

Understanding Large-Scale Dynamic Purchase Behavior

Bruno Jacobs*, Dennis Fok[†], Bas Donkers[‡]

August 1, 2020

In modern retail contexts, retailers sell products from vast product assortments to a large and heterogeneous customer base. Understanding purchase behavior in such a context is very important. Standard models cannot be used due to the high dimensionality of the data. We propose a new model that creates an efficient dimension reduction through the idea of purchase motivations. We only require customer-level purchase history data, which is ubiquitous in modern retailing. The model handles large-scale data and even works in settings with shopping trips consisting of few purchases. As scalability of the model is essential for practical applicability, we develop a fast, custom-made inference algorithm based on variational inference. Essential features of our model are that it accounts for the product, customer and time dimensions present in purchase history data; relates the relevance of motivations to customer- and shopping-trip characteristics; captures interdependencies between motivations; and achieves superior predictive performance. Estimation results from this comprehensive model provide deep insights into purchase behavior. Such insights can be used by managers to create more intuitive, better informed, and more effective marketing actions. We illustrate the model using purchase history data from a Fortune 500 retailer involving more than 4,000 unique products.

Keywords: dynamic purchase behavior; large-scale assortment; purchase history data; topic model; machine learning; variational inference

*Bruno Jacobs (brunojacobs@rhsmith.umd.edu) is an Assistant Professor of Marketing at the Robert H. Smith School of Business at the University of Maryland (corresponding author).

[†]Dennis Fok (dfok@ese.eur.nl) is Professor of Econometrics and Data Science at the Erasmus School of Economics, Erasmus University Rotterdam.

[‡]Bas Donkers (donkers@ese.eur.nl) is Professor of Marketing Research at the Erasmus School of Economics, Erasmus University Rotterdam.

1. Introduction

The value of purchase history data to improve marketing activities has since long been recognized by the field (Rossi et al. 1996). An explanation for the popularity of such data is that it is one of the few data sources about revealed customer preferences that is ubiquitous. It is available at virtually any retailer. Yet, the exponential growth in assortment size in many parts of the retail landscape – especially in online retailing – has made it difficult to extract valuable managerial insights from such data. Existing methods no longer suffice.

The primary challenge in analyzing a modern retailer’s purchase history data is accounting for the number and variety of products sold. On the one hand, the added value of using data to better understand customer needs and preferences increases with assortment size and diversity. Such understanding can improve marketing actions and support personalized communication. Examples are product recommendations, aiding navigation through and categorization of the assortment, and targeting activities such as personalized direct (e)mail campaigns. On the other hand, the large variety of products offered to – and purchased by – customers makes it increasingly difficult to understand and analyze such purchase behavior.

In this article we introduce a new model that enables marketers to gain in-depth insights from purchase history data in the context of large and varied assortments, while accounting for heterogeneity across customers and shopping trips. To ensure that the model can be applied to large, realistic, retailing settings we derive a custom-made scalable variational inference algorithm. We demonstrate the model using purchase history data from a Fortune 500 retailer that contains purchases from a large product assortment consisting of more than 4,000 products and close to 50,000 shopping trips.

Traditional applications of purchase history data have often involved fast-moving consumer goods (FMCG), for example, using scanner panel data in a supermarket, e.g. Guadagni and Little (1983), Gupta (1988), Manchanda et al. (1999). These three studies analyzed purchase behavior for only a small subset of the product assortment, considering respectively 8, 11, and 4 alternatives. Modern retail applications involve many more products and directly applying these methods to a large assortment is certainly not trivial, if not impossible. In retail settings outside of the FMCG context, two additional challenges surface. First, repeat purchases of the same product either occur very infrequently or will be non-existent for most customers.¹ For example, consider buying products at a home improvement store, ordering products from Amazon, or watching content on Netflix. Second, often only a few products are purchased per shopping trip, both online and offline.² The phenomenon of few products being purchased per shopping trip is further exacerbated by services like Amazon Prime, which guarantees free shipping with no minimum spending required.³ Taken together, these developments in modern retailing require new methods to harvest the value embedded in purchase history data.

In order to accurately describe purchase behavior, any quantitative method needs to account for at least three dimensions in the data (see Manchanda et al. (1999) for a similar argument). The first two dimensions are the products and the customers, with preferences varying across

¹ “Small baskets, large stores - how shopping behaviour is changing”, dunnhumby, March 20, 2017 (<https://www.dunnhumby.com/resources/reports/small-baskets-large-stores>)

² “Can Grocery Stores Embrace Change And Technology?”, Forbes, May 28, 2019 (<https://www.forbes.com/sites/lanabandoim/2019/05/28/can-grocery-stores-embrace-change-and-technology>)

³ “Amazon’s new weapon to crush competition: \$1 items delivered for free — by tomorrow”, Vox, October 14, 2019 (<https://www.vox.com/recode/2019/10/14/20906728/amazon-prime-low-price-products-add-on-one-day-delivery>)

customers. The third dimension relates to time, as a customer’s preferences and purchase behavior may vary across shopping trips. Such preference shifts could be intrinsic to the customer, e.g. due to evolving personal tastes, or driven by extrinsic contextual factors such as seasonality. These three dimensions (product, customer, and time) need to be accounted for simultaneously to properly capture the richness and complexity of purchase behavior.

Ideally, sufficient information is available to directly infer customer preferences at the product/customer/shopping trip-level. In practice, this is impossible as purchase data is very sparse across these dimensions due to several factors. First, product assortments are large and varied in modern retail environments. Second, a typical customer only purchases a very limited number of products from the complete assortment. Third, the scarce data that is available for a customer is spread out across shopping trips. The sparsity of the data, together with the size of the assortment and the need for interpretable outcomes, implies that the dimensionality of the problem needs to be reduced. This could easily be done by aggregating across products, customers or time. However, this eliminates the ability to learn anything about the removed dimension and requires ad-hoc aggregation rules, which may bias conclusions. For example, if the time dimension is ignored, seasonal products will be averaged out over time and as a result these products will be underexposed when they are in season.

In this paper we introduce a model that keeps all three dimensions at the original granularity, while specifying relations between products, customers and time in a lower dimensional space. This space consists of latent dimensions that each describe a salient pattern in the purchase data. Our identification of these dimensions is inspired by probabilistic topic models (Blei 2012), a modeling framework from the machine learning literature for text analysis. Jacobs et al. (2016) were the first to adapt this framework to purchase history data and labeled the resulting dimensions (topics) as latent *purchase motivations*. The idea is that motivations drive the observed purchase behavior in the customer base. For example, a motivation related to bathroom renovation would lead to purchases of products like PVC pipes, tiles, paint, and a bath tub, while a motivation related to gardening would lead to purchases of gardening supplies. The identification of such purchase motivations enables a marketing manager to reason about purchase behavior at a higher level, which can generate more insights than analyzing individual products separately, especially in the context of a large assortment.

In contrast to Jacobs et al. (2016), who aggregate over the time dimension, our model distinguishes between a customer’s shopping trips to provide a more nuanced and realistic representation of purchase behavior. This enables us to generate more detailed insights for retailers. For example, at the shopping-trip level our model can capture time-related effects like seasonality, where some products are more relevant during a certