu,\tau] = f^{-1}(\eta) = [(-2\eta_0)^{-1}\eta_1, (-2\eta_0)]$

The log normalizer as a function of μ and τ is given by $a(\mu,\tau) = \frac{1}{2}\left(\log 2\pi - \log\tau + \tau\mu^2\right)$. We can replace μ and τ with functions of η to obtain the log normalizer as a function of the natural parameter η: $a(\eta) = \frac{1}{2}\left(\log 2\pi - \log(-2\eta_1) + (-2\eta_1)^{-1}\eta_0^2\right)$.

Let q and p be two densities for the same class of distributions from the exponential family with natural parameters $\tilde{\eta}$ and η respectively. We examine their KL-divergence as it plays a central role in VI. In this KL-divergence, we first replace the log densities of q and p with their canonical forms:

$$\begin{aligned}KL(q(x|\tilde{\eta})||p(x|\eta)) &= E_q[\log q(x|\tilde{\eta}) - \log p(x|\eta)] \\ &= E_q\left[\log h_q(x) + t_q(x)^\top \tilde{\eta} - a_q(\tilde{\eta}) - \log h_p(x) - t_p(x)^\top \eta + a_p(\eta)\right] \\ &= E_q\left[\log h_q(x) + t_q(x)^\top \tilde{\eta} - \log h_p(x) - t_p(x)^\top \eta\right] - a_q(\tilde{\eta}) + a_p(\eta).\end{aligned} \tag{3.22}$$

Because q and p are densities for the same distribution their h, t, and a functions are equivalent, so we can drop the subscripts for these functions. By canceling and aggregating terms, we isolate the parts of the KL-divergence that are a function of $\tilde{\eta}$:

$$\begin{aligned}KL(q(x|\tilde{\eta})||p(x|\eta)) &= E_q\,[t(x)]^\top(\tilde{\eta} - \eta) - a(\tilde{\eta}) + a(\eta) \\ &\overset{\tilde{\eta}}{\propto} E_q\,[t(x)]^\top(\tilde{\eta} - \eta) - a(\tilde{\eta}).\end{aligned} \tag{3.23}$$

To evaluate this KL-divergence we need to be able to compute the expectation of the sufficient statistic $t(x)$ under the distribution q. One of the reasons for re-parameterizing the distribution in terms of the natural parameter is that it is easy to derive this expectation of the sufficient statistic $t(x)$: It is equal to the gradient of the log normalizer $a(\eta)$ with respect to η. We denote this gradient by $\nabla_\eta a(\eta)$, and it holds that:

$$\begin{aligned}\nabla_\eta a(\eta) = \nabla_\eta \log Z(\eta) = \frac{1}{Z(\eta)}\nabla_\eta Z(\eta) &= \frac{1}{Z(\eta)}\nabla_\eta \int_x \tilde{p}(x|\eta)dx \\ &= \frac{1}{Z(\eta)}\int_x h(x)\nabla_\eta \exp\left(\eta^\top t(x)\right)dx \\ &= \frac{1}{Z(\eta)}\int_x h(x)\exp\left(\eta^\top t(x)\right)t(x)dx \\ &= \int_x h(x)\exp\left(\eta^\top t(x) - a(\eta)\right)t(x)dx \\ &= \int_x p(x|\eta)t(x)dx \\ &\triangleq E[t(x)].\end{aligned} \tag{3.24}$$

This indeed shows that the mean of the sufficient statistic $t(x)$ is equal to the gradient of the log-normalizer $a(\eta)$ with respect to η. Plugging this general result back in (3.23), we obtain:

$$KL(q(x|\tilde{\eta})||p(x|\eta)) \overset{\tilde{\eta}}{\propto} \left[\nabla_{\tilde{\eta}} a(\tilde{\eta})\right](\tilde{\eta} - \eta) - a(\tilde{\eta}). \tag{3.25}$$

Exponential family - Expectation of the sufficient statistic

For a Normal distribution with natural parameter $\eta = [\tau\mu, -0.5\tau]$, we can derive the expectation of the sufficient statistic $t(x) = [x, x^2]$ by taking the gradient of the log normalizer $a(\eta) = \log Z(\eta) = \log \int_x \tilde{p}(x|\eta)dx$ with respect to η:

$$\begin{aligned}
\nabla_{\eta_0} a(\eta) &= (-2\eta_1)^{-1}\eta_0 = \mu = E[x], \\
\nabla_{\eta_1} a(\eta) &= (-2\eta_1)^{-1} + (-2\eta_1)^{-2}\eta_0^2 = \sigma^2 + \mu^2 = E[x^2].
\end{aligned}$$

Earlier in the chapter we have relied on the result that the KL-divergence is minimized if both distributions are equal, i.e. if we make q equal to p by setting $\tilde{\eta} = \eta$. Alternatively, we can also show directly that this property holds by setting the gradient of the KL-divergence with respect to $\tilde{\eta}$ to zero. This gradient is derived as follows:

$$\begin{aligned}
\nabla_{\tilde{\eta}} KL(q(x|\tilde{\eta})||p(x|\eta)) &= \nabla_{\tilde{\eta}} \left(\left[\nabla_{\tilde{\eta}} a(\tilde{\eta})\right](\tilde{\eta} - \eta) - a(\tilde{\eta}) \right) \\
&= \left[\nabla_{\tilde{\eta}} \left[\nabla_{\tilde{\eta}} a(\tilde{\eta})\right]\right](\tilde{\eta} - \eta) + \left[\nabla_{\tilde{\eta}} a(\tilde{\eta})\right] \nabla_{\tilde{\eta}}(\tilde{\eta} - \eta) - \nabla_{\tilde{\eta}} a(\tilde{\eta}) \\
&= \left[\nabla_{\tilde{\eta}\tilde{\eta}^\top} a(\tilde{\eta})\right](\tilde{\eta} - \eta) + \nabla_{\tilde{\eta}} a(\tilde{\eta}) - \nabla_{\tilde{\eta}} a(\tilde{\eta}) \\
&= \left[\nabla_{\tilde{\eta}\tilde{\eta}^\top} a(\tilde{\eta})\right](\tilde{\eta} - \eta),
\end{aligned} \tag{3.26}$$

where $\nabla_{\tilde{\eta}\tilde{\eta}^\top} a(\tilde{\eta})$ is the Hessian of the log normalizer $a(\tilde{\eta})$ with respect to the natural parameter $\tilde{\eta}$. Similar to the gradient of the log normalizer, this Hessian has a special property as well:

$$\begin{aligned}
\nabla_{\eta\eta^\top} a(\eta) = \nabla_\eta \nabla_{\eta^\top} a(\eta) = \nabla_\eta E[t(x)^\top] = \nabla_\eta \int_x p(x|\eta) t(x)^\top dx \\
&= \int_x h(x) \nabla_\eta \exp\left(\eta^\top t(x) - a(\eta)\right) t(x)^\top dx \\
&= \int_x h(x) \exp\left(\eta^\top t(x) - a(\eta)\right) \nabla_\eta \left(\eta^\top t(x) - a(\eta)\right) t(x)^\top dx \\
&= \int_x p(x|\eta) \left(t(x) - \nabla_\eta a(\eta)\right) t(x)^\top dx \\
&= \int_x p(x|\eta) (t(x) - E[t(x)]) t(x)^\top dx \\
&= \int_x p(x|\eta) \big(t(x) - E[t(x)]\big)\big(t(x) - E[t(x)]\big)^\top dx \\
&= E\left[\big(t(x) - E[t(x)]\big)\big(t(x) - E[t(x)]\big)^\top\right] \\
&\triangleq Cov[t(x)].
\end{aligned} \tag{3.27}$$

That is, the covariance matrix of the sufficient statistic $t(x)$ can be obtained by taking the Hessian of the log-normalizer $a(\eta)$ with respect to η. Plugging this result in the gradient of the KL-divergence in (3.26), we obtain:

$$\nabla_{\tilde{\eta}} KL(q(x|\tilde{\eta})||p(x|\eta)) = Cov_q[t(x)](\tilde{\eta} - \eta). \tag{3.28}$$

This shows that the gradient of $KL(q(x|\tilde{\eta})||p(x|\eta))$ has a special form when q and p are both densities for the same class of distributions in the exponential family. It is equal to the difference between the parameters of the two densities $\tilde{\eta} - \eta$, premultiplied by the covariance matrix of the sufficient statistic $t(x)$ under distribution q. It is straightforward that this gradient can be set to zero by setting $\tilde{\eta} = \eta$, which corresponds with the general property that the KL-divergence is minimized when q and p are equal.

3.2.6 *Conditionally conjugate models and MFVI*

With the above basics for the exponential family in place, it can be demonstrated how these distributions work in conjunction with VI. We consider mean-field variational inference for a specific subclass of models that involve exponential family distributions and that are *conditionally conjugate* (Blei et al., 2017). The general conjugacy property states that the posterior that results from a conjugate prior-likelihood pair is from the same family of distributions as the prior distribution. In addition to this the conditionally conjugate property ensures that the full-conditional distribution for each model parameter z_j is a member of the exponential family. Examples of such conditionally conjugate models are latent Dirichlet allocation (Blei et al., 2003), mixture models, hidden Markov models, and many more. The hierarchical Normal model that we focus on in this chapter is conditionally conjugate as well.

We combine the conditionally conjugate property with that of the Markov Blanket, displayed in (3.11), to obtain the following general expression for the full-conditional of z_j:

$$\begin{aligned} \log p(z_j|z_{\neg j}) &= \log p(z_j|MB_j) \\ &\overset{z_j}{\propto} \log p(z_j, MB_j) \\ &= \log h(z_j) + t(z_j)^\top \eta(MB_j), \end{aligned} \tag{3.29}$$

where the natural parameter $\eta(MB_j)$ is a function of the variables in the Markov Blanket z_j.

By plugging this result in the third line of (3.16) we obtain that the log kernel for the variational of z_j that minimizes the KL-divergence to $p(z_j|MB_j)$ can be written in the following form:

$$\begin{aligned}\log q_j^\star(z_j) &\overset{z_j}{\propto} E_{q(MB_j)}[\log p(z_j, MB_j)] \\ &= \log h(z_j) + t(z_j)^\top E_{q(MB_j)}[\eta(MB_j)],\end{aligned} \tag{3.30}$$

where $q_j^\star(z_j)$ is the optimal marginal variational distribution for z_j.

Summarizing this result: If mean-field variational inference is applied to conditionally conjugate models, then the optimal marginal variational distribution $q_j^\star(z_j)$ is from the same exponential family distribution as the full-conditional for z_j. The natural parameters of both distributions are closely related, as the natural parameter for $q_j^\star(z_j)$ is the variational expectation of the natural parameter for the full-conditional-distribution. That is, the natural parameter for $q_j^\star(z_j)$ is given by:

$$\tilde{\eta}_j^\star = E_{q(MB_j)}[\eta(MB_j)]. \tag{3.31}$$

With this result, we can derive the optimal marginal variational distribution for a focal variable z_j in three steps:

1. Define the Markov Blanket for z_j and write down the log joint density of z_j and the variables in its Markov Blanket.

2. From this joint, drop the terms that are constant with respect to z_j and rewrite it in the canonical form that matches with the prior for z_j.

3. Take the variational expectation of the natural parameter that results from this canonical form. This is the natural parameter for the optimal marginal variational distribution of z_j.

(3.31) shows that the variables in MB_j are potentially coupled in the full-conditional for z_j, i.e. if the variables in MB_j have interactions in $\eta(MB_j)$. If this is the case, a change in the model specification (for example, by adding or deleting a variable) will affect the solutions for the optimal marginal distributions of all variables in MB_j. This implies that one needs to adjust the update steps for all these variables in the optimization as well. This is clearly suboptimal, especially if one wants to quickly explore different model settings. In the next section we introduce a new result for a hierarchical Normal model which allows us to easily deal with this coupling. We achieve this by funneling all the dependencies between the elements in MB_j through a common error term.

Toy example - Deriving the optimal variational for μ

The optimal marginal variational distribution for μ in our toy example is derived using three steps.

1. The kernel of the log joint density of μ and its Markov Blanket MB_μ can be written as:

$$\log p(\mu, MB_\mu) \overset{\mu}{\propto} \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}^\top \begin{bmatrix} t_\mu m_\mu + \tau \sum_{i=1}^N y_i \\ -0.5(t_\mu + \tau N) \end{bmatrix}.$$

2. This is the kernel of a Normal distribution, where the natural parameter is a function of the elements in the Markov Blanket of μ.

3. By using the result in Equations (3.30) and (3.31), the optimal variational distribution for μ is a Normal distribution with natural parameter:

$$\begin{aligned} \tilde{\eta}_j^\star &= E_{q(MB_j)}[\eta(MB_j)] \\ &= \begin{bmatrix} t_\mu m_\mu + E_q[\tau] \sum_{i=1}^N y_i \\ -0.5(t_\mu + E_q[\tau] N) \end{bmatrix}. \end{aligned}$$

3.3 MFVI in hierarchical Normal models

In this section a new result for the estimation of hierarchical Normal models with mean-field variational inference is presented. The concepts discussed in Section 3.2 will be used extensively to arrive at these results. Our result shows that the dependence of parameters in a hierarchical Normal model can be funneled through a common error term. Before we arrive at this general result, we first illustrate the problem setting with examples of a linear model and a panel model.

3.3.1 *Linear model*

Consider the following linear regression model where a univariate dependent variable y_i is explained by the multivariate regressor $\mathbf{x}_i$:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + u_i, \tag{3.32}$$

where u_i is a Normally distributed error term with zero-mean and precision τ. We place a multivariate Normal prior distribution on $\boldsymbol{\beta}$, with mean-vector $\mathbf{m}_{\boldsymbol{\beta}}$ and precision-matrix $\mathbf{L}_{\boldsymbol{\beta}}$. By doing so, we have created a (simple) hierarchical Normal model, which is conditionally conjugate.

The goal is to infer $\boldsymbol{\beta}$ using mean-field variational inference. From our discussion in Section 3.2.6 we know that we can derive $q^\star(\boldsymbol{\beta})$, the optimal marginal variational distribution of $\boldsymbol{\beta}$, for a conditionally conjugate model in three steps: i) Determine $MB_{\boldsymbol{\beta}}$, the Markov Blanket for $\boldsymbol{\beta}$. ii) Derive $\boldsymbol{\eta}(MB_{\boldsymbol{\beta}})$, the natural parameter for the full-conditional distribution of $\boldsymbol{\beta}$. iii) Take the variational expectation over $\boldsymbol{\eta}(MB_{\boldsymbol{\beta}})$ to obtain the natural parameter for $q^\star(\boldsymbol{\beta})$.

For the first step, we consider the part of the log density of y_i that depends on $\boldsymbol{\beta}$. This can be written in the canonical form for a multivariate Normal distribution over $\boldsymbol{\beta}$:

$$\begin{aligned}\log p(y_i|\boldsymbol{\beta},\tau) &\overset{\boldsymbol{\beta}}{\propto} -\frac{1}{2}\tau\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 \\ &\overset{\boldsymbol{\beta}}{\propto} -\frac{1}{2}\tau\left(\mathbf{x}_i^\top \boldsymbol{\beta}\boldsymbol{\beta}^\top \mathbf{x}_i - 2y_i\mathbf{x}_i^\top \boldsymbol{\beta}\right) \\ &= \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}\boldsymbol{\beta}^\top \end{bmatrix} \cdot \begin{bmatrix} \tau \mathbf{x}_i y_i \\ -0.5\tau\mathbf{x}_i\mathbf{x}_i^\top \end{bmatrix},\end{aligned} \tag{3.33}$$

where $\cdot$ denotes the Frobenius inner-product, i.e. it calculates the dot product between two matrices as though they are vectors.

This result can be extended to the log joint density of $\boldsymbol{\beta}$ and its Markov Blanket:

$$\begin{aligned}\log p(\boldsymbol{\beta}, MB_{\boldsymbol{\beta}}) &\overset{\boldsymbol{\beta}}{\propto} \log p(\boldsymbol{\beta}|\mathbf{m}_{\boldsymbol{\beta}}, \mathbf{L}_{\boldsymbol{\beta}}) + \sum_{i=1}^{N} \log p(y_i|\boldsymbol{\beta},\tau) \\ &\overset{\boldsymbol{\beta}}{\propto} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}\boldsymbol{\beta}^\top \end{bmatrix} \cdot \begin{bmatrix} \mathbf{L}_{\boldsymbol{\beta}}\mathbf{m}_{\boldsymbol{\beta}} + \tau\sum_{i=1}^{N} \mathbf{x}_i y_i \\ -0.5(\mathbf{L}_{\boldsymbol{\beta}} + \tau\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top) \end{bmatrix}.\end{aligned} \tag{3.34}$$

The natural parameter for $q^\star(\boldsymbol{\beta})$, which we denote by $\tilde{\boldsymbol{\eta}}^\star_{\boldsymbol{\beta}}$, is obtained by taking the variational expectation of the natural parameter in (3.34). For the moment, the only stochastic parameter in $MB_{\boldsymbol{\beta}}$ is τ, resulting in:

$$\begin{aligned}\tilde{\boldsymbol{\eta}}^\star_{\boldsymbol{\beta}} &= E_{q(MB_{\boldsymbol{\beta}})}[\boldsymbol{\eta}(MB_{\boldsymbol{\beta}})] \\ &= \begin{bmatrix} \mathbf{L}_{\boldsymbol{\beta}}\mathbf{m}_{\boldsymbol{\beta}} + E_q[\tau]\sum_{i=1}^{N} \mathbf{x}_i y_i \\ -0.5(\mathbf{L}_{\boldsymbol{\beta}} + E_q[\tau]\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top) \end{bmatrix}.\end{aligned} \tag{3.35}$$

This natural parameter can be mapped to the more familiar mean-vector and covariance-matrix of a multivariate Normal:

$$\boldsymbol{\Sigma} = (-2\boldsymbol{\eta}_1)^{-1} \text{ and } \boldsymbol{\mu} = (-2\boldsymbol{\eta}_1)^{-1}\boldsymbol{\eta}_0,$$

and by applying this mapping to $\tilde{\boldsymbol{\eta}}^{\star}_{\boldsymbol{\beta}}$ we obtain:

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}^{\star}_{\boldsymbol{\beta}} &= (\mathbf{L}_{\boldsymbol{\beta}} + E_q[\tau]\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}, \\ \tilde{\boldsymbol{\mu}}^{\star}_{\boldsymbol{\beta}} &= (\mathbf{L}_{\boldsymbol{\beta}} + E_q[\tau]\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\top})^{-1}(\mathbf{L}_{\boldsymbol{\beta}}\mathbf{m}_{\boldsymbol{\beta}} + E_q[\tau]\sum_{i=1}^{N}\mathbf{x}_i y_i). \end{aligned} \tag{3.36}$$

3.3.2 *Panel linear regression model*

In case the linear model is extended to a panel model, the optimal marginal variational distribution can be obtained in a similar way. Consider that we now have $t = 1,\ldots,T_i$ observations for each individual i. The linear model is extended from (3.32) into a generic panel model:

$$y_{it} = \alpha_i + \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + u_{it}, \tag{3.37}$$

where the new systematic component α_i is endowed with a Normal prior, with mean m_α and precision t_α. This specification is a hierarchical Normal model as well.

We will show that the optimal variational distributions for α_i and $\boldsymbol{\beta}$ can be derived from the results for the linear model without any additional manual derivations. We first focus on $\boldsymbol{\beta}$ and notice that the following relation holds:

$$y_{it} - \alpha_i = \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + u_{it} \tag{3.38}$$

That is, the panel model in (3.37) can be rewritten as a linear model in which we explain $y_{it} - \alpha_i$ by $\mathbf{x}_{it}$. Here, $y_{it} - \alpha_i$ can be interpreted as the part of y_{it} that is not explained by α_i and hence, that needs to be explained by $\mathbf{x}_{it}$. More concretely, $y_{it} - \alpha_i$ is substituted for y_i in (3.34) to directly arrive at:

$$\begin{aligned} \log p(\boldsymbol{\beta}, MB_{\boldsymbol{\beta}}) &\overset{\boldsymbol{\beta}}{\propto} \log p(\boldsymbol{\beta}|\mathbf{m}_{\boldsymbol{\beta}}, \mathbf{L}_{\boldsymbol{\beta}}) + \sum_{i=1}^{I}\sum_{t=1}^{T_i}\log p(y_{it}|\alpha_i, \boldsymbol{\beta}, \tau) \\ &\overset{\boldsymbol{\beta}}{\propto} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}\boldsymbol{\beta}^{\top} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{L}_{\boldsymbol{\beta}}\mathbf{m}_{\boldsymbol{\beta}} + \tau\sum_{i=1}^{I}\sum_{t=1}^{T_i}\mathbf{x}_{it}(y_{it} - \alpha_i) \\ -0.5(\mathbf{L}_{\boldsymbol{\beta}} + \tau\sum_{i=1}^{I}\sum_{t=1}^{T_i}\mathbf{x}_{it}\mathbf{x}_{it}^{\top}) \end{bmatrix}. \end{aligned} \tag{3.39}$$

As for the linear model, the optimal natural parameter for $q^{\star}(\boldsymbol{\beta})$ is obtained by taking the variational expectation of the natural parameter in (3.39). The stochastic parameters in $MB_{\boldsymbol{\beta}}$ are given by α_i for $i = 1,\dots,I$ and τ. Note that the log kernel of y_{it} contains an interaction of α_i and τ, but because of the mean-field assumption their variational expectation factorizes $E_q[\alpha_i\tau] = E_q[\alpha_i]E_q[\tau]$:

$$\begin{aligned}\tilde{\boldsymbol{\eta}}^{\star}_{\boldsymbol{\beta}} &= E_{q(MB_{\boldsymbol{\beta}})}[\boldsymbol{\eta}(MB_{\boldsymbol{\beta}})]\\ &= \begin{bmatrix}\mathbf{L}_{\boldsymbol{\beta}}\mathbf{m}_{\boldsymbol{\beta}} + E_q[\tau]\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{x}_i(y_i - E_q[\alpha_i])\\ -0.5(\mathbf{L}_{\boldsymbol{\beta}} + E_q[\tau]\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{x}_i\mathbf{x}_i^{\top})\end{bmatrix}.\end{aligned} \tag{3.40}$$

For α_i a similar result can be obtained by replacing the prior parameters and noticing that the “coefficient” for α_i is 1:

$$\begin{aligned}\log p(\alpha_i, MB_{\alpha_i}) &\overset{\alpha_i}{\propto} \log p(\alpha_i|m_{\alpha_i}, t_{\alpha_i}) + \sum_{i=1}^{I}\sum_{t=1}^{T_i}\log p(y_{it}|\alpha_i, \boldsymbol{\beta}, \tau)\\ &\overset{\alpha_i}{\propto} \begin{bmatrix}\alpha_i\\ \alpha_i^2\end{bmatrix}^{\top}\begin{bmatrix}t_{\alpha_i}m_{\alpha_i} + \tau\sum_{i=1}^{I}\sum_{t=1}^{T_i}(y_{it} - \mathbf{x}_{it}^{\top}\boldsymbol{\beta})\\ -0.5(t_{\alpha_i} + \tau\sum_{i=1}^{I}\sum_{t=1}^{T_i}1)\end{bmatrix},\end{aligned} \tag{3.41}$$

with as optimal natural parameter for $q^{\star}(\alpha_i)$:

$$\begin{aligned}\tilde{\boldsymbol{\eta}}^{\star}_{\alpha_i} &= E_{q(MB_{\alpha_i})}[\boldsymbol{\eta}(MB_{\alpha_i})]\\ &= \begin{bmatrix}t_{\alpha_i}m_{\alpha_i} + E_q[\tau]\sum_{i=1}^{I}\sum_{t=1}^{T_i}(y_{it} - \mathbf{x}_{it}^{\top}E_q[\boldsymbol{\beta}])\\ -0.5(t_{\alpha_i} + E_q[\tau]\sum_{i=1}^{I}\sum_{t=1}^{T_i}1)\end{bmatrix}.\end{aligned} \tag{3.42}$$

These results will be generalized to arbitrary hierarchical Normal models.

3.3.3 *General hierarchical Normal model*

The results for the panel model can be extended to a more general setting. Consider a linear model for y_i where the systematic component, i.e. the mean specification, consists of $k = 1,\dots,K$ variables $\mathbf{c}_{ik}$. Each $\mathbf{c}_{ik}$ has a distinct effect on y_i, given by $\boldsymbol{\theta}_k$. This model of y_i can be written as:

$$y_i = \sum_{k=1}^{K}\mathbf{c}_{ik}\boldsymbol{\theta}_k + u_i, \tag{3.43}$$

with u_i a Normally distributed error term with zero-mean and precision τ. Each parameter $\boldsymbol{\theta}_k$ is endowed with a multivariate Normal prior distribution, with mean-vector $\mathbf{m}_{\boldsymbol{\theta}_k}$ and precision-matrix $\mathbf{L}_{\boldsymbol{\theta}_k}$. We may consider $\boldsymbol{\theta}_k$ to be a univariate parameter without loss of generality.

As in the previous examples, if we focus on a single parameter $\boldsymbol{\theta}_k$, we can consider the model where we explain the remainder $y_i - \sum_{j\neq k}\mathbf{c}_{ij}\boldsymbol{\theta}_j$ by $\mathbf{c}_{ik}$. Using this property, the joint density of $\boldsymbol{\theta}_k$ and its Markov Blanket $MB_{\boldsymbol{\theta}_k}$ can be written as the log kernel of a multivariate Normal distribution over $\boldsymbol{\theta}_k$:

$$\begin{aligned}\log p(\boldsymbol{\theta}_k, MB_{\boldsymbol{\theta}_k}) &\overset{\boldsymbol{\theta}_k}{\propto} \log p(\boldsymbol{\theta}_k | \mathbf{m}_{\boldsymbol{\theta}_k}, \mathbf{L}_{\boldsymbol{\theta}_k}) + \sum_{i=1}^{I} \log p(y_i | \{\mathbf{c}_{ik}, \boldsymbol{\theta}_k\}_{k=1}^{K}, \tau) \\ &\overset{\boldsymbol{\theta}_k}{\propto} \begin{bmatrix} \boldsymbol{\theta}_k \\ \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top \end{bmatrix} \cdot \begin{bmatrix} \mathbf{L}_{\boldsymbol{\theta}_k} \mathbf{m}_{\boldsymbol{\theta}_k} + \tau \sum_{i=1}^{I} \mathbf{c}_{ik} (y_i - \sum_{j \neq k} \mathbf{c}_{ij} \boldsymbol{\theta}_j) \\ -0.5(\mathbf{L}_{\boldsymbol{\theta}_k} + \tau \sum_{i=1}^{I} \mathbf{c}_{ik} \mathbf{c}_{ik}^\top) \end{bmatrix}.\end{aligned} \tag{3.44}$$

The result in (3.44) can be used to automatically write down the natural parameter of the multivariate Normal distribution that corresponds to the full-conditional distribution of $\boldsymbol{\theta}_k$. By itself this is a useful result, but it also exposes a general property of hierarchical Normal models: Each $\boldsymbol{\theta}_k$ parameter is coupled with all other parameters because of the $y_i - \sum_{j \neq k} \mathbf{c}_{ij} \boldsymbol{\theta}_j$ term in the natural parameter. Concretely, if the composition of the systematic component for y_i changes (i.e. a variable is added or removed), the natural parameters of all $\boldsymbol{\theta}_k$ will change as well.

This can be improved by considering that $y_i - \sum_{j \neq k} \mathbf{c}_{ij}$ is equivalent to the common error u_i plus the interaction of $\boldsymbol{\theta}_k$ with its coefficient $\mathbf{c}_{ik}$. This term is rewritten as a new auxiliary variable $\epsilon_i(\neg \mathbf{c}_{ik} \boldsymbol{\theta}_k)$:

$$\begin{aligned} y_i - \sum_{j \neq k} \mathbf{c}_{ij} &= y_i - \sum_{k=1}^{K} \mathbf{c}_{ij} \boldsymbol{\theta}_k + \mathbf{c}_{ik} \boldsymbol{\theta}_k \\ &= u_i + \mathbf{c}_{ik} \boldsymbol{\theta}_k \\ &\triangleq \epsilon_i(\neg \mathbf{c}_{ik} \boldsymbol{\theta}_k). \end{aligned} \tag{3.45}$$

The first line in (3.45) makes clear that $\mathbf{c}_{ik} \boldsymbol{\theta}_k$ cancels and hence, does not appear in $\epsilon_i(\neg \mathbf{c}_{ik} \boldsymbol{\theta}_k)$. Using this result, (3.44) can now be written in the following form:

$$\log p(\boldsymbol{\theta}_k, MB_{\boldsymbol{\theta}_k}) \overset{\boldsymbol{\theta}_k}{\propto} \begin{bmatrix} \boldsymbol{\theta}_k \\ \boldsymbol{\theta}_k \boldsymbol{\theta}_k^\top \end{bmatrix} \cdot \begin{bmatrix} \mathbf{L}_{\boldsymbol{\theta}_k} \mathbf{m}_{\boldsymbol{\theta}_k} + \tau \sum_{i=1}^{I} \mathbf{c}_{ik} \epsilon_i (\neg \mathbf{c}_{ik}^\top \boldsymbol{\theta}_k) \\ -0.5(\mathbf{L}_{\boldsymbol{\theta}_k} + \tau \sum_{i=1}^{I} \mathbf{c}_{ik} \mathbf{c}_{ik}^\top) \end{bmatrix}. \tag{3.46}$$

By taking the variational expectation of this natural parameter[2], we obtain the natural parameter for the optimal marginal variational distribution for $\boldsymbol{\theta}_k$ as

[2] We have assumed here that $\mathbf{c}_{ik}$ is not stochastic. If $\mathbf{c}_{ik}$ is stochastic, the same general result holds if $\mathbf{c}_{ik}$ only interacts with $\boldsymbol{\theta}_k$ in the model.

follows:

$$
\begin{aligned}
q^{\star}(\boldsymbol{\theta}_k) &= \begin{bmatrix} \mathbf{L}_{\boldsymbol{\theta}_k}\mathbf{m}_{\boldsymbol{\theta}_k} + E_q[\tau]\sum_{i=1}^{I}\mathbf{c}_{ik}E_q[\epsilon_i(\neg\mathbf{c}_{ik}^{\top}\boldsymbol{\theta}_k)] \\ -0.5(\mathbf{L}_{\boldsymbol{\theta}_k} + E_q[\tau]\sum_{i=1}^{I}\mathbf{c}_{ik}\mathbf{c}_{ik}^{\top}) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{L}_{\boldsymbol{\theta}_k}\mathbf{m}_{\boldsymbol{\theta}_k} + E_q[\tau]\sum_{i=1}^{I}\mathbf{c}_{ik}(E_q[u_i] + \mathbf{c}_{ik}^{\top}E_q[\boldsymbol{\theta}_k]) \\ -0.5(\mathbf{L}_{\boldsymbol{\theta}_k} + E_q[\tau]\sum_{i=1}^{I}\mathbf{c}_{ik}\mathbf{c}_{ik}^{\top}) \end{bmatrix}.
\end{aligned}
\tag{3.47}
$$

This result has two implications: First, the dependencies between $\boldsymbol{\theta}_k$ parameters are funneled in a function of the common error term u_i. In case we change the model (for example, by adding or deleting variables) we only have to update the specification of u_i to adapt this change for all variables in the model. Secondly, working with $\epsilon_i(\neg\mathbf{c}_{ik}^{\top}\boldsymbol{\theta}_k)$ has a computational advantage. It follows that the naive computation of $y - \sum_{j \neq k}\mathbf{c}_{ij}^{\top}\boldsymbol{\theta}_j$ requires $K-1$ operations. As there are K parameters in the mean specification for y_i, this results in $(K-1)\times K$ operations in total to calculate this term for every $\boldsymbol{\theta}_k$. In contrast, by using $\epsilon_i(\neg\mathbf{c}_{ik}^{\top}\boldsymbol{\theta}_k)$ instead we need much less operations: The computation of $u_i = y_i - \sum_{k=1}^{K}\mathbf{c}_{ik}^{\top}\boldsymbol{\theta}_k$ requires K operations and each $\epsilon_i(\neg\mathbf{c}_{ik}\boldsymbol{\theta}_k)$ for $k = 1,\ldots,K$, can be constructed from u_i using one additional operation. Hence, this approach requires only $2K$ operations in total and scales linearly with the number of parameters K. This is in stark contrast to the quadratic complexity of the naive approach.

3.4 Estimating the common precision matrix of a set of multivariate Normals with independent multivariate Normal variationals

In order to gain richer insights from the data, one might be interested in adding covariances to a model. However, adding such covariances is typically not without a (computational) cost. For example, the number of parameters in the covariance matrix of a multivariate Normal (MVN) distribution scales quadratically with the dimension K of the distribution. This problem is further amplified when the model is estimated using vanilla mean-field variational inference (MFVI). Consider that we have I random variables in the model that follow the same MVN distribution in the data generating process. In vanilla MFVI, each of these variables will be endowed with an MVN as marginal variational distribution, where the mean and covariance is specific to the focal variable. Clearly, the number of variational parameters grows very rapidly if either K or I increases.

Instead, in this section we deviate from vanilla MFVI and show how the common precision matrix for a set of MVN variables can be efficiently inferred. We obtain this gain in efficiency by endowing each of the I MVN variables with a variational distribution that consists of a set of independent (univariate)

Normals. This set of Normals can be considered to be a special case of an MVN distribution with a diagonal covariance matrix. We call such a distribution an independent multivariate Normal (iMVN). In this section, we will first formulate the problem setting. Subsequently, we provide the general closed-form solution of an iMVN that approximates an MVN and discuss the properties of this optimal approximation.

3.4.1 *Problem setting*

Consider the following simple hierarchical Normal model:

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}^\top \boldsymbol{\beta}_i + u_{it}, \\ \boldsymbol{\beta}_i &= \mathbf{m} + \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{v}_i, \end{aligned} \tag{3.48}$$

where u_{it} is a Normally distributed error term with some variance, and $\boldsymbol{v}_i$ is an MVN distributed error term with the identity as covariance matrix. Hence, $\boldsymbol{\beta}_i$ follows a multivariate Normal distribution with mean $\mathbf{m}$ and precision matrix $\boldsymbol{\Lambda}$. For simplicity we consider $\mathbf{m}$ to be known and fixed, this is without loss of generality. We are interested in estimating the unknown model parameters $\boldsymbol{\beta}_i$ for $i = 1, \ldots, I$ and $\boldsymbol{\Lambda}$ using MFVI.

Note that the conjugate prior for a precision matrix of an MVN distribution is the Wishart distribution. In other words, if we endow $\boldsymbol{\Lambda}$ with a Wishart prior with prior scale matrix $\mathbf{V}_\pi$ and prior degrees of freedom n_π, the full-conditional density of $\boldsymbol{\Lambda}$ conditioned on its Markov Blanket $MB_{\boldsymbol{\Lambda}}$ is proportional to the kernel of another Wishart. In the canonical form of the exponential family, the log of this kernel is given by:

$$\begin{aligned} \log p(\boldsymbol{\Lambda}|\{\boldsymbol{\beta}_i\}_{i=1}^I, \mathbf{m}, n_\pi, \mathbf{V}_\pi) &\overset{\boldsymbol{\Lambda}}{\propto} \log p(\boldsymbol{\Lambda}|n_\pi, \mathbf{V}_\pi) + \sum_{i=1}^{I} \log p(\boldsymbol{\beta}_i|\mathbf{m}, \boldsymbol{\Lambda}) \\ &\overset{\boldsymbol{\Lambda}}{\propto} \begin{bmatrix} \boldsymbol{\Lambda} \\ \log\det(\boldsymbol{\Lambda}) \end{bmatrix} \cdot \begin{bmatrix} -0.5[\mathbf{V}_\pi + \sum_{i=1}^I (\boldsymbol{\beta}_i - \mathbf{m})(\boldsymbol{\beta}_i - \mathbf{m})^\top] \\ 0.5(n_\pi + I - K - 1) \end{bmatrix}, \end{aligned} \tag{3.49}$$

where $\log\det(\boldsymbol{\Lambda})$ is the log determinant of the $\boldsymbol{\Lambda}$ matrix.

The $\sum_{i=1}^I (\boldsymbol{\beta}_i - \mathbf{m})(\boldsymbol{\beta}_i - \mathbf{m})^\top$ term in (3.49) can be recognized as I times the covariance matrix of the $\boldsymbol{\beta}_i$ vectors with mean $\mathbf{m}$. Notice that it captures the covariance *across* the $\boldsymbol{\beta}_i$ vectors, instead of the elements within a single $\boldsymbol{\beta}_i$ vector.

We can straightforwardly determine the optimal variational natural parameter for $q(\boldsymbol{\Lambda})$, which we denote by $\tilde{\boldsymbol{\eta}}_{\boldsymbol{\Lambda}}^\star$. It is obtained by taking the variational expectation of the natural parameter in (3.49), which results in:

$$\begin{aligned}\tilde{\boldsymbol{\eta}}^{\star}_{\boldsymbol{\Lambda}} &= E_{q(MB_{\boldsymbol{\Lambda}})}[\boldsymbol{\eta}(MB_{\boldsymbol{\Lambda}})] \\ &= \begin{bmatrix} -0.5[\mathbf{V}_{\pi} + \sum_{i=1}^{I} E_{q(\boldsymbol{\beta}_i)}[(\boldsymbol{\beta}_i - \mathbf{m})(\boldsymbol{\beta}_i - \mathbf{m})^{\top}]] \\ 0.5(n_{\pi} + I - K - 1) \end{bmatrix}.\end{aligned} \tag{3.50}$$

Naturally, the variational expectation of $(\boldsymbol{\beta}_i - \mathbf{m})(\boldsymbol{\beta}_i - \mathbf{m})^{\top}$ in (3.50) depends on $q(\boldsymbol{\beta}_i)$, the marginal variational distribution for $\boldsymbol{\beta}_i$. That is, if we change $q(\boldsymbol{\beta}_i)$ from an MVN to an iMVN distribution, this expectation is likely to change as well. Note, however, that this does not imply that the covariances *across* the $\boldsymbol{\beta}_i$ vectors become zero. This can be easily verified by considering the degenerate case of an iMVN where all (co)variances are zero, i.e. fixed values for $\boldsymbol{\beta}_i$ which would result in $\sum_{i=1}^{I} E_{q(\boldsymbol{\beta}_i)}[(\boldsymbol{\beta}_i - \mathbf{m})(\boldsymbol{\beta}_i - \mathbf{m})^{\top}] = \sum_{i=1}^{I}(\boldsymbol{\beta}_i - \mathbf{m})(\boldsymbol{\beta}_i - \mathbf{m})^{\top}$. In other words, as long as there is variation across the (variational posterior) means of the $\boldsymbol{\beta}_i$ vectors, the resulting precision matrix and hence, covariance matrix, will have non-zero off-diagonal elements. As degenerate distributions are unlikely to be accurate approximations for the posterior of $\boldsymbol{\beta}_i$, and full covariance MVNs are computationally very demanding, we outline in the next section the closed-form solution when approximating an MVN with an iMVN. This result can then be used within MFVI to obtain the optimal variational parameters for $\boldsymbol{\Lambda}$, as displayed in (3.50).

3.4.2 *Optimal iMVN to approximate an MVN*

In this section we will derive the optimal variational parameters for an iMVN that is used to approximate a generic MVN distribution in MFVI, i.e. the parameters that minimize the KL-divergence between the two distributions. In addition, we discuss the properties of these optimal parameters and contrast them against an MVN specification.

Let $\mathbf{x}$ be a K-dimensional *hierarchical* MVN random variable, so that its posterior is another MVN. The goal in variational inference is to minimize the KL-divergence between the variational distribution $q(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$, the posterior MVN distribution with mean-vector $\boldsymbol{\mu}$ and precision-matrix $\boldsymbol{\Lambda}$. Note that the result presented in this section holds for general MVN distributions, so without loss of generality $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ can refer to a prior MVN distribution as well.

This setting is not as straightforward as typically is the case when working with distributions from the exponential family in MFVI, as q and p are not from the same distribution: $q(\mathbf{x})$ is an iMVN distribution with diagonal covariance matrix, while $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is an MVN distribution with full covariance matrix. To minimize this KL-divergence we have to find the value of the variational parameters for $q(\mathbf{x})$ that sets the gradient of this KL-divergence to zero.

Before we can solve the gradient of $KL[q(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})] = E_{q(\mathbf{x})}[\log q(\mathbf{x}) - \log p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})]$ we first have to derive an expression for it. We start with the first term in this KL-divergence, the log density of $q(\mathbf{x})$ which is an iMVN variational distribution, and only consider the terms that depend on its variational natural parameter $\tilde{\boldsymbol{\eta}}$. The result, written in the canonical form of the exponential family, is given by:

$$\log q(\mathbf{x}) \overset{\tilde{\boldsymbol{\eta}}}{\propto} \begin{bmatrix} \mathbf{x} \\ \mathbf{x}^2 \end{bmatrix}^\top \begin{bmatrix} \tilde{\boldsymbol{\eta}}_0 \\ \tilde{\boldsymbol{\eta}}_1 \end{bmatrix} - a(\tilde{\boldsymbol{\eta}}) = t_q(\mathbf{x})^\top \begin{bmatrix} \tilde{\boldsymbol{\eta}}_0 \\ \tilde{\boldsymbol{\eta}}_1 \end{bmatrix} - a(\tilde{\boldsymbol{\eta}}), \tag{3.51}$$

where $\mathbf{x}^2$ is the vector that contains the element-wise squares of $\mathbf{x}$ and $t_q(\mathbf{x}) = [\mathbf{x}, \mathbf{x}^2]$ is the sufficient statistic of the iMVN distribution.

Contrast this against the canonical form of the log kernel for an MVN distribution, which is the second term in the KL-divergence:

$$\begin{aligned} \log p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}) &\overset{\mathbf{x}}{\propto} -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) \\ &\overset{\mathbf{x}}{\propto} \begin{bmatrix} \mathbf{x} \\ \mathbf{x}\mathbf{x}^\top \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} \\ -0.5\boldsymbol{\Lambda} \end{bmatrix} = t_p(\mathbf{x}) \cdot \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} \\ -0.5\boldsymbol{\Lambda} \end{bmatrix}, \end{aligned} \tag{3.52}$$

where the sufficient statistic of the MVN is given by $t_p(\mathbf{x}) = [\mathbf{x}, \mathbf{x}\mathbf{x}^\top]$.

The canonical form of the MVN does not match the one from the iMVN distribution, because the two sufficient statistics are different. We can rewrite the result from (3.52) by noting that it is possible to split the matrix dot product between $\mathbf{x}\mathbf{x}^\top$ and $\boldsymbol{\Lambda}$ in two parts. One part only involves second-order moments of $\mathbf{x}$ (the diagonal), while the other part only contains cross-terms of first-order moments (the off-diagonal elements). That is:

$$\mathbf{x}\mathbf{x}^\top \cdot [-0.5\boldsymbol{\Lambda}] = {\mathbf{x}^2}^\top [-0.5d(\boldsymbol{\Lambda})] - 0.5\mathbf{x}^\top [\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]\mathbf{x}, \tag{3.53}$$

where we have defined $d()$ as a function that works on both matrix and vector arguments: For a matrix argument, it returns a vector with the diagonal elements of the matrix. For a vector argument, it returns a diagonal matrix where the diagonal is populated with the elements of the vector.

By plugging this result back in (3.52), we obtain:

$$\begin{aligned}\log p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}) &\overset{\mathbf{x}}{\propto} \begin{bmatrix}\mathbf{x}\\ \mathbf{x}\mathbf{x}^\top\end{bmatrix} \cdot \begin{bmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu}\\ -0.5\boldsymbol{\Lambda}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{x}\\ \mathbf{x}^2\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - 0.5[\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]\mathbf{x}\\ -0.5d(\boldsymbol{\Lambda})\end{bmatrix},\end{aligned} \tag{3.54}$$

which now closely resembles the log of the kernel of an independent multivariate Normal distribution. Note that this is not an exponential family density in proper canonical form, as $\mathbf{x}$ appears in both terms of (3.54). This means that we cannot rely on result (3.31) to find the optimal variational parameters for $q(\mathbf{x})$.

Instead, we take the variational expectation[3] of the two log densities derived in (3.51) and (3.54) to obtain an expression for $KL[q(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})] = E_{q(\mathbf{x})}[\log q(\mathbf{x}) - \log p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})]$:

$$\begin{aligned}&KL[q(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})]\\ &\overset{q(\mathbf{x})}{\propto} -E_{q(\mathbf{x})}[\log p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}) - \log q(\mathbf{x})]\\ &\overset{q(\mathbf{x})}{\propto} -E_{q(\mathbf{x})}\begin{bmatrix}\mathbf{x}\\ \mathbf{x}^2\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - 0.5[\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]E_{q(\mathbf{x})}[\mathbf{x}] - \tilde{\boldsymbol{\eta}}_0\\ -0.5d(\boldsymbol{\Lambda}) - \tilde{\boldsymbol{\eta}}_1\end{bmatrix} - a(\tilde{\boldsymbol{\eta}}),\end{aligned} \tag{3.55}$$

where $a(\tilde{\boldsymbol{\eta}})$ is the log-normalizer of the iMVN variational distribution for $\mathbf{x}$. Note that we can distribute the variational expectation because $\mathbf{x}^\top[\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]\mathbf{x}$ only contains first order cross-moments of $\mathbf{x}$ and explicitly no higher order moments of $\mathbf{x}$.

By taking the gradient of (3.55) with respect to the variational natural parameter $\tilde{\boldsymbol{\eta}}$, we obtain:

$$\begin{aligned}&\nabla_{\tilde{\boldsymbol{\eta}}} KL[q(\mathbf{x})||p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})]\\ &\quad= -Cov_{q(\mathbf{x})}\begin{bmatrix}\mathbf{x}\\ \mathbf{x}^2\end{bmatrix} \begin{bmatrix}\boldsymbol{\Lambda}\boldsymbol{\mu} - [\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]E_{q(\mathbf{x})}[\mathbf{x}] - \tilde{\boldsymbol{\eta}}_0\\ -0.5d(\boldsymbol{\Lambda}) - \tilde{\boldsymbol{\eta}}_1\end{bmatrix},\end{aligned} \tag{3.56}$$

where we have used the exponential family identities that the gradient of the log-normalizer is equal to the mean of the sufficient statistic, and that the Hessian of the log-normalizer is equal to the covariance matrix of the sufficient statistic.

We can find the optimal value for $\tilde{\boldsymbol{\eta}}$ that minimizes the KL-divergence between $q(\mathbf{x})$ and $p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda})$, by setting this gradient to zero. This is equivalent to solving

[3] Formally, we have to take the variational expectations over $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. To avoid notational clutter we take $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ constant in these derivations. However, typically they are also inferred using mean-field variational inference and in these formulas they should be replaced with their variational expectations.

the following system of $2K$ equations:

$$\begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} - [\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]E_{q(\mathbf{x})}[\mathbf{x}] - \tilde{\boldsymbol{\eta}}_0 \\ -0.5d(\boldsymbol{\Lambda}) - \tilde{\boldsymbol{\eta}}_1 \end{bmatrix} = 0. \tag{3.57}$$

The optimal value for $\tilde{\boldsymbol{\eta}}_1$, denoted by $\tilde{\boldsymbol{\eta}}_1^\star$, can be straightforwardly derived by first solving the second set of K equations:

$$\tilde{\boldsymbol{\eta}}_1^\star = -0.5d(\boldsymbol{\Lambda}). \tag{3.58}$$

Solving the other set of K equations to find the optimal value for $\tilde{\boldsymbol{\eta}}_0$ requires some work. Notice that $E_{q(\mathbf{x})}[\mathbf{x}] = (-2\tilde{\boldsymbol{\eta}}_1)^{-1} \cdot \tilde{\boldsymbol{\eta}}_0$ and we can rewrite:

$$\begin{aligned} &\boldsymbol{\Lambda}\boldsymbol{\mu} - [\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]E_{q(\mathbf{x})}[\mathbf{x}] - \tilde{\boldsymbol{\eta}}_0 \\ &\quad = \boldsymbol{\Lambda}\boldsymbol{\mu} - [\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))](-2\tilde{\boldsymbol{\eta}}_1)^{-1} \cdot \tilde{\boldsymbol{\eta}}_0 - \tilde{\boldsymbol{\eta}}_0 \\ &\quad = \boldsymbol{\Lambda}\boldsymbol{\mu} - [\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]d(-2\tilde{\boldsymbol{\eta}}_1)^{-1}\tilde{\boldsymbol{\eta}}_0 - \tilde{\boldsymbol{\eta}}_0 \\ &\quad = \boldsymbol{\Lambda}\boldsymbol{\mu} - \left([\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]d(-2\tilde{\boldsymbol{\eta}}_1)^{-1} + I_K\right)\tilde{\boldsymbol{\eta}}_0, \end{aligned} \tag{3.59}$$

where I_K is the K-dimensional identity matrix. By replacing $\tilde{\boldsymbol{\eta}}_1$ with its solution $\tilde{\boldsymbol{\eta}}_1^\star = -0.5d(\boldsymbol{\Lambda})$, we can further expand:

$$\begin{aligned} &\boldsymbol{\Lambda}\boldsymbol{\mu} - [\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]E_{q(\mathbf{x})}[\mathbf{x}] - \tilde{\boldsymbol{\eta}}_0 \\ &\quad = \boldsymbol{\Lambda}\boldsymbol{\mu} - \left([\boldsymbol{\Lambda} - d(d(\boldsymbol{\Lambda}))]d(d(\boldsymbol{\Lambda}))^{-1} + I_K\right)\tilde{\boldsymbol{\eta}}_0 \\ &\quad = \boldsymbol{\Lambda}\boldsymbol{\mu} - \left(\boldsymbol{\Lambda}d(d(\boldsymbol{\Lambda}))^{-1} - d(d(\boldsymbol{\Lambda}))d(d(\boldsymbol{\Lambda}))^{-1} + I_K\right)\tilde{\boldsymbol{\eta}}_0 \\ &\quad = \boldsymbol{\Lambda}\boldsymbol{\mu} - \left(\boldsymbol{\Lambda}d(d(\boldsymbol{\Lambda}))^{-1} - I_K + I_K\right)\tilde{\boldsymbol{\eta}}_0 \\ &\quad = \boldsymbol{\Lambda}\boldsymbol{\mu} - \left(\boldsymbol{\Lambda}d(d(\boldsymbol{\Lambda}))^{-1}\right)\tilde{\boldsymbol{\eta}}_0 \end{aligned} \tag{3.60}$$

Using this expression, we can straightforwardly solve the first set of K equations from (3.57) which shows that the optimal value for $\tilde{\boldsymbol{\eta}}_0$ is given by:

$$\begin{aligned} \tilde{\boldsymbol{\eta}}_0^\star &= \left(\boldsymbol{\Lambda}d(d(\boldsymbol{\Lambda}))^{-1}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{\mu} \\ &= d(d(\boldsymbol{\Lambda}))\boldsymbol{\mu} \\ &= d(\boldsymbol{\Lambda}) \cdot \boldsymbol{\mu}. \end{aligned} \tag{3.61}$$

We can map $\tilde{\boldsymbol{\eta}}$ from the natural parameter space to the mean and covariance:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}^\star &= (-2\tilde{\boldsymbol{\eta}}_1^\star)^{-1} \cdot \tilde{\boldsymbol{\eta}}_0^\star = \boldsymbol{\mu}, \\ \boldsymbol{\sigma}^{2\star} &= (-2\tilde{\boldsymbol{\eta}}_1^\star)^{-1} = d(\boldsymbol{\Lambda})^{-1}. \end{aligned} \tag{3.62}$$

Contrast this result against the optimal values in case we would have specified an MVN for $q(\mathbf{x})$, which are naturally given by the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Lambda}^{-1}$ of distribution p. This shows that in MFVI, the iMVN variational q that best approximates an MVN distribution p has an intuitive solution. The random variable $\mathbf{x}$ has the same mean under q and p, which makes sense. The variances of the $\mathbf{x}$ elements under q are equal to $d(\boldsymbol{\Lambda})^{-1}$, which is the inverse of the diagonal elements of the precision matrix $\boldsymbol{\Lambda}$. Note that these values correspond to the conditional variances for $\mathbf{x}$ under p. The variances under p are given by $d(\boldsymbol{\Lambda}^{-1})$, i.e. the regular variances. It holds that $d(\boldsymbol{\Lambda})^{-1} \leq d(\boldsymbol{\Lambda}^{-1})$, where the equality only holds if $\boldsymbol{\Lambda}$ is a diagonal matrix. In words, the variances of an MVN will be underestimated if it is approximated with an iMVN, but the mean is accurately retrieved.

3.5 Conclusion

In this chapter we have reviewed (mean-field) variational inference and demonstrated how it can be used to infer the model structure of conditionally conjugate models that involve distributions from the exponential family. Besides the review of different concepts related to variational inference, we have provided extra insight that should make it easier to understand and dive into variational inference. This holds especially for the reader that is more familiar with traditional Bayesian statistical inference. In addition to this review, we presented two new results for a hierarchical Normal model in the context of mean-field variational inference.

First, we have shown that the dependencies of the systematic components, i.e. the parameters in the mean-specification of a Normally distributed variable can be funneled through the common error term. This results has two important implications: i) It allows for “automatic variational inference”, circumventing the need for manual derivations if the model structure is changed. ii) Directly working with the common error term has a lower computational complexity. This is especially relevant in the context of variational inference, as it is often employed in large and complex models.

Secondly, we presented a new result for mean-field variational inference to estimate a common precision matrix of a set of multivariate Normally distributed variables in a scalable way. The result uses a set of independent univariate Normal distributions to approximate the multivariate Normally distributed variables. We derived the closed-form analytical solution for this approximation that makes intuitive sense. This result enables the inference of covariance structures in high-dimensional models, without incurring the large increase in complexity that is typically accompanied with the introduction of a covariance matrix in mean-field variational inference.

These new results jointly make it much easier for a researcher to iteratively infer models that involve hierarchical Normals with variational inference. The need to manually derive model-specific results is taken away, reducing the time needed to employ a variational inference algorithm. In addition, these results could potentially be incorporated in an automatic VI algorithm, such as ADVI (Kucukelbir et al., 2017), to potentially improve its performance.

4

Model-Based Marketing Insights from High-Dimensional Purchase Data

4.1 Introduction

The size of the average product assortment has increased exponentially over the past two decades. This increase can primarily be explained by the rapid development of online retailers: The retailing landscape developed from a space constrained brick-and-mortar store to virtually unbounded online assortments. Naturally this change has had a profound effect on the way customers navigate through these assortments (Gourville and Soman, 2005), but the customer is not the only one affected. A retail manager is required to maintain such a large assortment and as the size increases, it arguably becomes more difficult to manage such an assortment, gain actionable insights, and develop effective marketing strategies. An additional challenge for retailers in this setting is to incorporate all available data to optimize their marketing efforts, which in particular holds for the online environment where typically data is abundant.

In this chapter a method is developed that can be used to gain insight in customer purchase behavior in large product assortments, accounting for dynamics, seasonality, and other information that may be available at the customer and purchase occasion level. The setting that is of interest concerns purchases from large and varied assortments that are challenging by its size. Typically such assortments are encountered in large brick-and-mortar stores and at online retailers. The goal is to develop a model-based tool that is able to provide insight at an aggregate level, such as seasonality patterns, but at the same time is able to provide insight when zoomed in on the individual customer or purchase occasion, e.g. identifying relevant products. These insights can be used by a manager to get a grip on the patterns underlying the purchase behavior at her firm. All of this is achieved in a scalable and computationally efficient manner, paving the way for advanced analytics on large-scale purchase history data.

To meet this goal, the size of the assortment is the primary challenge that has to be overcome as it is typically too large to be analyzed by traditional methods and unwieldy to directly analyze at the individual product level. A tempting idea might be to consider only a subset of the assortment but this approach is not without drawbacks. For example, in most retail settings product categories are defined a priori and our attention could be restricted to a single product category. However, with such an analysis one is precluded from gaining insight in purchase patterns that span across multiple categories. It is only natural to expect these patterns in real purchase behavior, as it seems unlikely that customers decide to restrict themselves to a single category when shopping. Moreover, the identification of such relations across categories might be one of the insights the retail manager is after. Another option is to restrict attention to just the most frequently purchased products in the data. In this case it can be argued that for retailers that hold a large assortment, the variety and depth of the product assortment on sale is one of their key strengths. By only considering the high-volume products we effectively ignore this. Next, even after excluding low-volume products the assortment size may still be very large. Put differently, this study is looking for a way to gain insight from the observed purchase behavior, without excluding any product a priori.

We choose to make the size of the assortment manageable by using a model that describes purchases using a lower dimensional set of drivers for behavior. In this way, we do not impose hard exclusions on any of the products or rely on an existing product category definition. To enable this, we build upon the work of Jacobs et al. (2016), discussed in Chapter 2, where purchases out of a large assortment of products are modeled and predicted by adapting the latent Dirichlet allocation (LDA) (Blei et al., 2003) model. The general idea here is that purchase behavior can be described by a set of *motivations*. These motivations are used to model the purchase behavior of all customers, and they are identified by the co-occurrence of products in the purchase histories of customers.

Each motivation captures a pattern of purchase behavior that can be used to describe the preference for a subset of related products in the assortment. It is represented in the model as a vector of probabilities over all products in the assortment. For a given motivation, the products with a high probability in this vector are relevant, and convey the "story" of the motivation. Motivations are able to span across categories: Examples are environmentally friendly products, products from a specific brand, or products for the price-sensitive customer. However, we want to emphasize that such attributes are latent, and that the motivations are solely inferred from the customers' purchase histories. In addition, the model allows customers to have heterogeneous preferences across purchase occasions.

Reducing the dimensionality by introducing these motivations also facilitates interpretation, as in general the number of motivations is much smaller than the size of the assortment. Each motivation can be (visually) inspected to understand the underlying purchase pattern it captures from the data. For example, the most probable products under a motivation can be listed, or alternatively a word cloud of product names can be created for each motivation (Liu and Toubia, 2017). This is similar to analyzing the components from a principal components analysis and in fact, LDA-type models have been labeled in the literature as *multinomial PCA* (Buntine, 2002).

This chapter alleviates a number of limitations of the modeling approach introduced in Jacobs et al. (2016), in order to provide more meaningful managerial insights that go beyond purchase prediction: First, it seems natural to have correlations among these motivations, a feature that the standard LDA model does not allow. We therefore rely on one of its extensions, the correlated topic model (CTM) introduced by Blei and Lafferty (2007), which allows for correlations across motivations. This is achieved by no longer modeling the motivation proportions with a Dirichlet, but instead to use a multivariate Normal that is transformed using a softmax[1] function. This takes the multivariate Normal from the real-vector space, and maps it to a distribution over probability vectors. In this way the probabilities in this distribution can have a non-trivial correlation structure, as the multivariate Normal has a covariance matrix. This is in contrast to the Dirichlet, where all correlations are small negative numbers by definition.

Secondly, each purchase occasion is considered separately, while accounting for the fact that these purchases are made by the same customer. This allows for preferences that can differ significantly across purchase occasions and thereby lets customer preferences vary over time. In addition, because we do not aggregate over purchase occasions, the date and time tied to a specific purchase occasion are preserved. This enables us to include seasonality effects in the model.

Lastly, by considering the purchase occasions separately it allows for state dependence between purchase occasions, i.e. a customer's current purchases will play a role in what this customer will be interested in on a next shopping trip. We include such dynamics in our framework by modeling the relevance of the motivations at a purchase occasion as a function of the motivation relevance in the previous purchase occasion. This allows us to quantify the persistence of each motivation over time, i.e. whether it will persist over multiple baskets.

However, these extensions are not without (computational) costs. One of the selling points of the model introduced in Jacobs et al. (2016), is its scalability, and

[1] The softmax is the transformation associated with the multinomial logit model.

by including correlations and dynamics the model will objectively become more complex. We acknowledge that scalability is an important factor in our problem domain as well, and therefore we offset the increase in complexity by a more efficient estimation routine. This is in line with the solutions recently proposed in Wedel and Kannan (2016), to enable marketing analytics in a big data setting.

We use a new estimation routine which allows us to infer our model parameters in a fast and scalable way. The algorithm relies on *variational inference*, an estimation technique that originates from the machine learning literature (Jordan et al., 1999, Wainwright and Jordan, 2008, Hoffman et al., 2013). The basic idea behind variational inference is to consider Bayesian inference as an optimization problem, and then to solve this problem using proven optimization techniques. In this way, results can generally be obtained much faster compared to traditional sampling techniques such as MCMC (Ansari et al., 2016, Kucukelbir et al., 2017). In addition, the estimation of a hierarchical model such as the model we will introduce can be trivially parallelized across a cluster of computers.

The estimation routine relies on the results derived in Chapter 3 and is structured in such a way that it is easily extended without the requirement of additional derivations or major adjustments to the computer code. Both are time-intensive and require specialized knowledge. Examples of these generic extensions are the addition/alteration of dynamics, seasonality, or additional customer-information. This property is especially attractive if the model is applied in a dynamic setting such as a dashboard. Over time, the focus of the manager may shift towards other explanatory variables, or perhaps new variables become available over time. In particular, this makes our algorithm easily adaptable to new data sets with varying types of explanatory variables, aiding in the adoption of our model in practice.

Although the research application and model specification are very different from this chapter, the goal outlined by Dew and Ansari (2017) is applicable to our research as well. In that paper the authors develop a semi-parametric model-based approach for the analysis of purchase timing. Arguably this is a fairly complex method. However, the authors focus on converting the model's output to actionable results that are presented in a dashboard-like environment. In our research we face a similar challenge: First, a model that can be used to model purchase decisions in the setting of a large product assortment is developed. Subsequently, we want to leverage the model structure to gain insights in the patterns underlying the purchase behavior.

Trusov et al. (2016) have recently proposed a modeling approach that is similar to ours. They adapt the correlated topic model (CTM) to track and profile browsing behavior of users across website categories. Although CTM is also used as the foundation of our model, there are a few key differences that set the two modeling approaches apart: First of all is the scale of the application. In their application, Trusov et al. (2016) consider visits to 29 distinct website categories. This number is in stark contrast to the number of products for sale at a large retailer, and in our application we consider thousands of products. Secondly, for inference the authors rely on traditional MCMC sampling methods. In case we are dealing with high-dimensional data in a complex hierarchical model, it is questionable whether MCMC sampling methods are able to adequately infer the posterior distribution in a reasonable amount of time. As discussed above, our solution to this problem is to turn to variational inference, which is a different estimation technique to infer the model parameters.

All these extensions enable one to obtain additional insight in the purchase patterns in our data, both on an aggregate level as well as for an individual customer. On the customer-base level we can now link the relevance of each motivation to predictor variables such as seasonality and, in addition, to customer- and basket-specific predictor variables. This can facilitate the design of marketing strategies as for example, we can learn which products to advertise to which customers at which point in time.

On the individual level, a customer's journey at the retailer can be tracked. We can infer the general preferences of a customer, but as we consider the purchase occasions separately we are able to examine which motivations were relevant for the customer at a specific moment in time. This does not only allow us to look at purchases in hindsight; we can also predict what a customer might be interested in at the next shopping trip. For example, this opens up opportunities for targeting, where customers are selected based on which motivations are likely to be relevant for them.

The layout of the remainder of this chapter is as follows: In Section 4.2 the methodology for our modeling framework is introduced. We start with the basic model that can be used to describe purchase behavior using motivations. Subsequently, this model is extended with dynamics and explanatory variables. Next, Section 4.3 describes the data that will be used. The data deals with purchases at a retailer that wishes to remain anonymous. The results of our model applied to this data set are displayed and discussed in detail in Section 4.4. Finally, we conclude in Section 4.5 with an overview, discuss the managerial implications, and provide topics for further research.

4.2 Methods

In this section our method and its details are presented. We start with the basics of our modeling approach which can be used to discover latent motivations from purchase history data. Our method is gradually extended to allow for correlations between motivations, the inclusion of explanatory variables at the basket- and customer-level, and for persistence of motivations. As our extensions enter the model in non-linear ways, we discuss how their effects can be interpreted. We conclude by providing details of our estimation routine which is based on variational inference, an estimation technique from the machine learning literature.

To discover the global patterns that explain the purchases in our data we follow the work of Jacobs et al. (2016) in which a scalable approach is introduced to model high-dimensional purchase data. This method is an adaptation of latent Dirichlet allocation (LDA): The most commonly used topic model introduced in the seminal paper of Blei et al. (2003). The basic idea is that just as a document can be viewed as a collection of words, a purchase history can be considered to be a set of purchases. As topics are incongruous in the context of purchase data, they are relabeled to *motivations*. In the cited paper, the adapted model is applied to purchases of non-food fast moving consumer goods and some example motivations that are inferred are a preference for products that are Eco-friendly, for diet products, or products for the sensitive skin.

In our application, a motivation can be thought of as a specific *project*. Examples that come to mind are projects such as gardening and renovating the kitchen, or alternatively, projects that are more driven by a season such as Christmas. What such motivations have in common is that each of them typically relates to a (small) subset of the products in the assortment, making it easy to infer the narrative of the motivation, using the descriptions of products that receive high probability under a motivation. In a more technical sense the motivations can be considered to be latent components that are used to describe our high-dimensional purchase data in a lower dimensional space. This reduction in dimensionality facilitates interpretation and allows us to more easily uncover patterns in the data to gain insight in the purchase behavior.

In our model, each motivation is represented by a probability vector over all J products in the assortment. The products that are likely under a given motivation convey the story of this motivation, while the products that receive a low probability are less important for this specific motivation. We index the motivations by $m = 1,\dots,M$. The generative process for motivation m's probability vector is given by:

$$\boldsymbol{\phi}_m \sim \text{DIRICHLET}_J(\boldsymbol{\alpha} = \mathbf{1}\zeta). \tag{4.1}$$

Here $\mathbf{1}$ is a J-dimensional vector of ones, and the positive scalar parameter ζ determines the sparseness of the Dirichlet: Sparse probability vectors in which just a few products receive the majority of the probability mass are favored with smaller values of ζ, while conversely, more uniformly distributed vectors are more likely for larger values of ζ. In our model, ζ is treated as a formal parameter that has to be estimated.

The motivations have to be linked in the model to the actual purchases that we observe for each of the $i = 1,\ldots,I$ customers. This is achieved by introducing latent motivation assignments on the level of individual purchases, that connect each observed product purchase to one of the motivations. More specifically, for customer i each item in basket $b = 1,\ldots,B_i$ is assigned to one of the M motivations. By doing so a flexible model that enables a shopping trip to serve multiple purposes is created, as the items in a basket can be assigned to different motivations. The items in basket b are indexed by $n = 1,\ldots,N_{ib}$ and the latent motivation assignment to item n is denoted by z_{ibn} and modeled as:

$$z_{ibn} \sim \textsc{Categorical}_M(\boldsymbol{\theta} = \boldsymbol{\theta}_{ib}), \quad \text{for } n = 1,\ldots,N_{ib}. \tag{4.2}$$

Note that $\boldsymbol{\theta}_{ib}$, the probability vector that describes the likelihood of each motivation is both customer- and basket-specific. This extends Jacobs et al. (2016), where the baskets of a customer are aggregated to a single purchase history. As a customer may have different goals in mind for each shopping trip, it intuitively makes sense to let the relevance of each motivation vary over baskets.

With the motivations $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_M$ and the latent assignment z_{ibn} in place, we now need to specify a generative model for the purchases. Let us assume for the moment that the motivation for purchase n in basket b is equal to m, i.e. $z_{ibn} = m$. Under this condition, it is likely that the product corresponding to this purchase, denoted by y_{ibn}, is a product that has high probability under motivation m, i.e. the products that have high values in $\boldsymbol{\phi}_m$. Formalizing this we have:

$$y_{ibn}|z_{ibn} = m \sim \textsc{Categorical}_J(\boldsymbol{\theta} = \boldsymbol{\phi}_m). \tag{4.3}$$

Summarizing the generative process for the purchases:

1. For each motivation $m = 1, \dots, M$:

   Draw a probability vector $\boldsymbol{\phi}_m \sim \text{DIRICHLET}_J(\boldsymbol{\alpha} = \mathbf{1}\zeta)$

2. For each customer $i = 1, \dots, I$:

   For each of i's baskets $b = 1, \dots, B_i$:

   For each purchased item $n = 1, \dots, N_{ib}$ in basket b:

   i. Draw a motivation assignment:
   $z_{ibn} \sim \text{CATEGORICAL}_M(\boldsymbol{\theta} = \boldsymbol{\theta}_{ib})$

   ii. Draw a product using the motivation assignment:
   $y_{ibn} | z_{ibn} \sim \text{CATEGORICAL}_J(\boldsymbol{\theta} = \boldsymbol{\phi}_{z_{ibn}})$

In the process outlined above we have not yet specified a model for $\boldsymbol{\theta}_{ib}$, the basket-specific vector of probabilities over the M motivations. In the vanilla implementation of LDA, $\boldsymbol{\theta}_{ib}$ follows a Dirichlet distribution, i.e. $\boldsymbol{\theta}_{ib} \sim \text{DIRICHLET}_M(\boldsymbol{\alpha})$, where in the literature $\boldsymbol{\alpha}$ is often considered to be a fixed value (Wallach et al., 2009). A convenient property of the Dirichlet is that it is the conjugate prior for the Categorical distribution. In our model the z_{ibn} are Categorically distributed, so that we could enjoy the conjugacy between $\mathbf{z}_{ib} = \{z_{ib1}, \dots, z_{ibN_{ib}}\}$ and $\boldsymbol{\theta}_{ib}$. Most notably, the joint density of $\mathbf{z}_{ib}$ and $\boldsymbol{\theta}_{ib}$ would be proportional to the kernel of a Dirichlet, which would allow us to conveniently marginalize $\boldsymbol{\theta}_{ib}$ out of the model while still retaining a closed-form joint density. However, in our application this convenient property of the Dirichlet is offset by a computational and conceptual drawback.

First, we cover the disadvantage that is of computational nature: The normalizer of the Dirichlet distribution is the Beta function which is relatively complex to evaluate. By retaining a Dirichlet prior on $\boldsymbol{\theta}_{ib}$ we have to evaluate this function many times for each basket in the data. In case we want to link available explanatory variables to the relevance of a motivation, we have to make the parameter of the Dirichlet a function of these explanatory variables. This approach was advocated in Jacobs et al. (2016), but further increases the number of times the Beta function has to be evaluated. This severely slows down inference and thereby hampers the scalability of the model.

The other disadvantage involves the limited correlation structure of the Dirichlet distribution. By definition, the correlations between all elements of a Dirichlet are negative, which is a direct result of the fact that these elements sum to unity. In addition, the correlations tend to be small for large enough M, such that the elements of $\boldsymbol{\theta}_{ib}$ are nearly independent. In practice, this is not a realistic modeling assumption as it is likely that some motivations will positively co-occur (e.g. working in the garden and building an outdoor pool), while other motivation may be uncorrelated (for example, garden work and a kitchen renovation).

To overcome these challenges we drop the Dirichlet assumption for $\boldsymbol{\theta}_{ib}$ and extend upon the correlated topic model (CTM) (Blei and Lafferty, 2007), which itself is an extension of LDA. The Dirichlet prior $\boldsymbol{\theta}_{ib}$ is replaced by a deterministic function of $\boldsymbol{\alpha}_{ib}$, where $\boldsymbol{\alpha}_{ib}$ is a parameter vector that describes the preferences over motivations in the b-th basket for customer i. As we need to ensure that $\boldsymbol{\theta}_{ib}$ remains a valid probability vector, a natural choice is to apply the softmax transformation to $\boldsymbol{\alpha}_{ib}$:

$$\boldsymbol{\theta}_{ib}(\boldsymbol{\alpha}_{ib}) = \textsc{Softmax}(\boldsymbol{\alpha}_{ib}) = \frac{\exp(\boldsymbol{\alpha}_{ib})}{\sum_k \exp(\alpha_{ibk})}. \tag{4.4}$$

The softmax transformation is reminiscent of the multinomial logit function and can be interpreted in a similar fashion: if α_{ibm} is large compared to the other values in $\boldsymbol{\alpha}_{ib}$, this translates to a large value for θ_{ibm}, increasing the probability for motivation m in basket b. The reverse holds in case α_{ibm} is relatively small compared to the other values in $\boldsymbol{\alpha}_{ib}$.

Put differently, $\boldsymbol{\alpha}_{ib}$ can be interpreted as a measure for the relevance of the motivations in basket b. Each element in $\boldsymbol{\alpha}_{ib}$ is modeled separately by a Normal distribution. For the moment the mean of this distribution, denoted by μ_{ibm}, is set equal to the baseline preference of customer i for motivation m, given by κ_{im}. However, we foreshadow here that in the remainder of this section we will extend the specification for μ_{ibm} such that it becomes heterogeneous among customers and across baskets. For now we have the following model for the relevance of motivation m in basket b:

$$\alpha_{ibm} \sim \textsc{Normal}(\mu = \mu_{ibm} = \kappa_{im}, \tau = \tau_m), \tag{4.5}$$

i.e. a Normal distribution with mean κ_{im} and precision τ_m. This is equivalent to the following linear model for α_{ibm}:

$$\alpha_{ibm} = \mu_{ibm} + \sqrt{\tau_m^{-1}}\epsilon_{ibm} = \kappa_{im} + \sqrt{\tau_m^{-1}}\epsilon_{ibm}, \tag{4.6}$$

where ϵ_{ibm} is standard Normally distributed and we have substituted κ_{im} for μ_{ibm}.

Note that in this model all correlations between elements of $\boldsymbol{\alpha}_{ib}$ are implicitly set to zero. Instead, we allow correlations between motivations to enter the model at the customer level. For customer i the relevance of each motivation is captured in the vector $\boldsymbol{\kappa}_i = [\kappa_{i1}, \ldots, \kappa_{iM}]$, for which we specify a multivariate Normal, with mean $\boldsymbol{\mu}_\kappa$ and full precision matrix $\boldsymbol{\Lambda}_\kappa$. The motivation for placing correlations on $\boldsymbol{\kappa}_i$ instead of $\boldsymbol{\alpha}_{ib}$ is that we expect to see more systematic correlations at the customer-level, compared to the basket-level.

Let us motivate this by an example: Consider that customers with an outdoor pool are more likely to work in the garden as well. Assume we have discovered a pool motivation and a garden motivation. It is likely that these motivations are correlated at the customer-level. However, the same reasoning does not have to hold at the basket level. It seems reasonable to assume that one shopping trip can serve the gardening needs of the customer, while another purchase occasion is focused on the outdoor pool. In our application this is further amplified by the fact that the number of purchased products for an average basket tends to be low, which does not provide a lot of room for correlations to be identified.

This concludes the specification of our baseline model. In the next section we will discuss how the model can be extended in several ways by expanding the systematic component for α_{ibm}.

4.2.1 *Adding explanatory variables to the model*

One of the goals in this chapter is to create a dashboard-like environment that can be used by a manager to gain insight in variations in purchase behavior for different segments of the customer base and over time. In this section we present our methodology for incorporating customer- and basket-specific effects in the model, complemented with an autoregressive specification for the motivation relevance at the basket level. These additional variables give the manager extra controls to zoom in on specific details of the purchase behavior at her firm. For example, for specific customer segments or at specific points in time.

In purchase history data each basket is associated with a date and time of day. An important implication of modeling each basket separately, instead of aggregating them to a single purchase history, is that the time stamp of the basket is retained. By including this information in the model we can relate changes in purchase behavior to seasonality patterns, and uncover how the relevance of motivations shift accordingly. For example, one expects different shopping baskets during the December month, i.e. the holiday season, then during a summer month.

Let there be K_x basket-specific predictor variables that are denoted by $\mathbf{x}_{ib}$. A prototypical example of a variable contained in $\mathbf{x}_{ib}$ is a month dummy. By adding these dummies to the specification of α_{ibm}, we allow the relevance of motivation m to differ over the months. Each of these variables has its own motivation-specific effect and for motivation m we collect these in the K_x-dimensional vector $\boldsymbol{\beta}_m$. In addition to the customer-specific intercept κ_{im}, the α_{ibm} specification is now defined as:

$$\alpha_{ibm} = \kappa_{im} + \boldsymbol{\beta}_m^\top \mathbf{x}_{ib} + \sqrt{\tau_m^{-1}}\epsilon_{ibm}.$$

Naturally, we can include any information that is available at the customer-level in a similar way. Such predictor variables are especially important as they provide the manager with a way to analyze the customer base. In turn, this information could be used to create a targeting strategy. For example, if the manager is interested in promoting a product that has a high probability under one of the motivations, one strategy might be to time this promotion such that the corresponding motivation is likely relevant.

The K_w customer-specific variables are summarized in the vector $\mathbf{w}_i$. Examples of such variables are age and gender. Similar to the basket-variables above, we collect for motivation m the K_w effects in the vector $\boldsymbol{\gamma}_m$, and further extend the α_{ibm} specification as follows:

$$\alpha_{ibm} = \kappa_{im} + \boldsymbol{\beta}_m^\top \mathbf{x}_{ib} + \boldsymbol{\gamma}_m^\top \mathbf{w}_i + \sqrt{\tau_m^{-1}}\epsilon_{ibm}.$$

Finally, we would like to infer if certain motivations are persistent across shopping trips. To determine this the α_{ibm} is extended to include a lagged term: $\alpha_{i,b-1,m}$. Effectively, this creates an autoregressive model for α_{ibm} which can be used to assess the level of persistence. We endow each motivation m with its own autocorrelation coefficient ρ_m. This allows for varying degrees of persistence over the motivations, as one motivation may be more persistent than another. Adding this to the model we obtain the complete specification for α_{ibm}:

$$\alpha_{ibm} = \kappa_{im} + \boldsymbol{\beta}_m^\top \mathbf{x}_{ib} + \boldsymbol{\gamma}_m^\top \mathbf{w}_i + \rho_m \alpha_{i,b-1,m} + \sqrt{\tau_m^{-1}}\epsilon_{ibm}. \tag{4.7}$$

Note that for the initial baskets the lagged values are missing and hence to obtain an unconditional mean of α_{ibm} that is consistent across baskets, we need to account for these missing lags. With the following specification we allow for shifts in both level and slopes to compensate for the missing lagged value in the initial periods:

$$\alpha_{i1m} = \delta_{0m} + \delta_{1m}\kappa_{im} + \delta_{2m}(\boldsymbol{\beta}_m^\top \mathbf{x}_{ib}) + \delta_{3m}(\boldsymbol{\gamma}_m^\top \mathbf{w}_i) + \sqrt{\tau_m^{-1}}\epsilon_{ibm}. \tag{4.8}$$

This section is concluded with a remark on identification in the model: Note that $\boldsymbol{\theta}_{ib}$ is invariant to shifts in the level of $\boldsymbol{\alpha}_{ib} = [\alpha_{ib1}, \ldots, \alpha_{ibM}]$. That is, if we add a scalar constant c to each α_{ibm} element, the result of the SOFTMAX transformation will again be $\boldsymbol{\theta}_{ib}$. One approach to work around this is to set one of the motivations as a baseline (Trusov et al., 2016). We choose not to follow this approach, primarily because the covariance structure over the other $M-1$ motivations is dependent on which motivation is chosen as baseline. In our case this is undesired, as one of the focal points of our analysis is interpretation. Instead, as we are working in the Bayesian setting, we rely on priors to identify the level of $\boldsymbol{\alpha}_{ib}$. Its prior uniformly influences all motivations, which identifies the level of $\boldsymbol{\alpha}_{ib}$ in the model.

4.2.2 *Interpretation of the effect sizes*

The specification for α_{ibm} defined above contains several explanatory components, i.e. $\mathbf{x}_{ib}$, $\mathbf{w}_i$, and $\alpha_{i,b-1,m}$. Each of these have their own effect on the value of α_{ibm}, given by $\boldsymbol{\beta}_m$, $\boldsymbol{\gamma}_m$, and ρ_m respectively. However, α_{ibm} is a latent construct in our model for which changes in its level are difficult to interpret directly. Instead, we examine the effect of these variables on $\boldsymbol{\theta}_{ib}$, the probabilities of the motivations at the basket level. In the model these probabilities are a non-linear function of $\boldsymbol{\alpha}_{ib}$, so they will be affected by the explanatory variables in a non-linear manner as well.

Because of the non-linearity of the Softmax transformation, cf. (4.4), the effect of a focal explanatory variable on $\boldsymbol{\theta}_{ib}$ is dependent on the values of the other variables in the specification for α_{ibm}. Remember that α_{ibm} is specified as $\mu_{ibm} + \sqrt{\tau_m^{-1}}\epsilon_{ibm}$, i.e. it is a combination of a systematic component μ_{ibm} and a standard Normally distributed disturbance term ϵ_{ibm}. Hence, to interpret the effect of the focal variable on $\boldsymbol{\theta}_{ib}$, we need a baseline value for μ_{ibm}, which can be computed using baseline values for $\mathbf{x}_{ib}$, $\mathbf{w}_i$, and $\alpha_{i,b-1,m}$ for $m = 1,\ldots,M$. For the explanatory variables in $\mathbf{x}_{ib}$ and $\mathbf{w}_i$ that are continuous, an intuitive baseline is given by their values for the average customer, while for discrete variables a natural choice is to set them to their reference level. A baseline for $\alpha_{i,b-1,m}$ can be created by averaging the posterior means over all baskets in the data. For ϵ_{ibm} we draw samples from a Normal distribution with variance equal to the posterior mean of τ_m^{-1}. These draws combined with the baseline for μ_{ibm} can be used to compute values for $\boldsymbol{\theta}_{ib}$. The reference value for $\boldsymbol{\theta}_{ib}$ is obtained by averaging over these values.

We can measure the partial effect of one of the variables in $\mathbf{x}_{ib}$ and $\mathbf{w}_i$ by increasing the level of the focal variable with one unit and measure how the motivation probabilities change as a result. This applies to both discrete and continuous variables. Our approach to measure the effect of ρ_m on $\boldsymbol{\theta}_{ib}$ is similar. We apply a shift in the value of $\alpha_{i,b-1,m}$ (which is set to the average value) and determine how the motivation probabilities are affected by this shift.

4.2.3 Model inference

In order to examine the motivations in our purchase history data and analyze how different explanatory variables affect the relevance of each motivation, we first need to estimate the model parameters and latent variables. In this section we discuss our estimation technique, variational inference, and provide the prior specification of our model.

Estimation

For the estimation of our model we rely on variational inference (VI), a technique developed in the machine learning literature (Jordan et al., 1999) to approximate posterior densities. In this section we will cover the basics of variational inference. Readers looking for a general review of the technique are referred to Bishop (2006), Blei et al. (2017), and Chapter 3. For additional technical details, including a stochastic extension to VI that uses subsampling to speed up inference by several orders of magnitude, we refer to Hoffman et al. (2013).

The general idea underlying VI is that we can reformulate the probabilistic inference problem as an optimization problem that can be optimized with well-known techniques from the optimization literature. This is in contrast to the inference methods that are based on MCMC sampling such as the Metropolis-Hastings algorithm (Hastings, 1970) and the Gibbs sampler (Gelfand and Smith, 1990), that over the past decade have become prevalent in marketing in order to estimate hierarchical Bayesian models (Rossi et al., 2012). The promise of VI is that because it involves a deterministic optimization problem, it can be performed in a fraction of the time it would cost to estimate the model using a stochastic MCMC method, as typically VI converges much faster (Ansari et al., 2016, Kucukelbir et al., 2017). In addition, VI is ready directly after convergence. This is in contrast to a sampling-based inference method such as MCMC, which requires additional draws after the Markov Chain has reached its stationary distribution.

Returning to the inference problem, we are interested in examining the posterior distribution of our model which is defined as the density of our unknown model components conditioned on the data:

$$p(\mathbf{z},\boldsymbol{\phi},\zeta,\boldsymbol{\alpha},\boldsymbol{\kappa},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\rho},\boldsymbol{\delta},\boldsymbol{\tau},\boldsymbol{\mu}_K,\boldsymbol{\Lambda}_K|\mathbf{y};\mathbf{x},\mathbf{w}) = p(\boldsymbol{\Omega}|y),$$

where we let $\boldsymbol{\Omega}$ serve as a placeholder for all unknown model components. The goal in VI is to approximate this posterior with another distribution, which is called the *variational distribution* and denoted by $q(\boldsymbol{\Omega})$. The fit of $q(\boldsymbol{\Omega})$ to $p(\boldsymbol{\Omega}|y)$

is measured in terms of the Kullback-Leibler (KL) divergence, defined as:

$$\mathrm{KL}[q(\mathbf{\Omega})||p(\mathbf{\Omega}|y)] \triangleq \mathrm{E}_q[\log q(\mathbf{\Omega}) - \log p(\mathbf{\Omega}|y)], \tag{4.9}$$

where $\triangleq$ denotes “defined as” and E_q is the variational expectation operator, i.e. the expected value under the variational distribution q.

The objective in VI is to find $q(\mathbf{\Omega})$ that is at close as possible to $p(\mathbf{\Omega}|y)$ in terms of KL-divergence. More formally, we need to solve the following minimization problem:

$$q^\star(\mathbf{\Omega}) = \underset{q(\mathbf{\Omega})}{\arg\min}\, \mathrm{KL}[q(\mathbf{\Omega})||p(\mathbf{\Omega}|y)] \tag{4.10}$$

Note that the objective in (4.10) is minimized and equal to zero if we let $q(\mathbf{\Omega}) = p(\mathbf{\Omega}|y)$ (Kullback and Leibler, 1951). That result is not very helpful, as $p(\mathbf{\Omega}|y)$ is the posterior that we are trying to infer. However this result does show that in essence, variational inference does not have to result in an approximation of the posterior (Bishop, 2006).

To facilitate estimation restrictions are placed on $q(\mathbf{\Omega})$ and one of the most commonly used restrictions is the mean-field assumption. It states that the variational distribution $q(\mathbf{\Omega})$ should factorize over all unknown model components:

$$q(\mathbf{\Omega}) = \prod_{\omega_u \in \mathbf{\Omega}} q(\omega_u),$$

where each $q(\omega_u)$ has its own set of variational parameters $\tilde{\eta}_u$. In addition, we assume that the class of each variational factor $q(\omega_u)$ matches the distribution we specified for ω_u in our data generating process. For example, $\boldsymbol{\phi}_m$ follows a Dirichlet distribution in the model so we let $q(\boldsymbol{\phi}_m)$ be a Dirichlet as well. As a result of this mean-field restriction, the variational typically underestimates the variance of the posterior density (Blei et al., 2017), while the posterior means are recovered accurately.

Factorizing the variational distribution does not imply that we simply ignore the dependencies between parameters in the model. The mean-field assumption does not affect the model's posterior $p(\mathbf{\Omega}|y)$, it only affects our approximation $q(\mathbf{\Omega})$. The goal is still to fit $q(\boldsymbol{\phi}_m)$ to $p(\boldsymbol{\phi}_m|\mathbf{y})$, the marginal posterior of $\boldsymbol{\phi}_m$ in the model:

$$\begin{aligned} q(\boldsymbol{\phi}_m) &\approx p(\boldsymbol{\phi}_m|\mathbf{y}) \\ &= \sum_{\mathbf{z}} \int_{\zeta} p(\boldsymbol{\phi}_m, \mathbf{z}, \zeta|\mathbf{y}) d\zeta \\ &\overset{\boldsymbol{\phi}_m}{\propto} \sum_{\mathbf{z}} \int_{\zeta} p(\mathbf{y}|\mathbf{z} = m, \boldsymbol{\phi}_m) p(\boldsymbol{\phi}_m|\zeta) d\zeta, \end{aligned}$$

which clearly depends on the other parameters $\mathbf{z}$ and ζ in the model.

A coordinate descent algorithm is employed to iteratively optimize for each of the variational parameters $\tilde{\eta}_u$ in turn, while holding all other variational parameters fixed. Each of these local problems is concave, guaranteeing that this algorithm converges to a (local) optimum of the variational objective function (Bishop, 2006).

The prior specification

The complete model contains several parameters that have a prior distribution, i.e. whose parameters are not estimated as part of the model. Below we list each parameter that is endowed with a prior distribution, interpret the parameter's role in the model, and provide their prior distribution and our rationale for that choice.

The ζ parameter controls the sparseness of the Dirichlet distribution for the $\boldsymbol{\phi}_m$ vectors, i.e. the probability distributions that correspond to the M motivations. Smaller values for ζ favor sparse distributions, while larger values favor distributions that are more diffuse. As ζ is a parameter for a Dirichlet, it is restricted to positive values. For this reason, we place a log-Normal prior distribution on ζ, with parameters $\mu = -\log(J)\frac{J+1}{J}$ and $\sigma^2 = 2\log(J)\frac{1}{J}$, where J is the number of products in the assortment. The mean of this log-Normal is equal to $\frac{1}{J}$ and hence, favors sparse $\boldsymbol{\phi}_m$ vectors. This prior was chosen as sparse vectors are typically easier to interpret. However, we note that this prior carries the weight equivalent to that of a single purchase in the data and hence, its effect is negligible compared to the total number of purchases.

For all parameters in the systematic component of the α_{ibm} specification we specify (multivariate) Normal priors. This choice is primarily driven by computational reasons: α_{ibm} is itself Normally distributed, and the conjugate prior for a Normal is another Normal.

The effects of the basket-specific $\mathbf{x}_{ib}$ and the customer-specific $\mathbf{w}_i$ variables on the relevance of motivation m at a purchase occasion are given by the parameter vectors $\boldsymbol{\beta}_m$ and $\boldsymbol{\gamma}_m$. A priori we cannot expect a certain effect for the explanatory variables as the motivations are latent prior to inference. Hence, each of the $\boldsymbol{\beta}_m$

and $\boldsymbol{\gamma}_m$ parameter vectors is endowed with a multivariate Normal prior that has zero mean and identity covariance. We remark that a variance of one in this setting corresponds to a broad prior distribution, as the effect of $\boldsymbol{\beta}_m$ and $\boldsymbol{\gamma}_m$ on $\boldsymbol{\theta}_{ib}$ is exponential.

In addition, we have an alternative specification for the α_{ibm} values that relate to the initial baskets of the customers. This is because in the first basket the lagged value for α_{ibm} is missing. This is corrected for by a shift in the level and slope of the α_{i1m} specification in (4.8), where the shifts are relative to the regular α_{ibm} specification in (4.7). We have parameters that shift the level $(\delta_{0m}, \delta_{1m})$ and parameters that shift the effect of explanatory variables $(\delta_{2m}, \delta_{3m})$. Again, we specify Normal distributions with unit variance for these parameters. For δ_{0m} we set $\mu = 0$ and for δ_{1m} we set $\mu = 1$, which effectively corresponds to no shift in level. Similarly, for δ_{2m} and δ_{3m} we set $\mu = 1$, which corresponds to no shift in effect for the explanatory variables.

Another parameter in the α_{ibm} specification for which we need to specify a prior distribution is τ_m, which is the precision (inverse covariance) for the α_{ibm} values. Again, we choose a conjugate prior, which in this case is the Gamma distribution. As parameters we set $\alpha = 2M$ and $\beta = 2M$. This corresponds to a mean of 1 and a mode that is approximately 1 for large values of M. The weight of this prior is negligible compared to the data, as typical applications will have tens of thousands of baskets.

The last set of parameters that require a prior distribution are $\boldsymbol{\mu}_\kappa$ and $\boldsymbol{\Lambda}_\kappa$: The mean vector and precision matrix (inverse covariance) of the multivariate Normal distribution for the $\boldsymbol{\kappa}_i$ vectors. Remember that $\boldsymbol{\kappa}_i$ describes the baseline relevance of the M motivations for customer i. The conjugate priors for the mean and precision of a multivariate Normal are given by another multivariate Normal and the Wishart distribution, respectively. For $\boldsymbol{\mu}_\kappa$ we specify a multivariate Normal with zero mean and as covariance we take the identity matrix. Again, this variance of one corresponds to a broad prior distribution as these parameters indirectly affect $\boldsymbol{\theta}_{ib}$ exponentially. In addition, note that this zero mean vector has an important role in the model, as it is the primary identification for the level of both α_{ibm} and $\boldsymbol{\kappa}_i$. For $\boldsymbol{\Lambda}_\kappa$ we specify a Wishart distribution with scale matrix $\mathbf{V} = \frac{1}{2} I_M$, where I_M is the $M \times M$ identity matrix, and degrees of freedom $n = 2M$. This sets the mean of the Wishart equal to I_M, while the mode rapidly approaches $\frac{1}{2} I_M$ for all but the smallest settings of M.

4.3 Data

Our method is applied to purchase history data that is made available to us by a retailer wishing to remain anonymous.[2] The data set contains detailed information on the shopping basket contents for a subset of the retailer's customers, that are tracked over a 24-month window (ranging from March 2012 - March 2014). We focus on those customers who made a purchase in one of the Florida-stores of this retailer. In this subsection several of the details concerning the data set and the single preprocessing step taken are discussed.

The retailer currently has a hierarchical product taxonomy in place that classifies the products in the assortment. The taxonomy consists of three levels that range from generic to specific, namely: Group, Class, Subclass. A product is described by the combination of these three levels and a unique product description. In total, the dataset contains purchases from an assortment that is comprised of 29,027 distinct products and we note that many of these products are rarely purchased. Theoretically, our model works for such products as well. However, we are interested in gaining substantive insights from the data instead of capturing purchase patterns that are driven by just a few co-occurrences. To achieve this, we choose to reduce the assortment size using the available product taxonomy.

Naturally, we want to retain the detailed information that is available to us for the frequently purchased products. That is, we only choose to aggregate products that are bought very infrequently. We define an infrequent product as one that is purchased at less than 10 purchase occasions in two years time. In this way 3,301 products in the data are identified as frequent. The other infrequent products are "rolled-up" one level in the hierarchy, i.e. the infrequent products within a certain Subclass are aggregated to a single product that is specific to this Subclass. By doing so we lose some details of the data, i.e. the unique product descriptions of the aggregated products, but we retain much more information compared to just deleting the infrequent products from the data. This latter approach is typically taken in applications of topic models to deal with infrequent terms. If after this aggregation step some infrequent products remain, we apply the same procedure to aggregate from the Subclass- to the Class-level, and from the Class- to the Group-level. In this way we are able to retain all but 19 infrequent products. These 19 products are removed from the data, together with their 34 purchases. 4,266 products remain with a combined total of 139,622 purchases. We provide some descriptives of this aggregation step in Table 4.1, where we divide the 4,266 products by the three levels of the aggregation step.

[2] We thank the Wharton Customer Analytics Initiative (WCAI) for setting up the research opportunity that has connected us to this retailer.

TABLE 4.1 – *Descriptives for the three levels of the aggregation step.*

Aggregation level	Unique products	Aggregated products	Purchases
Unaggregated	3,301	3,301	80,902
Aggregated to Subclass	24,563	875	56,991
Aggregated to Class	911	78	1,382
Aggregated to Group	233	12	347
Total	29,008	4,266	139,622

The data tracks the purchase behavior of 2,259 customers that made 47,568 shopping trips over the 24-month window. This totals to about 21 baskets on average per customer. In Figure 4.1 the frequencies of the number of baskets per customer are displayed. Some of the customers visit the retailer once during this period, but the vast majority of the customers returns multiple times. There are 139,622 purchases made in total, indicating relatively small baskets with on average just under 3 products per basket.

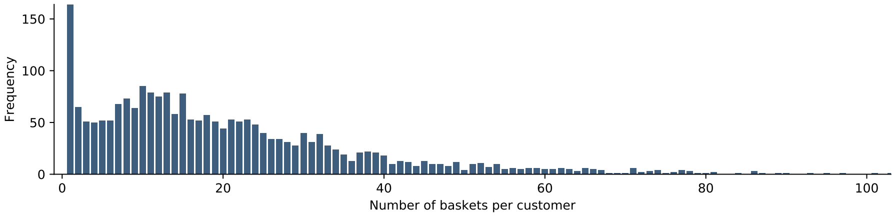


FIGURE 4.1 – *Frequency counts for the number of baskets per customer.*

In the model both customer-specific and basket-specific characteristics are included as explanatory variables. The predictors on the basket level all are all time related:

- `Year`: A binary variable that indicates if a basket is purchased during the first year (baseline) or second year of the observed period.
- `Month`: A dummy for the month in which the shopping trip was made (baseline is March).
- `Type of day`: A binary variable that differentiates baskets between weekdays (baseline) and weekends.
- `Time of purchase`: A binary variable that indicates if a shopping trip was made before 5pm (00:00–16:59, baseline) or after 5pm (17:00–23:59).

This set of variables could possibly be extended with external data sources, e.g. weather related variables such as temperature or amount of rainfall.
As predictors on the customer level we include the following characteristics:

- `Age`: Customer's age binned in 6 categories [0–34, 34–44, 45–54 (baseline), 55–64, 65+, Unknown].
- `Gender`: [Female, Male (baseline), Unknown].
- `Household size`: A binary variable that serves as a proxy for the size of the household [Small household: less than 2 individuals (baseline), Large household: 2 or more individuals].

4.4 Results

In this section the results for our estimated model structure are displayed and discussed. The model parameters are inferred via variational inference. We initialize the motivation-specific $\boldsymbol{\phi}_m$ vectors using the standard LDA model. This initialization approach is advocated in the literature (Gopalan et al., 2014). The results displayed are obtained after 15,000 iterations of our variational estimation routine, after which our optimization seems converged. All results displayed are based on the (variational) posterior means of the parameters, unless specified otherwise. Because the model is multi-faceted we will analyze the results by considering the model components separately and show how these results can be interpreted to gain insight in the high-dimensional purchase data. The setup of this section is as follows: We first describe a few of the motivations and highlight some of the motivation properties. Subsequently, we focus on the correlation structure among motivations. Next, we highlight the seasonality patterns for some of the motivations and finally, we discuss the effects of the remaining predictor variables and examine the persistence of the motivations.

4.4.1 *Motivations*

The number of motivations M is set by the researcher and typically should strike a balance between the interpretability of the model results (favoring lower values of M) and descriptive power of the model (typically favoring higher values of M). In our application we set $M = 100$, which is a common choice in the machine learning literature (Hoffman et al., 2013, Gopalan et al., 2014). We note that this is not a hard restriction and in case one is interested in finding the *right* number of motivations there are alternatives, such as using a hold-out sample of the data to determine the number of topics (Jacobs et al., 2016). However, this comes at a cost as the computational complexity will increase.

Each latent motivation in the model is represented by its own high-dimensional discrete distribution over the complete product assortment. Manually examining this entire vector for one of the motivations is both challenging and time-consuming as the assortment consists of 4266 products in total. However, we can typically infer the gist of a focal motivation by considering just the products that have received the highest probability under that motivation. In Tables 4.2 and 4.3 we display the 5 most likely products for two of the motivations that we inferred from the purchase data. Besides the product descriptions we also use the product taxonomy as specified by the retailer, consisting of the product's Group, Class, and Subclass. This information facilitates the interpretation and labeling of the motivations. If information such as a product taxonomy is not readily available, another option is to visually inspect a motivation's distribution by means of a word cloud, as suggested in Liu and Toubia (2017). For example, a world cloud could be created based on the words in the descriptions of the products in a motivation, where the size of each word is determined by the probability distribution for a motivation.

TABLE 4.2 – *The 5 most likely products for motivation 95.*

Motivation 95: Paint equipment and supplies

Group	Class	Subclass	Description	ϕ_{95}
PAINT	APPLICATORS	CHIP/FOAM APPLICATORS	CHIP 2.0 FLAT BRUSH	0.1392
PAINT	APPLICATORS	TRAYS/LINERS	9 IN PLASTIC TRAY LINER - WHITE	0.1284
PAINT	APPLICATORS	CHIP/FOAM APPLICATORS	CHIP 3.0 FLAT BRUSH	0.0607
PAINT	APPLICATORS	FRAMES/GRIDS	9 IN HD ROLLER FRAME - ORG HNDLE	0.0600
PAINT	APPLICATORS	TRAYS/LINERS	9 IN PLASTIC TRAY LINER 10PK - WHITE	0.0492

Motivation 95, displayed in Table 4.3, clearly captures a purchase pattern that is related to painting as the 5 most likely products are paint equipment and supplies. This is in line with the product taxonomy created by the retailer, as the 5 products are all placed in the paint group. We reiterate that the product taxonomy itself is not part of the model and hence it is not used to infer the motivations. Instead, these products all receive a high probability under the same motivation because they are relatively often purchased together across shopping baskets in the data.

Table 4.3 shows that the likely products for motivation 69 are related to gardening. However, note that the LEATHER PALM GLOVE - LARGE product, which is the fifth-most likely under this motivation, is not part of one of the gardening groups specified by the retailer. Instead, it is placed in the hardware group. This indicates two intrinsic properties of a motivation: i) We can derive extra meaning from the purchase of a general purpose product, such as protective gloves, when

TABLE 4.3 – *The 5 most likely products for motivation 69.*

Motivation 69: Gardening

Group	Class	Subclass	Description	ϕ_{69}
GARDEN/OUTDOOR	SOILS AND MULCH	BARKS & MULCHES	2 CU FT BARK NUGGETS	0.1636
GARDEN/OUTDOOR	SOILS AND MULCH	AMENDMENTS/COMPOST	40 LB COMPOSTED MANURE	0.1520
GARDEN/OUTDOOR	SOILS AND MULCH	POTTING/SPECIALTY SOILS	40 LB POTTING SOIL	0.1179
GARDEN/INDOOR	CLEANING	CLEANING CHEMICALS	CLOROX OUTDOOR BLEACH	0.1013
HARDWARE	SECURITY/SAFETY	WORK GLOVES	LEATHER PALM GLOVE - LARGE	0.0818

purchased in combination with gardening related products. ii) A motivation is able to capture purchase patterns that span across multiple product categories. Both properties of customer purchase behavior are easily overlooked if one solely focuses on individual products, or when one only examines products from a single category.

Similar to what we did above for motivation 69 and 95, we can label each inferred motivation by examining the products that receive the highest probability under that motivation. This provides a high-level insight into the general patterns that underlie the purchase data and in turn, it enables us to reason about purchase behavior in the motivation space instead of the product space. This significantly helps in interpretation, as by doing so we reduce the dimensionality from 4,266 products to just the 100 inferred motivations.

However, the motivations do not only differ in the products that they place emphasis on. They also vary in the number of relevant products that describe the motivation. The story that they convey can be a narrow one that involves just a few products from the assortment, or a more general one that concerns more products. To examine this we consider the cumulative probability for the 5 most likely products under a motivation. This figure serves as a proxy for the sparseness of a motivation's probability distribution, where a higher value suggests that the majority of the probability mass is distributed across fewer products, resulting in a topic that is focused on only a small set of products.

In Figure 4.2 we display the cumulative probability of the 5 most likely products within each motivation. The motivations are not sorted in a particular order. For most motivations, this figure is situated between 0.25 and 0.90, which indicates that they represent sparse to very sparse distributions over the products in the assortment. That is, the gist for each of these motivations can be described by a small set of products that receive high probability. Returning to our previous two motivations: For motivation 95 this cumulative probability is close to 0.40 while for motivation 69 it is over 0.60. These figures suggest that both motivations are sparse and allocate the majority of their mass to just a few products. The notable exception is motivation 100, for which the cumulative probability of the 5 most likely product is about 0.02. This suggests a distribution that is intrinsically different from the other 99 motivations, as its mass is spread out over many products in the assortment.

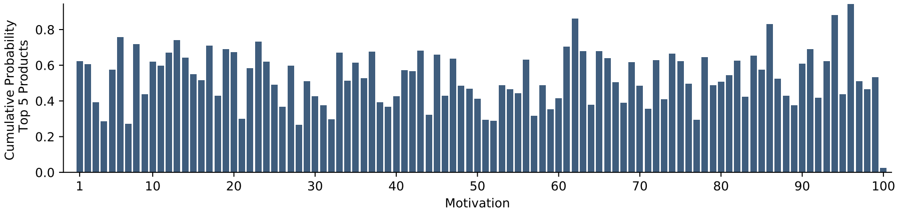
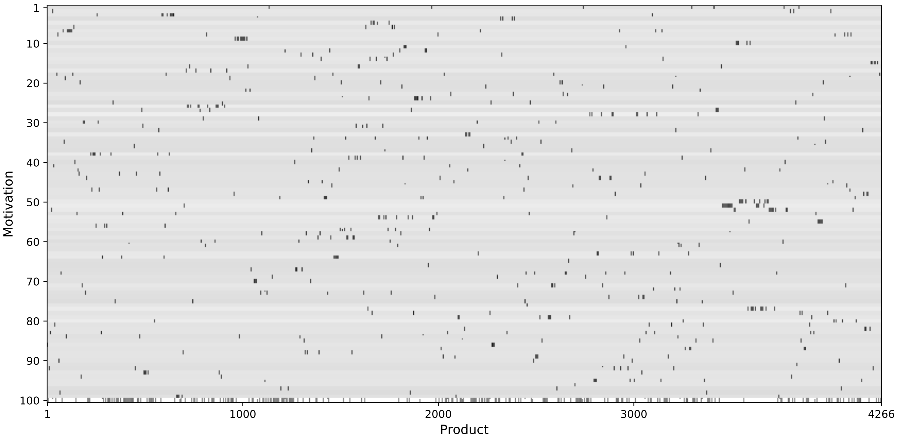


FIGURE 4.2 – *Cumulative probability of the 5 most likely products per motivation, using the posterior mean of each* $\boldsymbol{\phi}_m$ *vector.*

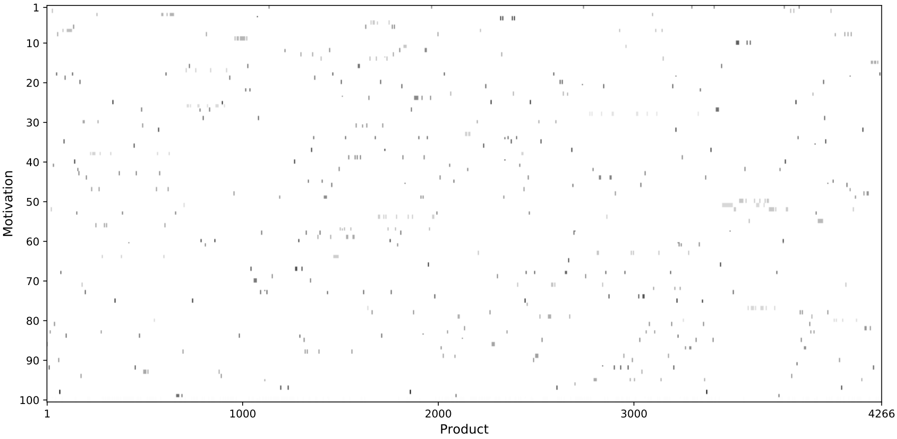
We can gain more insight in the sparsity of motivations by plotting a heat map of the $\boldsymbol{\phi}$ matrix, and we display two heat maps in Figure 4.3. Each row on the vertical axis represents one of the 100 motivations, while the horizontal axis represents the 4,266 products in the assortment, which are sorted in alphabetical order of their Group – Class – Subclass – Description.

The heat map we display in Figure 4.3a is the heat map for the natural log of the elements in $\boldsymbol{\phi}$. Here, a few things are noteworthy: First, for the majority of the rows, i.e. motivations,a limited number of dark bars per row is observed. This reconfirms that most motivations are sparse distributions, where just a few products receive significant probability mass. Second, for each row of the heat map the dark bars can either be clustered or they can be spread out more uniformly over the length of the bar. Take motivation 55 for example, for which we observe many dark bars clustered together. As the products are sorted in alphabetical order starting with the group name, this indicates that for motivation 55 the most likely products are probably from the same product group. On the other hand, for motivation 79 the dark bars are more scattered across the row. This indicates that the purchase pattern captured by this motivation is sparse, but likely spans across multiple product categories. Third, the heat map shows that motivation 100 is represented by a distribution that is diffuse, as we suspected. It places relatively high probability on a lot of different products in the assortment, displayed by many dark bars scattered across the bottom row.

The next heat map in Figure 4.3b is for the probabilities in $\boldsymbol{\phi}$ relative to the frequency of the products in the data. This measure provides additional insight by what factor a product's probability is *lifted* under a motivation, compared to the baseline in the data. We create such a lift measure by dividing the probabilities for product j for each of the 100 motivations by the relative frequency of product



(A) *Heat map of $\boldsymbol{\phi}$. For display purposes we increase the contrast by displaying the natural log of the probabilities.*



(B) *Heat map where each value in $\boldsymbol{\phi}$, i.e. ϕ_{mj} the probability of product j under motivation m, is divided by product j's relative frequency in the data set.*

FIGURE 4.3 – *Heat maps related to the values in $\boldsymbol{\phi}$, the 100×4266 matrix containing the probability distributions over the product assortment for each of the 100 motivations. The products on the horizontal axis are sorted in ascending order of their Group – Class – Subclass – Description. Darker shades in the heat map indicate larger values, while lighter shades represent lower values.*

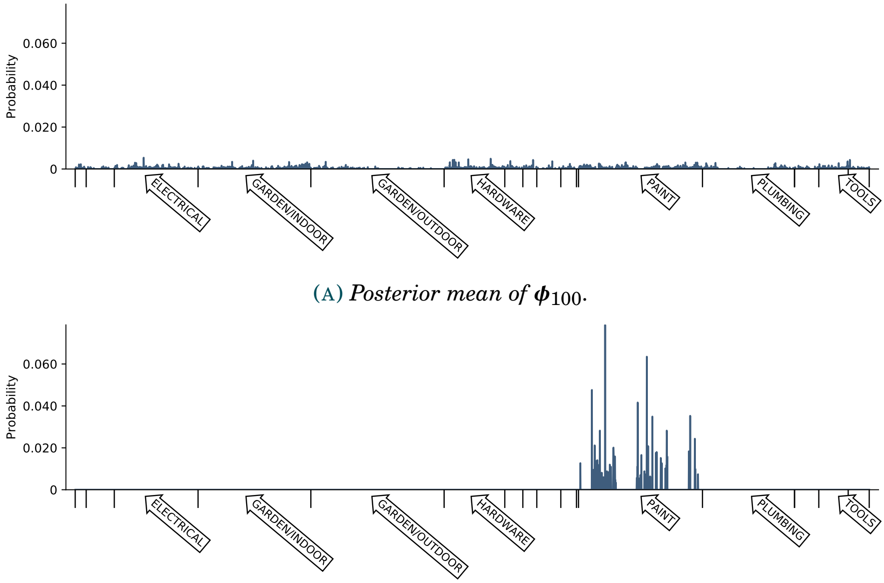
j in the data. The resulting ratios inform us by what factor product j is more or less likely under each of the motivations. The heat map of these ratios is displayed in Figure 4.3b. Lighter bars correspond to a ratio closer to zero, which indicates a decrease (or very small lifts) in probability compared to the relative frequency in the data. On the other hand, darker bars correspond to larger lifts, indicating that this product is very informative for that motivation.

The general outline is similar to the heat map in Figure 4.3a because the location of the colored bars is approximately the same. However, the intensity of the darker bars can now be different as the magnitude of a product's probability under a motivation does not necessarily translate one-to-one to the magnitude of its lift. Consider motivation 100 for example: This distribution spreads its mass across a lot of products, captured in Figure 4.3a by dark bars for these products. However, for the same products we observe very light, close to white, bars in Figure 4.3b. This indicates that the lift in probability for these products under motivation 100 is not large. Combining this with the observation that these products are displayed by dark bars in Figure 4.3a, their relative frequency in the data should be relatively high. This seems to suggest that motivation 100 is not focused on any products from the assortment in the particular.

In case one is interested in a specific motivation, the probability distribution over products for that motivation can be examined. Effectively, this is the same as zooming in on one of the rows of the heat map of $\boldsymbol{\phi}$. Since motivation 100 seems to be an outlier we inspect $\boldsymbol{\phi}_{100}$, its probability vector over the products, in more detail in Figure 4.4a. This is contrasted against $\boldsymbol{\phi}_{28}$ for motivation 28 in Figure 4.4b. As before, the products on the horizontal axis are sorted alphabetically on Group – Class – Subclass – Description. That is, the smaller the distance between two products on the horizontal axis, the more likely it is they belong to the same group or (sub)class. For further insight we have annotated the plots in Figure 4.4 with the 7 product groups that contain the most products in the assortment.

For motivation 28 we observe that virtually all its probability mass is located in the paint product group. In addition, within the paint group the likely products appear to be clustered, which suggests that this motivation is concerned with a subset of the product classes that fall under the paint group. After manually inspecting motivation 28, it seems to concern both interior paint and paint applicator products. This is in contrast to motivation 100, where the mass is more evenly spread over the assortment.

The last property of the motivations that we consider is their relative size in the customer base. Intuitively it makes sense that some motivations may be more generic and appeal to more customers than others. A measure for the size of a motivation on the population level is the posterior mean of $\boldsymbol{\mu}_{\kappa}$, which captures the baseline relevance of the motivations. We transform this posterior mean using the Softmax transformation to proportions between [0, 1]. The result



(A) *Posterior mean of* $\boldsymbol{\phi}_{100}$.

(B) *Posterior mean of* $\boldsymbol{\phi}_{28}$.

FIGURE 4.4 – *Bar plots of the product probabilities under motivations* 100 *and* 28*. The products on the horizontal axis are sorted in ascending order of their Group – Class – Subclass – Description. In addition, the horizontal axis is divided by the 17 product groups from the product taxonomy as specified by the retailer. The 7 groups that contain the most products are annotated.*

is displayed in Figure 4.5.

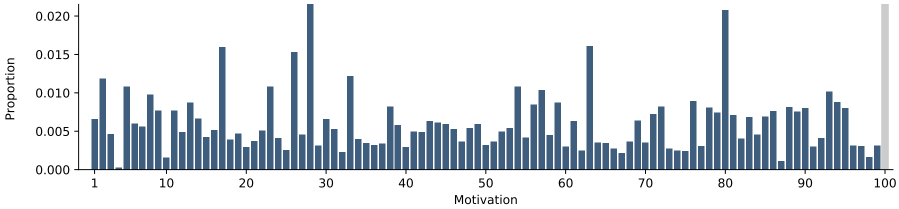


FIGURE 4.5 – *Size of motivations measured by* SOFTMAX($\boldsymbol{\mu}_\kappa$) *for the variational posterior mean of* $\boldsymbol{\mu}_\kappa$*. Note: The proportion of motivation 100 (equal to 0.398) is clipped in this figure.*

The majority of the proportions range from 0.002 to 0.02 and if we contrast this to the baseline proportion ($M^{-1} = 0.01$) this seems reasonable. For example, one of the larger motivations is 80, which is about general tools such as wrenches and screw drivers. It seems likely that these are relevant for many customers. One of the smaller motivations is motivation 87 which is related to moving as it describes products such as moving boxes and packaging tape. This result indicates that these products are only relevant for a small set of customers. We remark that these reported proportions are on the population level and it can be expected that we will observe more sparse proportions the closer we move to the data, e.g. if we examine the proportions for a customer or on the basket-level.

The notable exception in Figure 4.5 is again motivation 100, which is represented by a shaded bar. This bar is clipped because motivation 100 has a proportion of 0.398, which is an order of magnitude larger than the other motivations that we have discovered. We can combine this observation with our other findings about motivation 100, i.e. that it is represented by a diffuse distribution that spreads its mass across many products that tend to be relatively frequently purchased in the data, to provide a possible explanation for the occurrence of motivation 100 in our results: It serves as an aggregate motivation in the model, which might be needed to compensate (or allow) for the more extreme distributions we find that belong to the other 99 motivations.

4.4.2 *Motivation correlations*

The covariance matrix for $\boldsymbol{\kappa}_i$, $\boldsymbol{\Sigma}_\kappa = \boldsymbol{\Lambda}_\kappa^{-1}$, describes which motivations are likely to be jointly (ir)relevant for an average customer. For interpretation purposes it is convenient to consider correlations instead of the covariances directly. The dimension of this correlation matrix is 100×100 as we inferred 100 motivations in total. Clearly, it is not feasible to separately evaluate each of the correlation pairs in this matrix.

Instead we again use a heat map to visualize this large matrix and get an intuition for the general correlation structure underlying the motivations. This heat map of the motivation correlation matrix is displayed in Figure 4.6. Roughly speaking, the resulting heat map can be separated in three groups: The positive correlations are indicated by shades of red, the negative correlations by blue, and the correlations that are absent are colored green. Note that the majority of the cells in Figure 4.6 are displayed in green, describing pairs of motivations that have a low correlation, i.e. if one motivation out of a pair is relevant for a customer, that information is not informative about whether or not the other motivation will be relevant as well.

The more interesting motivation pairs are those that exhibit strong positive correlations. After manually examining the motivations that have some of the highest correlations, we notice that many of them revolve around gardening such as motivations 5, 6, 11, 12, 13, and 14. Remember that the correlations

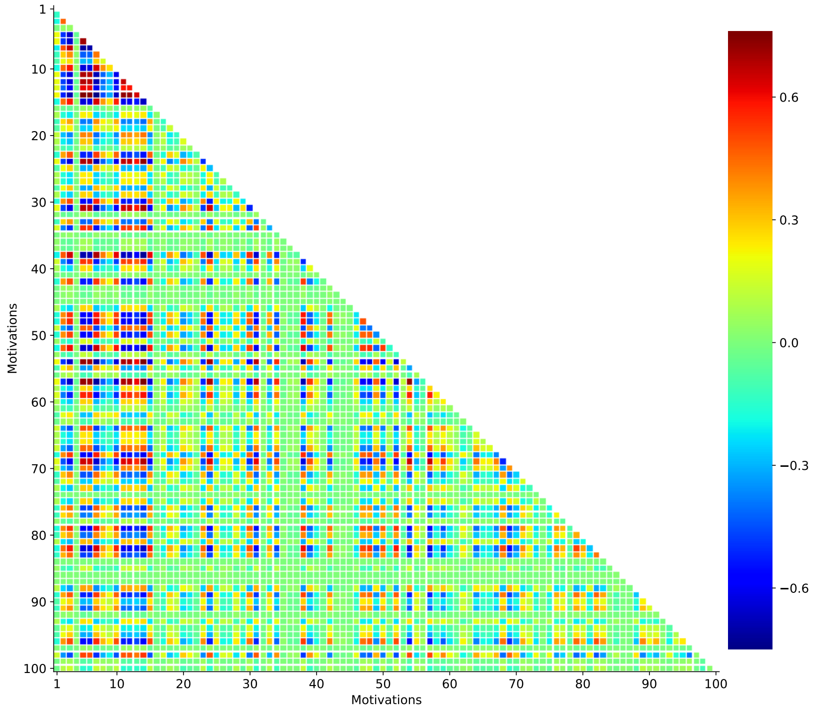


FIGURE 4.6 – *Heat map of the lower-diagonal of the correlation matrix that is obtained using the posterior mean of $\boldsymbol{\Sigma}_\kappa$, which is the covariance matrix of $\boldsymbol{\kappa}_i$. Each cell in the heat map represents the correlation between two motivations. If the color shade of the cell is closer to red, this indicates a high correlation between the motivations, and if it is closer to blue it indicates a lower correlation. Motivations that are uncorrelated are displayed by shades of green.*

are specified on the customer-base level and one explanation for these positive correlations is that a customer who is active in his garden may be interested in multiple gardening related “projects”. Apparently these separate garden projects are not aggregated to a single comprehensive motivation in the model. Recall that the identification of the motivations is driven by the co-occurrence of products in a shopping basket. Hence, a reasonable explanation for this observation is that these gardening projects are spread out over several shopping trips. For example, because there is limited space in the garden and the gardening project needs time to complete, or because the gardening project is bound to a certain season. This suggests that even though more than one gardening motivation may be relevant across a customer’s entire purchase history, this is not necessarily the case at the level of a purchase occasion. In fact, some of these gardening motivations may have a negative correlation on the shopping basket level because of the reasons outlined above. Furthermore, it makes intuitive sense that gardening is a recurrent activity, making it relevant for multiple shopping baskets in a customer's purchase history. This is in contrast to a motivation that may only be relevant once for a customer, such as a bathroom renovation.

The motivation correlation structure can also be used to predict which other motivations might be of interest to a customer. To determine this we consider a plot inspired by an ROC (Receiver Operating Characteristic) curve that uses the $\textsc{Softmax}(\boldsymbol{\alpha}_{ib})$ values, i.e. the probabilities over motivations for the b-th basket of customer i. The intuition here is that a pair of motivations A and B that is highly correlated at the customer-base level, i.e. represented by a large value in $\boldsymbol{\Sigma}_{\kappa}$, will have predictive power across baskets of a customer. We assess this by dividing the baskets in our data set between the first and second year of our sample. The customers are sorted in descending order of average probability for motivation A in the first year, where the average is taken over a customer's baskets in the first year. Subsequently, for the second year the cumulative sum of average probabilities for motivation B is calculated, where the aforementioned ordering to calculate the cumulative sum is used. These results can be used to construct an ROC-like curve. We transform the values on the vertical axis to the $[0,1]$-interval, by dividing the values on the vertical axis by the total sum of average probabilities for motivation B. The values on the horizontal axis are transformed to the $[0,1]$-interval as well, such that they represent the customer percentiles obtained by ordering on the average probability for motivation A in the first year.

An ROC curve is created for a specific pair of positively correlated motivations. The motivations considered are 6, which is about the construction of gypsum walls, and 9, which is involved with bathroom renovation. The posterior mean of the correlation between these motivations is 0.63. This number is relatively high, especially as it is not to be expected that this correlation is perfect: Gypsum walls can be needed in other projects, and a bathroom can also be renovated without constructing new walls.

The resulting ROC curve is displayed in Figure 4.7 by the dark graph and it is clearly situated above the 45-degree-line through the origin. To put this lift in perspective a similar curve where we plotted the average probability for motivation 6 in the first year against the second year is created. This can be interpreted as an "auto"-ROC curve for motivation 6. It is displayed by the light gray graph in Figure 4.7. We note that the ROC curve for motivations 6 and 9 is relatively close to this auto-ROC curve. This indicates that customers who have a high probability for motivation 6 in the first year, tend to have a high probability for motivation 9 in the second year. This finding supports the claim that the motivation correlations on the customer-base level are informative on the basket-level as well.

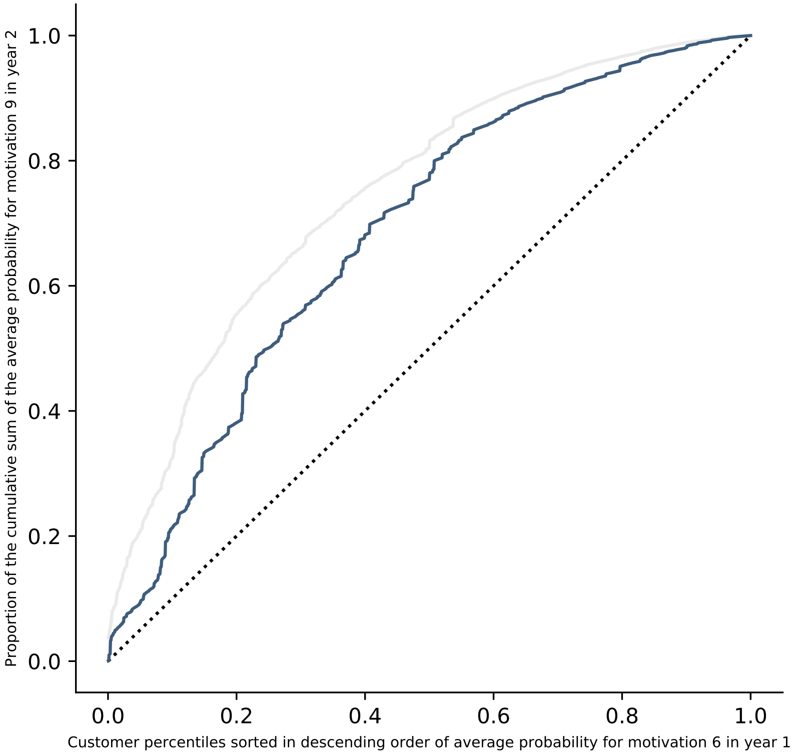


Figure 4.7 – *The dark graph represents the ROC curve between motivation 6 (construction of gypsum walls) and motivation 9 (bathroom renovation). The customer percentiles on the horizontal axis are based in descending order on a customer's average probability for motivation 6 in the first year of the data set. The vertical axis represents the proportion of the cumulative sum for the average probabilities of motivation 9 in the second year of the data set. The light gray graph displays the ROC curve for motivation 6 in the first year against itself in the second year.*

4.4.3 *Motivation seasonality*

In the model, the baskets of a customer are not aggregated to a single purchase history. Instead we consider them as distinct shopping trips. This has enabled us to include monthly seasonality effects as a predictor for $\boldsymbol{\alpha}_{ib}$, the relevance of each motivation in the b-th basket of customer i. However, directly evaluating the monthly effects for each of the 100 motivations is not practical. Instead, we choose to highlight the seasonality pattern over the 12 months for 6 of the motivations in Figure 4.8. In these plots, the horizontal axis represents the 12 calendar months, starting in January and ending in December. The vertical axis is the ratio of the average probability for a motivation in each month, with respect to the average probability of the motivation. The averages are taken over the $\textsc{Softmax}(\boldsymbol{\alpha}_{ib})$ values across baskets and customers. The labels applied to these 6 motivations are obtained in a similar way as at the beginning of Section 4.4.1, i.e. for each motivation we examined the most likely products to infer meaning.

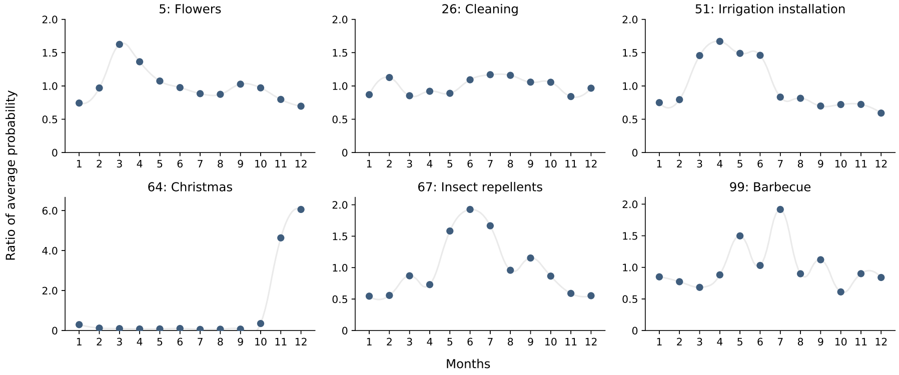


FIGURE 4.8 – *Plots for six motivations that display the ratio of a motivation's average monthly probability, with respect to the average probability for that motivation. The average probability is calculated based on the* $\textsc{Softmax}(\boldsymbol{\alpha}_{ib})$ *values across baskets and customers.*

The motivation with the strongest seasonality effect is number 64, which is about products related to Christmas and the holiday season. We observe that this motivation has a very low baseline relevance outside the holiday season. We discover a different seasonality pattern for motivation 67, which contains insect and mosquito repellents. The plot shows that the relevance of this motivation peaks during the summer months and intuitively, it makes sense that these products are in higher demand during these months. For motivation 26, which places emphasis on general cleaning products, we hardly see any variation across the year. This is in agreement with the notion that cleaning is not restricted to a certain season. The other motivations can be interpreted in a similar way: Flowers are planted during spring, irrigation systems tend to be installed before summer really starts and barbecue products are in higher demand during summer compared to winter. All these results show that some topics exhibit (strong) seasonality patterns that can be used to improve effectiveness of marketing campaigns related to these motivations.

4.4.4 *Effect of other explanatory variables*

Besides the seasonality effects discussed in Section 4.4.3, the model contains several other explanatory variables at the basket-level as well as variables that are customer-specific. Each of these variables is able to shift $\boldsymbol{\alpha}_{ib}$, the baseline relevance of the motivations for the b-th basket of customer i. If such a shift is positive (negative) for a motivation, it indicates that the explanatory variable has a positive (negative) effect on the relevance of the focal motivation. This provides the retail manager with additional input to determine her marketing strategies. For example, a group of customers with a certain characteristic can be targeted in a promotion because they are on average more interested in a certain motivation.

The effect of an explanatory variable is calculated using the approach outlined in Section 4.2.2. As each of our explanatory variables is a dummy, we interpret these effects relative to that of the baseline value for the dummy. This allows us to interpret the effect as a "partial" effect that describes the shift in relevance for a motivation induced by the focal variable, relative to its baseline level, ceteris paribus.

We first consider the basket-specific explanatory variables that we included in the model besides the seasonality effects which we discussed in Section 4.2.2. Three dummy variables remain, namely `Year 2` vs. `Year 1`, `Weekend` vs. `Weekday` and `17:00-23:59` vs. `00:00-16:59`. The corresponding shifts in average probability for each of the motivations relative to their baseline values are displayed in Figure 4.9. Note that the scale, i.e. effect size, on the horizontal axis differs across the subplots.

The first dummy discussed is `Weekend` vs. `Weekday`. It captures the shift in motivation relevance between shopping trips during a regular day of the week (Monday to Friday) compared to a day in the weekend (Saturday and Sunday). One of the motivations that has a large negative shift is motivation 87 which relates to products involved in moving. In this motivation, products such as moving boxes, packaging tape, and bubble cushion are important. Intuitively, it makes sense to assume that on average people prepare for a move on weekdays, and perform the actual move during the weekend. We can contrast this finding

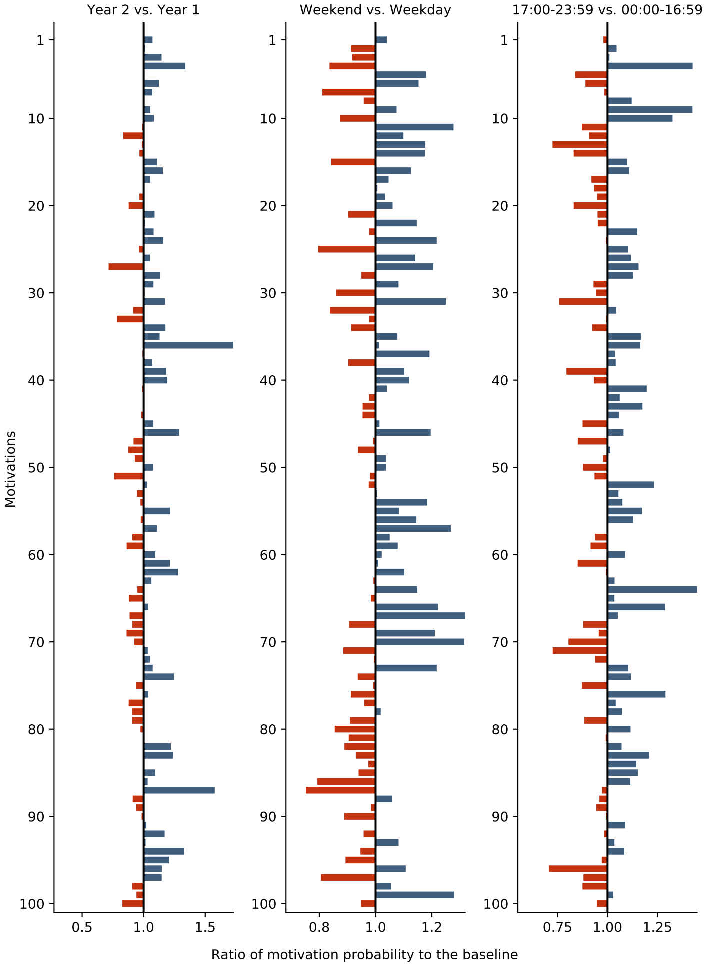


FIGURE 4.9 – *Impact on motivation likelihood for the basket-specific explanatory variables versus the baseline likelihood. The horizontal axis represents the ratio of these probabilities. Note that the scale for this axis differs across the subplots.*

against motivation 69, which relates to pool maintenance. This motivation has the largest positive shift for this dummy variable, suggesting that supplies for pool maintenance are on average more often purchased during weekends than during weekdays.

The second basket-specific variable that we consider is `17:00-23:59` vs. `00:00-16:59`, which divides the shopping trips in a group that occurred before 5pm and a group that occurred after 5pm. The largest positive shift is observed for motivation 8, which places emphasis on food products such as candy bars and sodas. It seems plausible that these products are in higher demand after 5pm as these shopping trips co-occur with dinnertime.

For the explanatory variables on the individual level a similar analysis is performed. Because of space constrains we will not discuss all the variables but instead focus on the following 4 dummy variables: `Age 0-34` vs. `Age 45-54`, `Age 65+` vs. `Age 45-54`, `Female` vs. `Male`, and `Large household` vs. `Small household`. The shifts in the average probabilities for each of these variables relative to their baseline values are displayed in Figure 4.10.

We first consider the effects for the age related dummies that segment the customer base in different age groups: `Age 0-34` vs. `Age 45-54` and `Age 65+` vs. `Age 45-54`. These dummies have a substantial impact on the likelihood of motivations and we notice that for a large share of the motivations the effect for these two dummies are reversed. For example, consider motivation 14, where the average probability is doubled for customers in the 65+ age group compared to the baseline age group of $45-54$, while in contrast it is lowered by 20% for the youngest age group. After examining the contents of this motivation, it turns out that it places emphasis on perennial flowers, which are flowers that are generally easy to maintain. This finding suggests that these type of flowers are more in line with the gardening needs of the elderly.

The next variable separates the customer base on gender: `Female` vs. `Male`. Large positive shifts for motivations 5, 14, and 24 are observed, which are all related to gardening, suggesting that these motivations are more relevant for the female customers. In contrast, large negative shifts are observed for motivation 36, which is related to interior lighting, and motivations 38 and 47, which are related to electrical tools and electrical boxes, conduits, and fittings. These findings are well aligned with prevailing gender stereotypes.

The last customer characteristic that we consider describes the household size: `Large household` vs. `Small household`. The motivations that receive large positive shifts for large households seem primarily concerned with outdoor "projects" such as gardening and barbecuing. Perhaps this variable acts as a proxy for home size or type, as larger families tend to live in family homes that often come with a garden, in contrast to an apart which is more likely to house a smaller household.

To summarize this section, the explanatory variables in the model clearly provide new insights in the latent purchase patterns of the data. It would be difficult to obtain these insights by just considering the raw data. These extra insights are primarily enabled through our use of motivations, which are coherent groups of products, instead of analyses focused on single products from

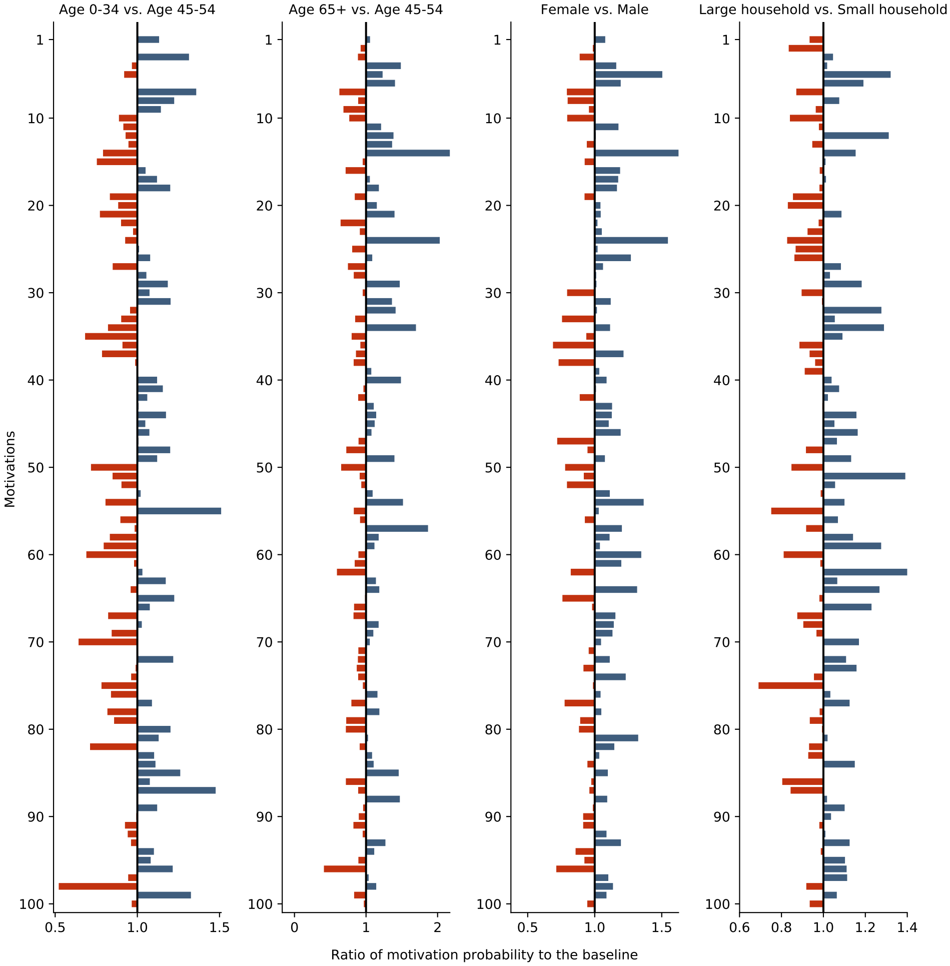


FIGURE 4.10 – *Impact on motivation likelihood for the customer-specific explanatory variables versus the baseline likelihood. The horizontal axis represents the ratio of these probabilities. Note that the scale for this axis differs across the subplots.*

the assortment. In this way, it becomes easier to discern distinctive patterns and find plausible explanations for the purchase behavior in the data, and connect these patterns to explanatory variables. Finally, we want to emphasize that the effects reported in this section are all "partial". A retail manager can study all possible combinations of variables needed to answer a research question at hand.

4.4.5 *Motivation persistence*

The last set of effects that we consider concerns the persistence of motivations across shopping trips. These effects provide insight in how the relevance of a motivation in a current basket affects the relevance of that motivation in the next basket. Motivations with a higher persistence tend to be relevant across more shopping trips and “last longer”. For each motivation we measure the persistence as follows: We first create an average baseline as discussed in Section 4.2.2. Subsequently, we shift the focal motivation’s relevance with 2 posterior standard deviations and measure how the average relevance changes compared to the baseline. The results are displayed in Figure 4.11.

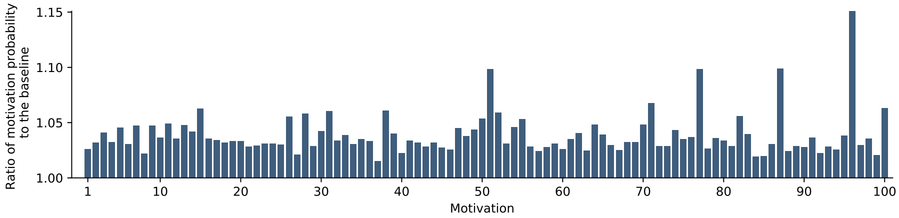


FIGURE 4.11 – *Impact on motivation likelihood, by applying a shift of two times the posterior mean of the standard deviation* $\sqrt{\tau_m^{-1}}$ *for each motivation* m. *The vertical axis represents the ratio of these probabilities versus the baseline.*

Note that it holds for all motivations that a positive shock results in positive effects. That is, the average probability for the motivation is increased in the next period, in line with our expectations. The motivation that displays the largest persistence effect is motivation 96. The likely products under this motivation relate to rental tools, so it could be that motivation 96 actually describes a multi-basket purchase pattern. In the first shopping trip, equipment is rented and a deposit is paid. In the second trip, the equipment is returned, rental fees are charged and the deposit is refunded to the customer. The motivation with the second-largest persistence effect is motivation 51, which is about the installation of an irrigation system in the garden. This seems like a large project, that could easily require multiple trips to the retailer. The important point to take away is that even though the level of persistence varies over motivations, it is an additional component of the model that can be used to predict which motivations are relevant during the next shopping trip. Being able to accurately predict what a customer wants at a given point in time is a vital point in retail management.

4.5 Managerial implications and future research

In this chapter we have presented a novel method that can be used to gain insight in a retailing context where product assortments are so large that traditional analysis methods fall short because of this dimensionality. This typically happens when dealing with more than a hundred products. The method that we have presented scales well with the size of the assortment, providing the opportunity to analyze very large assortments. The application in this chapter involved purchases from an assortment of over 4000 products. In principle, however, the method scales well with the number of products and should work with assortments of a larger order of magnitude.

The cornerstone of our method is to infer a limited number of motivations that can be used to adequately describe the purchase behavior for all customers across all products in the assortment. For each motivation we can describe which products are relevant, given that the focal motivation is active. The intuition behind the scalability of our method is that the number of motivations is much smaller than the size of the product assortment, effectively reducing the dimensionality of our problem.

The model was applied to a large data set containing purchase data from a retailer that wishes to remain anonymous. The data consists of about $50,000$ shopping baskets that contain purchases from an assortment of over 4,000 products. The results are intuitive and show a high degree of internal consistency, providing face validity to the model. In addition, in Jacobs et al. (2016) it has been shown that latent Dirichlet allocation (LDA), which underlies our model as well, is well-suited for predicting a customer's next purchase with high accuracy. However, in this chapter we have specified a much richer model structure that can be leveraged to gain insight in the purchase behavior of a customer base that goes beyond inferred motivations and purchase prediction. From these results several managerial implications can be derived, which will be discussed next.

First, we consider the separate purchase occasions in the model as they occur in the data set. We specifically do not aggregate shopping baskets into a single purchase history. The implication of this step is that each purchase occasion is assigned a unique time stamp. In turn, this enables us to connect the variation in relevance of a motivation to different time periods in the year, e.g. months or seasons. For a manager, this provides insight in when a motivation becomes "active" and relevant. Some of these seasonality effects may be known to the manager a priori, but a model-based approach allows for the discovery of unknown patterns, as well as quantifying when the seasonality effect starts and ends. In addition, we are able to capture time effects during the day. Which motivations are likely active for shoppers during working hours, and which after working hours? These results can be used to create marketing campaigns that advertise or promote the right products at the right time.

Second, the model specification includes additional explanatory variables at the customer level such as a customer's gender and age. This enables us to derive which motivations are (ir)relevant across segments of the customer base. A retail manager can use this information to create personalized marketing messages for customers, based on their observed characteristics and their inferred motivation relevance. This helps in the identification and selection of products that are likely relevant for the customer. This approach can also be reversed: If the manager is interested in promoting a specific product from the assortment for which we know that it is relevant under a specific motivation, the manager can target the customers in the segments for which this motivation is relevant and hence, who are likely interested in the product.

Third, we estimate the persistence of motivations in our model. This enables us to infer if a motivation is likely relevant for multiple adjacent shopping trips of a customer. Not only does this potentially improve the predictive power of the model, it also provides a way to assess the involvement of a customer in a certain motivation. Based on this information a retail manager could entice a customer to "complete" the motivation, for example by offering a discount.

Fourth, we allow the motivations in the model to correlate. From these correlations the motivations can be determined that are (un)likely to both be relevant for a customer. This information could be used by a marketing manager to identify potential cross-selling opportunities, where products that are relevant in highly correlated motivations are jointly promoted. In addition, the correlations increase the explanatory power of the model. This holds in particular for a customer for which we only observe a limited number of purchases. By leveraging the motivation correlation structure, we can identify another motivation that can be of interest to the customer, even in case we have not observed any purchases that relate to that motivation.

The aforementioned points mainly focus on effects at the population level. However, the model captures heterogeneity in purchase behavior as well, as customers are individually represented in the model. This allows us to create customer-specific predictions for products a customer could be interested in. As the baskets are considered separately, we are able to visualize a customer's journey at the retailer. In this way we can track a focal customer over time, and examine how his preferences change over time.

Although our approach already provides many valuable insights, there are several interesting extensions that can be addressed in future research. First of all, we fixed the number of motivations in the model to be equal to 100. Ideally, this number is determined in a more sophisticated way and substantiated by empirical evidence such as a hold-out likelihood to prevent overfitting to the data. However, even if one is not too concerned with overfitting, the determination of the number of motivations remains a complex task. A goal of this chapter was to show that the (latent) motivations can be used to provide insight in high-dimensional purchase data. By increasing the number of motivations without bound, this insight might become blurry and more difficult to grasp. That is, we believe that the *right* number of motivations in the model should strike a balance between model performance, i.e. how well the model is able to describe the observed data, and the usability and implementability of the model results by a manager. At the moment, this is an open question but the development of techniques that are able to strike this balance provides an interesting avenue for further research.

Second, in this chapter we considered purchase history data and explanatory variables at the basket- and customer-level. However, typically there is much more relevant data available, especially at online retailers. On the customer side, additional information might be collected, such as browse and search history and product ratings. On the other hand for products, additional information is often available in the form of product features that can be more quantitative (brand, price, etc.), or qualitative such as product descriptions and reviews. Incorporating these additional sources of information in a meaningful way could further improve both interpretability and actionability of the results.

Finally, the rich model structure specified in this chapter lends itself for more sophisticated research applications. In this chapter we have mainly focused on an exploratory analysis of the high-dimensional purchase data. This is the first step of the research funnel when working with a complex data set that is concerned with a large product assortment that is unwieldy to analyze using traditional methods. The logical next step in this funnel would be to apply these unique results in a predictive or prescriptive analysis, leading to actionable insights and added benefits for both the customer and the firm.

Bibliography

Ansari, A., Li, Y., and Zhang, J.Z. Probabilistic Topic Model for Hybrid Recommender Systems: A Stochastic Variational Bayesian Approach. SSRN Scholarly Paper ID 2916514, Social Science Research Network, Rochester, NY, 2016.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y.W. On Smoothing and Inference for Topic Models. In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 27–34. 2009.

Baydin, A.G., Pearlmutter, B.A., Radul, A.A., and Siskind, J.M. Automatic differentiation in machine learning: a survey. *arXiv:1502.05767 [cs]*, 2015. ArXiv: 1502.05767.

Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Blei, D.M. Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annual Review of Statistics and Its Application*, 1(1):203–232, 2014.

Blei, D.M., Kucukelbir, A., and McAuliffe, J.D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Blei, D.M. and Lafferty, J.D. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Bodapati, A.V. Recommendation Systems with Purchase Data. *Journal of Marketing Research*, 45(1):77–93, 2008.

Box, G.E.P. and Hunter, W.G. A Useful Method for Model-Building. *Technometrics*, 4(3):301–318, 1962.

Buntine, W. Variational Extensions to EM and Multinomial PCA. In: T. Elomaa, H. Mannila, and H. Toivonen, editors, *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings*, pp. 23–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. DOI: 10.1007/3-540-36755-1_3.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017.

Damien, P., Wakefield, J., and Walker, S. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(2):331–344, 1999.

Dew, R. and Ansari, A. Bayesian Nonparametric Customer Base Analysis with Model-based Visualizations. SSRN Scholarly Paper ID 2692307, Social Science Research Network, Rochester, NY, 2017.

Fader, P.S. Integrating the Dirichlet-Multinomial and Multinomial Logit Models of Brand Choice. *Marketing Letters*, 4(2):99–112, 1993.

Fader, P.S. and Hardie, B.G.S. Modeling Consumer Choice Among SKUs. *Journal of Marketing Research*, 33(4):442–452, 1996.

Fader, P.S. and Schmittlein, D.C. Excess Behavioral Loyalty for High-Share Brands: Deviations from the Dirichlet Model for Repeat Purchasing. *Journal of Marketing Research*, 30(4):478–493, 1993.

Fleder, D. and Hosanagar, K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science*, 55(5):697–712, 2009.

Gelfand, A.E. and Smith, A.F.M. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2007.

Ghose, A., Ipeirotis, P.G., and Li, B. Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue. *Management Science*, 60(7):1632–1654, 2014.

Goodhardt, G.J., Ehrenberg, A.S.C., and Chatfield, C. The Dirichlet: A Comprehensive Model of Buying Behaviour. *Journal of the Royal Statistical Society. Series A (General)*, 147(5):621–655, 1984.

Gopalan, P.K., Charlin, L., and Blei, D. Content-based recommendations with Poisson factorization. In: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 3176–3184. Curran Associates, Inc., 2014.

Gourville, J.T. and Soman, D. Overchoice and Assortment Type: When and Why Variety Backfires. *Marketing Science*, 24(3):382–395, 2005.

Griffiths, T.L. and Steyvers, M. Finding scientific topics. In: *Proceedings of the National Academy of Sciences*, volume 101, pp. 5228–5235. 2004.

Grover, R. and Srinivasan, V. A Simultaneous Approach to Market Segmentation and Market Structuring. *Journal of Marketing Research*, 24(2):139–153, 1987.

Guadagni, P.M. and Little, J.D.C. A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science*, 2(3):203–238, 1983.

Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

Hoffman, M.D., Blei, D.M., Wang, C., and Paisley, J. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Hoffman, M.D. and Gelman, A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Jacobs, B.J.D., Donkers, B., and Fok, D. Model-Based Purchase Predictions for Large Assortments. *Marketing Science*, 35(3):389–404, 2016.

Jain, D., Bass, F.M., and Chen, Y.M. Estimation of Latent Class Models with Heterogeneous Choice Probabilities: An Application to Market Structuring. *Journal of Marketing Research*, 27(1):94–101, 1990.

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

Jeuland, A.P., Bass, F.M., and Wright, G.P. A Multibrand Stochastic Model Compounding Heterogeneous Erlang Timing and Multinomial Choice Processes. *Operations Research*, 28(2):255–277, 1980.

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., and Saul, L.K. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D.M. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18:14:1–14:45, 2017.

Kullback, S. and Leibler, R.A. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Liu, D.R., Lai, C.H., and Wang-Jung. A hybrid of sequential rules and collaborative filtering for product recommendation. *Information Sciences*, 179(20):3505–3519, 2009.

Liu, J. and Toubia, O. A Semantic Approach for Estimating Consumer Content Preferences from Online Search Queries. SSRN Scholarly Paper ID 2705069, Social Science Research Network, Rochester, NY, 2017.

Maddala, G. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, 1983.

McFadden, D. The Choice Theory Approach to Market Research. *Marketing Science*, 5(4):275–297, 1986.

Mimno, D., Hoffman, M.D., and Blei, D.M. Sparse stochastic inference for latent Dirichlet allocation. In: *Proceedings of the Twenty-Ninth International Conference on Machine Learning*, pp. 1599–1606. 2012.

Minka, T.P. Expectation Propagation for Approximate Bayesian Inference. In: *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pp. 362–369. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

Naik, P., Wedel, M., Bacon, L., Bodapati, A., Bradlow, E., Kamakura, W., Kreulen, J., Lenk, P., Madigan, D., and Montgomery, A. Challenges and opportunities in high-dimensional choice data analyses. *Marketing Letters*, 19(3):201–213, 2008.

Neal, R.M. *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., and Granka, L. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 2007.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

Ramage, D., Dumais, S., and Liebling, D. Characterizing Microblogs with Topic Models. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 130–137. 2010.

Ranganath, R., Gerrish, S., and Blei, D.M. Black Box Variational Inference. *arXiv:1401.0118 [cs, stat]*, 2013. ArXiv: 1401.0118.

Rossi, P.E., Allenby, G.M., and McCulloch, R. *Bayesian statistics and marketing*. John Wiley & Sons, 2012.

Salganik, M.J., Dodds, P.S., and Watts, D.J. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856, 2006.

Steyvers, M. and Griffiths, T. Handbook of Latent Semantic Analysis. pp. 424–440. Psychology Press, 2013.

Train, K.E. *Discrete choice methods with simulation*. Cambridge University Press, 2 edition, 2009.

Tran, D., Kucukelbir, A., Dieng, A.B., Rudolph, M., Liang, D., and Blei, D.M. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv:1610.09787 [cs, stat]*, 2016. ArXiv: 1610.09787.

Trusov, M., Ma, L., and Jamal, Z. Crumbs of the Cookie: User Profiling in Customer-Base Analysis and Behavioral Targeting. *Marketing Science*, 35(3):405–426, 2016.

Wagner, U. and Taudes, A. A Multivariate Polya Model of Brand Choice and Purchase Incidence. *Marketing Science*, 5(3):219–244, 1986.

Wainwright, M. and Jordan, M. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends(r) Mach. Now Publishers, 2008.

Wallach, H.M., Mimno, D., and McCallum, A. Rethinking LDA: Why Priors Matter. In: *Advances in Neural Information Processing Systems 22*, pp. 1973–1981. 2009.

Wedel, M. and Kamakura, W.A. *Market segmentation: conceptual and methodological foundations*. Kluwer Academic Publishers, 2 edition, 2000.

Wedel, M. and Kannan, P.K. Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, 80(6):97–121, 2016.

Xu, Y.C. and Kim, H.W. Order Effect and Vendor Inspection in Online Comparison Shopping. *Journal of Retailing*, 84(4):477–486, 2008.

Zanutto, E.L. and Bradlow, E.T. Data pruning in consumer choice models. *Quantitative Marketing and Economics*, 4(3):267–287, 2006.

Summary

Over the past two decades online retailing has become ubiquitous and today's large online retailers enable customers to purchase virtually any product. As a consequence product assortments at such retailers are of a different order of magnitude compared to the traditional brick-and-mortar stores. In this dissertation model-based methods are presented that can be used to model purchase decisions in such high-dimensional product assortments. These methods are able to accurately predict at the individual customer level which product will be purchased next out of the large assortment. In addition, the methods provide substantive insights in the patterns that underlie the observed purchase behavior. The applicability of such methods in practice hinges on their scalability and this holds especially true for online retailers. Model results should be rapidly obtained and the estimation time should not significantly increase in case the customer base or product assortment expands. Scalability is therefore a focal point in this dissertation. The methods introduced are adaptations and extensions of fast scalable methods from the machine learning literature that make these methods also suitable for the online retailing context. This ensures that estimation times remain feasible even if the size of the retailer increases and opens the way for advanced model-based marketing analytics in high-dimensional assortments.

Samenvatting

De afgelopen twee decennia is online winkelen steeds populairder geworden en vandaag de dag maken grote online winkels het mogelijk voor hun klanten om vrijwel elk product te kopen. Het gevolg hiervan is dat de productassortimenten van online winkels vele malen groter zijn dan die van traditionele winkels. In dit proefschrift worden modelmatige methoden gepresenteerd die gebruikt kunnen worden om aankopen uit zulke hoog dimensionale assortimenten te modelleren. Deze methodes zijn in staat om op het niveau van een individuele klant nauwkeurig te voorspellen welk product uit een groot assortiment gekocht gaat worden. Daarnaast leveren de methoden een dieper inzicht in de onderliggende patronen van het geobserveerde aankoopgedrag. De toepasbaarheid van zulke methoden in de praktijk hangt sterk samen met de schaalbaarheid van de methode en dit geldt met name voor online winkels. Resultaten uit het model moeten snel verkregen kunnen worden en de rekentijd mag niet significant toenemen als de hoeveelheid klanten of het aantal producten stijgt. Schaalbaarheid is daarom een van de aandachtspunten in dit proefschrift. De geïntroduceerde methoden zijn aanpassingen en uitbreidingen van snelle en schaalbare methoden uit de machinaal leren literatuur, zodat deze geschikt worden om te gebruiken in de context van online winkelen. Dit zorgt ervoor dat de rekentijd beperkt blijft zelfs als de winkel groeit, waarmee de weg wordt vrij gemaakt voor geavanceerde modelmatige marketing analyses in hoog dimensionale assortimenten.

About the Author

Bruno Jacobs (1988) holds a Master's degree in Econometrics and Management Science from Erasmus University Rotterdam. In 2012 he joined the Erasmus Research Institute of Management (ERIM) as a PhD student under the supervision of prof. dr. Bas Donkers and prof. dr. Dennis Fok. He carried out his research within the Marketing department and the Econometric Institute at the Erasmus School of Economics. In 2015, Bruno was a visiting scholar at Columbia University in the City of New York, hosted by prof. dr. Asim Ansari. Bruno currently works as an Assistant Professor in Marketing at the Robert H. Smith School of Business, University of Maryland, College Park.

Bruno's research is situated at the intersection of marketing, machine learning, and econometrics. Currently, he focuses on the development of new methods to describe, understand, and predict customer decisions in high-dimensional product assortments. Scalability is one of the focal points in his research. Bruno has worked with national and international retailers for his research. His work has been published in Marketing Science and presented at the Wharton School, multiple INFORMS Marketing Science Conferences, and a Marketing Dynamics Conference.

Portfolio

Education

- M.Sc. Econometrics and Management Science: Specialization in Quantitative Marketing, *Erasmus University Rotterdam*, 2012.
- B.Sc. Econometrics and Operations Research, *Erasmus University Rotterdam*, 2011.

Research Visits

- Columbia University in the City of New York, hosted by prof. dr. Asim Ansari, 2015.

Publications

- Bruno J.D. Jacobs, Bas Donkers, Dennis Fok (2016), “Model-Based Purchase Predictions for Large Assortments,” *Marketing Science*, 35 (3), 389–404.

Selected Conference Presentations

- Bruno Jacobs, Bas Donkers, Dennis Fok. “Model-Based Project Discovery,” *Marketing Science Conference*, Baltimore, 2015.
- Bruno Jacobs, Bas Donkers, Dennis Fok. “Model-Based Project Discovery,” *WCAI Symposium at the Wharton School*, Philadelphia, 2015.
- Bruno Jacobs, Bas Donkers, Dennis Fok. “Product Recommendations Based on Latent Purchase Motivations,” *Marketing Science Conference*, Atlanta, 2014.

Teaching and Advising Experience

At Erasmus University Rotterdam:

- Teaching Assistant for "Econometrics 2," 2013–2016
- Teaching Assistant for "Advanced Programming in MATLAB," 2011–2012
- Guest Lecturer for "New Research Methods in Marketing," 2011–2012
- Advisor for "M.Sc. Thesis in the Marketing Master Program," 2012–2016
- Advisor for "Seminar in Business Analytics and Quantitative Marketing," 2015–2016

Honors, Awards, and Grants

Honors

- AMA-Sheth Foundation Doctoral Consortium 2015, Fellow
- Marketing Science Doctoral Consortium 2015, Fellow
- Marketing Science Doctoral Consortium 2013, Fellow

Awards

- Accepted for the WCAI research project: Using Purchase History to Identify Customer "Projects"

Grants

- Erasmus Trustfonds Grant
- ERIM Research Visit Grant
- ERIM Travel Grants

ERIM Publications List

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

Dissertations in the last five years

Abbink, E.J., *Crew Management in Passenger Rail Transport*, Promotors: Prof. L.G. Kroon & Prof. A.P.M. Wagelmans, EPS-2014-325-LIS, http://repub.eur.nl/pub/76927.

Acar, O.A., *Crowdsourcing for Innovation: Unpacking Motivational, Knowledge and Relational Mechanisms of Innovative Behavior in Crowdsourcing Platforms*, Promotor: Prof. J.C.M. van den Ende, EPS-2014-321-LIS, http://repub.eur.nl/pub/76076.

Akemu, O., *Corporate Responses to Social Issues: Essays in Social Entrepreneurship and Corporate Social Responsibility*, Promotors: Prof. G.M. Whiteman & Dr. S.P. Kennedy, EPS-2017-392-ORG, https://repub.eur.nl/pub/95768.

Akin Ates, M., *Purchasing and Supply Management at the Purchase Category Level: Strategy, structure and performance*, Promotors: Prof. J.Y.F. Wynstra & Dr. E.M. van Raaij, EPS-2014-300-LIS, http://repub.eur.nl/pub/50283.

Akpinar, E., *Consumer Information Sharing*, Promotor: Prof. A. Smidts, EPS-2013-297-MKT, http://repub.eur.nl/pub/50140.

Alexander, L., *People, Politics, and Innovation: A Process Perspective*, Promotors: Prof. H.G. Barkema & Prof. D.L. van Knippenberg, EPS-2014-331-S&E, http://repub.eur.nl/pub/77209.

Alexiou, A., *Management of Emerging Technologies and the Learning Organization: Lessons from the Cloud and Serious Games Technology*, Promotors: Prof. S.J. Magala, Prof. M.C. Schippers and Dr. I. Oshri, EPS-2016-404-ORG, http://repub.eur.nl/pub/93818.

Almeida e Santos Nogueira, R.J. de, *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*, Promotors: Prof. U. Kaymak & Prof. J.M.C. Sousa, EPS-2014-310-LIS, http://repub.eur.nl/pub/51560.

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-frequency Data*, Promotor: Prof. D.J.C. van Dijk, EPS-2013-273-F&A, http://repub.eur.nl/pub/38240.

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promotors: Prof. H.W. Volberda & Prof. F.A.J. van den Bosch, EPS-2013-278-S&E, http://repub.eur.nl/pub/39128.

Benschop, N., *Biases in Project Escalation: Names, frames & construal levels*, Promotors: Prof. K.I.M. Rhode, Prof. H.R. Commandeur, Prof. M. Keil & Dr. A.L.P. Nuijten, EPS-2015-375-S&E, http://repub.eur.nl/pub/79408.

Berg, W.E. van den, *Understanding Salesforce Behavior using Genetic Association Studies*, Promotor: Prof. W.J.M.I. Verbeke, EPS-2014-311-MKT, http://repub.eur.nl/pub/51440.

Beusichem, H.C. van, *Firms and Financial Markets: Empirical Studies on the Informational Value of Dividends, Governance and Financial Reporting*, Promotors: Prof. A. de Jong & Dr. G. Westerhuis, EPS-2016-378-F&A, http://repub.eur.nl/pub/93079.

Bliek, R. de, *Empirical Studies on the Economic Impact of Trust*, Promotor: Prof. J. Veenman & Prof. Ph.H.B.F. Franses, EPS-2015-324-ORG, http://repub.eur.nl/pub/78159.

Boons, M., *Working Together Alone in the Online Crowd: The Effects of Social Motivations and Individual Knowledge Backgrounds on the Participation and Performance of Members of Online Crowdsourcing Platforms*, Promotors: Prof. H.G. Barkema & Dr. D.A. Stam, EPS-2014-306-S&E, http://repub.eur.nl/pub/50711.

Bouman, P., *Passengers, Crowding and Complexity: Models for Passenger Oriented Public Transport*, Prof. L.G. Kroon, Prof. A. Schöbel & Prof. P.H.M. Vervest, EPS-2017-420-LIS, https://repub.eur.nl/pub/100767.

Brazys, J., *Aggregated Marcoeconomic News and Price Discovery*, Promotor: Prof. W.F.C. Verschoor, EPS-2015-351-F&A, http://repub.eur.nl/pub/78243.

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit, Coworker Satisfaction, and Relational Models*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-292-ORG, http://repub.eur.nl/pub/41508.

Cancurtaran, P., *Essays on Accelerated Product Development*, Promotors: Prof. F. Langerak & Prof. G.H. van Bruggen, EPS-2014-317-MKT, http://repub.eur.nl/pub/76074.

Caron, E.A.M., *Explanation of Exceptional Values in Multi-dimensional Business Databases*, Promotors: Prof. H.A.M. Daniels & Prof. G.W.J. Hendrikse, EPS-2013-296-LIS, http://repub.eur.nl/pub/50005.

Carvalho, L. de, *Knowledge Locations in Cities: Emergence and Development Dynamics*, Promotor: Prof. L. Berg, EPS-2013-274-S&E, http://repub.eur.nl/pub/38449.

Chammas, G., *Portfolio Concentration*, Promotor: Prof. J. Spronk, EPS-2017-410-F&E, https://repub.eur.nl/pub/94975.

Cranenburgh, K.C. van, *Money or Ethics: Multinational corporations and religious organisations operating in an era of corporate responsibility*, Promotors: Prof. L.C.P.M. Meijs, Prof. R.J.M. van Tulder & Dr. D. Arenas, EPS-2016-385-ORG, http://repub.eur.nl/pub/93104.

Consiglio, I., *Others: Essays on Interpersonal and Consumer Behavior*, Promotor: Prof. S.M.J. van Osselaer, EPS-2016-366-MKT, http://repub.eur.nl/pub/79820.

Cox, R.H.G.M., *To Own, To Finance, and To Insure – Residential Real Estate Revealed*, Promotor: Prof. D. Brounen, EPS-2013-290-F&A, http://repub.eur.nl/pub/40964.

Darnihamedani, P., *Individual Characteristics, Contextual Factors and Entrepreneurial Behavior*, Promotors: Prof. A.R. Thurik & S.J.A. Hessels, EPS-2016-360-S&E, http://repub.eur.nl/pub/93280.

Dennerlein, T., *Empowering Leadership and Employees' Achievement Motivations: the Role of Self-Efficacy and Goal Orientations in the Empowering Leadership Process*, Promotors: Prof. D.L. van Knippenberg & Dr. J. Dietz, EPS-2017-414-ORG, https://repub.eur.nl/pub/98438.

Deng, W., *Social Capital and Diversification of Cooperatives*, Promotor: Prof. G.W.J. Hendrikse, EPS-2015-341-ORG, http://repub.eur.nl/pub/77449.

Depecik, B.E., *Revitalizing brands and brand: Essays on Brand and Brand Portfolio Management Strategies*, Promotors: Prof. G.H. van Bruggen, Dr. Y.M. van Everdingen and Dr. M.B. Ataman, EPS-2016-406-MKT, http://repub.eur.nl/pub/93507.

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promotor: Prof. A.P.M. Wagelmans, EPS-2013-272-LIS, http://repub.eur.nl/pub/38241.

Duyvesteyn, J.G., *Empirical Studies on Sovereign Fixed Income Markets*, Promotors: Prof. P. Verwijmeren & Prof. M.P.E. Martens, EPS-2015-361-F&A, https://repub.eur.nl/pub/79033.

Duursema, H., *Strategic Leadership: Moving Beyond the Leader-Follower Dyad*, Promotor: Prof. R.J.M. van Tulder, EPS-2013-279-ORG, http://repub.eur.nl/pub/39129.

Elemes, A, *Studies on Determinants and Consequences of Financial Reporting Quality*, Promotor: Prof. E. Peek, EPS-2015-354-F&A, https://repub.eur.nl/pub/79037.

Ellen, S. ter, *Measurement, Dynamics, and Implications of Heterogeneous Beliefs in Financial Markets*, Promotor: Prof. W.F.C. Verschoor, EPS-2015-343-F&A, http://repub.eur.nl/pub/78191.

Erlemann, C., *Gender and Leadership Aspiration: The Impact of the Organizational Environment*, Promotor: Prof. D.L. van Knippenberg, EPS-2016-376-ORG, http://repub.eur.nl/pub/79409.

Eskenazi, P.I., *The Accountable Animal*, Promotor: Prof. F.G.H. Hartmann, EPS-2015-355-F&A, http://repub.eur.nl/pub/78300.

Evangelidis, I., *Preference Construction under Prominence*, Promotor: Prof. S.M.J. van Osselaer, EPS-2015-340-MKT, http://repub.eur.nl/pub/78202.

Faber, N., *Structuring Warehouse Management*, Promotors: Prof. M.B.M. de Koster & Prof. A. Smidts, EPS-2015-336-LIS, http://repub.eur.nl/pub/78603.

Feng, Y., *The Effectiveness of Corporate Governance Mechanisms and Leadership Structure: Impacts on strategic change and firm performance*, Promotors: Prof. F.A.J. van den Bosch, Prof. H.W. Volberda & Dr. J.S. Sidhu, EPS-2017-389-S&E, https://repub.eur.nl/pub/98470.

Fernald, K., *The Waves of Biotechnological Innovation in Medicine: Interfirm Cooperation Effects and a Venture Capital Perspective*, Promotors: Prof. E. Claassen, Prof. H.P.G. Pennings & Prof. H.R. Commandeur, EPS-2015-371-S&E, http://hdl.handle.net/1765/79120.

Fisch, C.O., *Patents and trademarks: Motivations, antecedents, and value in industrialized and emerging markets*, Promotors: Prof. J.H. Block, Prof. H.P.G. Pennings & Prof. A.R. Thurik, EPS-2016-397-S&E, http://repub.eur.nl/pub/94036.

Fliers, P.T., *Essays on Financing and Performance: The role of firms, banks and board*, Promotor: Prof. A. de Jong & Prof. P.G.J. Roosenboom, EPS-2016-388-F&A, http://repub.eur.nl/pub/93019.

Fourne, S.P., *Managing Organizational Tensions: A Multi-Level Perspective on Exploration, Exploitation and Ambidexterity*, Promotors: Prof. J.J.P. Jansen & Prof. S.J. Magala, EPS-2014-318-S&E, http://repub.eur.nl/pub/76075.

Gaast, J.P. van der, *Stochastic Models for Order Picking Systems*, Promotors: Prof. M.B.M de Koster & Prof. I.J.B.F. Adan, EPS-2016-398-LIS, http://repub.eur.nl/pub/93222.

Giurge, L.M., *A Test of Time; A temporal and dynamic approach to power and ethics*, Promotors: Prof. M.H. van Dijke & Prof. D. De Cremer, EPS-2017-412-ORG, https://repub.eur.nl/pub/98451.

Glorie, K.M., *Clearing Barter Exchange Markets: Kidney Exchange and Beyond*, Promotors: Prof. A.P.M. Wagelmans & Prof. J.J. van de Klundert, EPS-2014-329-LIS, http://repub.eur.nl/pub/77183.

Hekimoglu, M., *Spare Parts Management of Aging Capital Products*, Promotor: Prof. R. Dekker, EPS-2015-368-LIS, http://repub.eur.nl/pub/79092.

Heyde Fernandes, D. von der, *The Functions and Dysfunctions of Reminders*, Promotor: Prof. S.M.J. van Osselaer, EPS-2013-295-MKT, http://repub.eur.nl/pub/41514.

Hogenboom, A.C., *Sentiment Analysis of Text Guided by Semantics and Structure*, Promotors: Prof. U. Kaymak & Prof. F.M.G. de Jong, EPS-2015-369-LIS, http://repub.eur.nl/pub/79034.

Hogenboom, F.P., *Automated Detection of Financial Events in News Text*, Promotors: Prof. U. Kaymak & Prof. F.M.G. de Jong, EPS-2014-326-LIS, http://repub.eur.nl/pub/77237.

Hollen, R.M.A., *Exploratory Studies into Strategies to Enhance Innovation-Driven International Competitiveness in a Port Context: Toward Ambidextrous Ports*, Promotors: Prof. F.A.J. Van Den Bosch & Prof. H.W.Volberda, EPS-2015-372-S&E, http://repub.eur.nl/pub/78881.

Hout, D.H. van, *Measuring Meaningful Differences: Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling*, Promotors: Prof. P.J.F. Groenen & Prof. G.B. Dijksterhuis, EPS-2014-304-MKT, http://repub.eur.nl/pub/50387.

Houwelingen, G.G. van, *Something To Rely On*, Promotors: Prof. D. de Cremer & Prof. M.H. van Dijke, EPS-2014-335-ORG, http://repub.eur.nl/pub/77320.

Hurk, E. van der, *Passengers, Information, and Disruptions*, Promotors: Prof. L.G. Kroon & Prof. P.H.M. Vervest, EPS-2015-345-LIS, http://repub.eur.nl/pub/78275.

Iseger, P. den, *Fourier and Laplace Transform Inversion with Applications in Finance*, Promotor: Prof. R. Dekker, EPS-2014-322-LIS, http://repub.eur.nl/pub/76954.

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promotor: Prof. R. Dekker, EPS-2013-288-LIS, http://repub.eur.nl/pub/39933.

Khanagha, S., *Dynamic Capabilities for Managing Emerging Technologies*, Promotor: Prof. H.W. Volberda, EPS-2014-339-S&E, http://repub.eur.nl/pub/77319.

Khattab, J., *Make Minorities Great Again: a contribution to workplace equity by identifying and addressing constraints and privileges*, Prof. D.L. van Knippenberg & Dr. A. Nederveen Pieterse, EPS-2017-421-ORG, https://repub.eur.nl/pub/99311.

Kil, J., *Acquisitions Through a Behavioral and Real Options Lens*, Promotor: Prof. H.T.J. Smit, EPS-2013-298-F&A, http://repub.eur.nl/pub/50142.

Klooster, E. van't, *Travel to Learn: the Influence of Cultural Distance on Competence Development in Educational Travel*, Promotors: Prof. F.M. Go & Prof. P.J. van Baalen, EPS-2014-312-MKT, http://repub.eur.nl/pub/51462.

Koendjbiharie, S.R., *The Information-Based View on Business Network Performance: Revealing the Performance of Interorganizational Networks*, Promotors: Prof. H.W.G.M. van Heck & Prof. P.H.M. Vervest, EPS-2014-315-LIS, http://repub.eur.nl/pub/51751.

Koning, M., *The Financial Reporting Environment: The Role of the Media, Regulators and Auditors*, Promotors: Prof. G.M.H. Mertens & Prof. P.G.J. Roosenboom, EPS-2014-330-F&A, http://repub.eur.nl/pub/77154.

Konter, D.J., *Crossing Borders with HRM: An Inquiry of the Influence of Contextual Differences in the Adoption and Effectiveness of HRM*, Promotors: Prof. J. Paauwe & Dr. L.H. Hoeksema, EPS-2014-305-ORG, http://repub.eur.nl/pub/50388.

Korkmaz, E., *Bridging Models and Business: Understanding Heterogeneity in Hidden Drivers of Customer Purchase Behavior*, Promotors: Prof. S.L. van de Velde & Prof. D. Fok, EPS-2014-316-LIS, http://repub.eur.nl/pub/76008.

Krämer, R., *A license to mine? Community organizing against multinational corporations*, Promotors: Prof. R.J.M. van Tulder & Prof. G.M. Whiteman, EPS-2016-383-ORG, http://repub.eur.nl/pub/94072.

Kroezen, J.J., *The Renewal of Mature Industries: An Examination of the Revival of the Dutch Beer Brewing Industry*, Promotor: Prof. P.P.M.A.R. Heugens, EPS-2014-333-S&E, http://repub.eur.nl/pub/77042.

Kysucky, V., *Access to Finance in a Cros-Country Context*, Promotor: Prof. L. Norden, EPS-2015-350-F&A, http://repub.eur.nl/pub/78225.

Lee, C.I.S.G, *Big Data in Management Research: Exploring New Avenues*, Promotors: Prof. S.J. Magala & Dr. W.A. Felps, EPS-2016-365-ORG, http://repub.eur.nl/pub/79818.

Legault-Tremblay, P.O., *Corporate Governance During Market Transition: Heterogeneous responses to Institution Tensions in China*, Promotor: Prof. B. Krug, EPS-2015-362-ORG, http://repub.eur.nl/pub/78649.

Lenoir, A.S., *Are You Talking to Me? Addressing Consumers in a Globalised World*, Promotors: Prof. S. Puntoni & Prof. S.M.J. van Osselaer, EPS-2015-363-MKT, http://repub.eur.nl/pub/79036.

Leunissen, J.M., *All Apologies: On the Willingness of Perpetrators to Apologize*, Promotors: Prof. D. de Cremer & Dr. M. van Dijke, EPS-2014-301-ORG, http://repub.eur.nl/pub/50318.

Li, D., *Supply Chain Contracting for After-sales Service and Product Support*, Promotor: Prof. M.B.M. de Koster, EPS-2015-347-LIS, http://repub.eur.nl/pub/78526.

Li, Z., *Irrationality: What, Why and How*, Promotors: Prof. H. Bleichrodt, Prof. P.P. Wakker, & Prof. K.I.M. Rohde, EPS-2014-338-MKT, http://repub.eur.nl/pub/77205.

Liu, N., *Behavioral Biases in Interpersonal Contexts*, Supervisors: Prof. A. Baillon & Prof. H. Bleichrodt, EPS-2017-408-MKT, https://repub.eur.nl/pub/95487.

Liang, Q.X., *Governance, CEO Identity, and Quality Provision of Farmer Cooperatives*, Promotor: Prof. G.W.J. Hendrikse, EPS-2013-281-ORG, http://repub.eur.nl/pub/39253.

Liket, K., *Why 'Doing Good' is not Good Enough: Essays on Social Impact Measurement*, Promotors: Prof. H.R. Commandeur & Dr. K.E.H. Maas, EPS-2014-307-STR, http://repub.eur.nl/pub/51130.

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research*, Promotors: Prof. A.R. Thurik, Prof. P.J.F. Groenen, & Prof. A. Hofman, EPS-2013-287-S&E, http://repub.eur.nl/pub/40081.

Lu, Y., *Data-Driven Decision Making in Auction Markets*, Promotors: Prof. H.W.G.M. van Heck & Prof. W. Ketter, EPS-2014-314-LIS, http://repub.eur.nl/pub/51543.

Ma, Y., *The Use of Advanced Transportation Monitoring Data for Official Statistics*, Promotors: Prof. L.G. Kroon and Dr. J. van Dalen, EPS-2016-391-LIS, http://repub.eur.nl/pub/80174.

Manders, B., *Implementation and Impact of ISO 9001*, Promotor: Prof. K. Blind, EPS-2014-337-LIS, http://repub.eur.nl/pub/77412.

Mell, J.N., *Connecting Minds: On The Role of Metaknowledge in Knowledge Coordination*, Promotor: Prof. D.L. van Knippenberg, EPS-2015-359-ORG, http://hdl.handle.net/1765/78951.

Meulen,van der, D., *The Distance Dilemma: the effect of flexible working practices on performance in the digital workplace*, Promotors: Prof. H.W.G.M. van Heck & Prof. P.J. van Baalen, EPS-2016-403-LIS, http://repub.eur.nl/pub/94033.

Micheli, M.R., *Business Model Innovation: A Journey across Managers' Attention and Inter-Organizational Networks*, Promotor: Prof. J.J.P. Jansen, EPS-2015-344-S&E, http://repub.eur.nl/pub/78241.

Milea, V., *News Analytics for Financial Decision Support*, Promotor: Prof. U. Kaymak, EPS-2013-275-LIS, http://repub.eur.nl/pub/38673.

Moniz, A., *Textual Analysis of Intangible Information*, Promotors: Prof. C.B.M. van Riel, Prof. F.M.G de Jong & Dr. G.A.J.M. Berens, EPS-2016-393-ORG, http://repub.eur.nl/pub/93001.

Mulder, J., *Network design and robust scheduling in liner shipping*, Promotors: Prof. R. Dekker & Dr. W.L. van Jaarsveld, EPS-2016-384-LIS, http://repub.eur.nl/pub/80258.

Naumovska, I., *Socially Situated Financial Markets: A Neo-Behavioral Perspective on Firms, Investors and Practices*, Promotors: Prof. P.P.M.A.R. Heugens & Prof. A. de Jong, EPS-2014-319-S&E, http://repub.eur.nl/pub/76084.

Neerijnen, P., *The Adaptive Organization: the socio-cognitive antecedents of ambidexterity and individual exploration*, Promotors: Prof. J.J.P. Jansen, P.P.M.A.R. Heugens & Dr. T.J.M. Mom, EPS-2016-358-S&E, http://repub.eur.nl/pub/93274.

Okbay, A., *Essays on Genetics and the Social Sciences*, Promotors: Prof. A.R. Thurik, Prof. Ph.D. Koellinger & Prof. P.J.F. Groenen, EPS-2017-413-S&E, https://repub.eur.nl/pub/95489.

Oord, J.A. van, *Essays on Momentum Strategies in Finance*, Promotor: Prof. H.K. van Dijk, EPS-2016-380-F&A, http://repub.eur.nl/pub/80036.

Peng, X., *Innovation, Member Sorting, and Evaluation of Agricultural Cooperatives*, Promotor: Prof. G.W.J. Hendriks, EPS-2017-409-ORG, https://repub.eur.nl/pub/94976.

Pennings, C.L.P., *Advancements in Demand Forecasting: Methods and Behavior*, Promotors: Prof. L.G. Kroon, Prof. H.W.G.M. van Heck & Dr. J. van Dalen, EPS-2016-400-LIS, http://repub.eur.nl/pub/94039.

Peters, M., *Machine Learning Algorithms for Smart Electricity Markets*, Promotor: Prof. W. Ketter, EPS-2014-332-LIS, http://repub.eur.nl/pub/77413.

Pocock, M., *Status Inequalities in Business Exchange Relations in Luxury Markets*, Promotors: Prof. C.B.M. van Riel & Dr. G.A.J.M. Berens, EPS-2017-346-ORG, https://repub.eur.nl/pub/98647.

Porck, J., *No Team is an Island: An Integrative View of Strategic Consensus between Groups*, Promotors: Prof. P.J.F. Groenen & Prof. D.L. van Knippenberg, EPS-2013-299-ORG, http://repub.eur.nl/pub/50141.

Pozharliev, R., *Social Neuromarketing: The role of social context in measuring advertising effectiveness*, Promotors: Prof. W.J.M.I. Verbeke & Prof. J.W. van Strien, EPS-2017-402-MKT, https://repub.eur.nl/pub/95528.

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promotors: Prof. H.J.H.M. Claassen & Prof. H.R. Commandeur, EPS-2013-282-S&E, http://repub.eur.nl/pub/39654.

Protzner, S., *Mind the gap between demand and supply: A behavioral perspective on demand forecasting*, Promotors: Prof. S.L. van de Velde & Dr. L. Rook, EPS-2015-364-LIS, http://repub.eur.nl/pub/79355.

Pruijssers, J.K., *An Organizational Perspective on Auditor Conduct*, Promotors: Prof. J. van Oosterhout & Prof. P.P.M.A.R. Heugens, EPS-2015-342-S&E, http://repub.eur.nl/pub/78192.

Retel Helmrich, M.J., *Green Lot-Sizing*, Promotor: Prof. A.P.M. Wagelmans, EPS-2013-291-LIS, http://repub.eur.nl/pub/41330.

Rietdijk, W.J.R., *The Use of Cognitive Factors for Explaining Entrepreneurship*, Promotors: Prof. A.R. Thurik & Prof. I.H.A. Franken, EPS-2015-356-S&E, http://repub.eur.nl/pub/79817.

Rietveld, N., *Essays on the Intersection of Economics and Biology*, Promotors: Prof. A.R. Thurik, Prof. Ph.D. Koellinger, Prof. P.J.F. Groenen & Prof. A. Hofman, EPS-2014-320-S&E, http://repub.eur.nl/pub/76907.

Rösch, D., *Market Efficiency and Liquidity*, Promotor: Prof. M.A. van Dijk, EPS-2015-353-F&A, http://repub.eur.nl/pub/79121.

Roza, L., *Employee Engagement in Corporate Social Responsibility: A collection of essays*, Promotor: L.C.P.M. Meijs, EPS-2016-396-ORG, http://repub.eur.nl/pub/93254.

Rubbaniy, G., *Investment Behaviour of Institutional Investors*, Promotor: Prof. W.F.C. Verschoor, EPS-2013-284-F&A, http://repub.eur.nl/pub/40068.

Schoonees, P., *Methods for Modelling Response Styles*, Promotor: Prof. P.J.F. Groenen, EPS-2015-348-MKT, http://repub.eur.nl/pub/79327.

Schouten, M.E., *The Ups and Downs of Hierarchy: the causes and consequences of hierarchy struggles and positional loss*, Promotors; Prof. D.L. van Knippenberg & Dr. L.L. Greer, EPS-2016-386-ORG, http://repub.eur.nl/pub/80059.

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promotor: Prof. G.M.H. Mertens, EPS-2013-283-F&A, http://repub.eur.nl/pub/39655.

Smit, J., *Unlocking Business Model Innovation: A look through the keyhole at the inner workings of Business Model Innovation*, Promotor: H.G. Barkema, EPS-2016-399-S&E, http://repub.eur.nl/pub/93211.

Sousa, M.J.C. de, *Servant Leadership to the Test: New Perspectives and Insights*, Promotors: Prof. D.L. van Knippenberg & Dr. D. van Dierendonck, EPS-2014-313-ORG, http://repub.eur.nl/pub/51537.

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promotor: Prof. R. Dekker, EPS-2013-293-LIS, http://repub.eur.nl/pub/41513.

Staadt, J.L., *Leading Public Housing Organisation in a Problematic Situation: A Critical Soft Systems Methodology Approach*, Promotor: Prof. S.J. Magala, EPS-2014-308-ORG, http://repub.eur.nl/pub/50712.

Stallen, M., *Social Context Effects on Decision-Making: A Neurobiological Approach*, Promotor: Prof. A. Smidts, EPS-2013-285-MKT, http://repub.eur.nl/pub/39931.

Szatmari, B., *We are (all) the champions: The effect of status in the implementation of innovations*, Promotors: Prof J.C.M & Dr. D. Deichmann, EPS-2016-401-LIS, http://repub.eur.nl/pub/94633.

Tarakci, M., *Behavioral Strategy: Strategic Consensus, Power and Networks*, Promotors: Prof. D.L. van Knippenberg & Prof. P.J.F. Groenen, EPS-2013-280-ORG, http://repub.eur.nl/pub/39130.

Tuijl, E. van, *Upgrading across Organisational and Geographical Configurations*, Promotor: Prof. L. van den Berg, EPS-2015-349-S&E, http://repub.eur.nl/pub/78224.

Tuncdogan, A., *Decision Making and Behavioral Strategy: The Role of Regulatory Focus in Corporate Innovation Processes*, Promotors: Prof. F.A.J. van den Bosch, Prof. H.W. Volberda, & Prof. T.J.M. Mom, EPS-2014-334-S&E, http://repub.eur.nl/pub/76978.

Uijl, S. den, *The Emergence of De-facto Standards*, Promotor: Prof. K. Blind, EPS-2014-328-LIS, http://repub.eur.nl/pub/77382.

Vagias, D., *Liquidity, Investors and International Capital Markets*, Promotor: Prof. M.A. van Dijk, EPS-2013-294-F&A, http://repub.eur.nl/pub/41511.

Valogianni, K., *Sustainable Electric Vehicle Management using Coordinated Machine Learning*, Promotors: Prof. H.W.G.M. van Heck & Prof. W. Ketter, EPS-2016-387-LIS, http://repub.eur.nl/pub/93018.

Vandic, D., *Intelligent Information Systems for Web Product Search*, Promotors: Prof. U. Kaymak & Dr. Frasincar, EPS-2017-405-LIS, https://repub.eur.nl/pub/95490.

Veelenturf, L.P., *Disruption Management in Passenger Railways: Models for Timetable, Rolling Stock and Crew Rescheduling*, Promotor: Prof. L.G. Kroon, EPS-2014-327-LIS, http://repub.eur.nl/pub/77155.

Venus, M., *Demystifying Visionary Leadership: In search of the essence of effective vision communication*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-289-ORG, http://repub.eur.nl/pub/40079.

Vermeer, W., *Propagation in Networks:The impact of information processing at the actor level on system-wide propagation dynamics*, Promotor: Prof. P.H.M.Vervest, EPS-2015-373-LIS, http://repub.eur.nl/pub/79325.

Versluis, I., *Prevention of the Portion Size Effect*, Promotors: Prof. Ph.H.B.F. Franses & Dr. E.K. Papies, EPS-2016-382-MKT, http://repub.eur.nl/pub/79880.

Vishwanathan, P., *Governing for Stakeholders: How Organizations May Create or Destroy Value for their Stakeholders*, Promotors: Prof. J. van Oosterhout & Prof. L.C.P.M. Meijs, EPS-2016-377-ORG, http://repub.eur.nl/pub/93016.

Visser, V.A., *Leader Affect and Leadership Effectiveness: How leader affective displays influence follower outcomes*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-286-ORG, http://repub.eur.nl/pub/40076.

Vlaming, R. de., *Linear Mixed Models in Statistical Genetics*, Prof. A.R. Thurik, Prof. P.J.F. Groenen & Prof. Ph.D. Koellinger, EPS-2017-416-S&E, https://repub.eur.nl/pub/100428.

Vries, J. de, *Behavioral Operations in Logistics*, Promotors: Prof. M.B.M de Koster & Prof. D.A. Stam, EPS-2015-374-LIS, http://repub.eur.nl/pub/79705.

Wagenaar, J.C., *Practice Oriented Algorithmic Disruption Management in Passenger Railways*, Promotors: Prof. L.G. Kroon & Prof. A.P.M. Wagelmans, EPS-2016-390-LIS, http://repub.eur.nl/pub/93177.

Wang, P., *Innovations, status, and networks*, Promotors: Prof. J.J.P. Jansen & Dr. V.J.A. van de Vrande, EPS-2016-381-S&E, http://repub.eur.nl/pub/93176.

Wang, R., *Corporate Environmentalism in China*, Promotors: Prof. P.P.M.A.R Heugens & Dr. F.Wijen, EPS-2017-417-S&E, https://repub.eur.nl/pub/99987.

Wang, T., *Essays in Banking and Corporate Finance*, Promotors: Prof. L. Norden & Prof. P.G.J. Roosenboom, EPS-2015-352-F&A, http://repub.eur.nl/pub/78301.

Wang, Y., *Corporate Reputation Management: Reaching Out to Financial Stakeholders*, Promotor: Prof. C.B.M. van Riel, EPS-2013-271-ORG, http://repub.eur.nl/pub/38675.

Weenen, T.C., *On the Origin and Development of the Medical Nutrition Industry*, Promotors: Prof. H.R. Commandeur & Prof. H.J.H.M. Claassen, EPS-2014-309-S&E, http://repub.eur.nl/pub/51134.

Wessels, C., *Flexible Working Practices: How Employees Can Reap the Benefits for Engagement and Performance*, Promotors: Prof. H.W.G.M. van Heck, Prof. P.J. van Baalen & Prof. M.C. Schippers, EPS-2017-418-LIS, https://repub.eur.nl/.

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value*, Promotor: Prof. A. de Jong, EPS-2013-277-F&A, http://repub.eur.nl/pub/39127.

Yang, S., *Information Aggregation Efficiency of Prediction Markets*, Promotor: Prof. H.W.G.M. van Heck, EPS-2014-323-LIS, http://repub.eur.nl/pub/77184.

Ypsilantis, P., *The Design, Planning and Execution of Sustainable Intermodal Port-hinterland Transport Networks*, Promotors: Prof. R.A. Zuidwijk & Prof. L.G. Kroon, EPS-2016-395-LIS, http://repub.eur.nl/pub/94375.

Yuferova, D., *Price Discovery, Liquidity, Provision, and Low-Latency Trading*, Promotors: Prof. M.A. van Dijk & Dr. D.G.J. Bongaerts, EPS-2016-379-F&A, http://repub.eur.nl/pub/93017.

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promotor: Prof. M.B.M. de Koster, EPS-2013-276-LIS, http://repub.eur.nl/pub/38766.

Zuber, F.B., *Looking at the Others: Studies on (un)ethical behavior and social relationships in organizations*, Promotor: Prof. S.P. Kaptein, EPS-2016-394-ORG, http://repub.eur.nl/pub/94388.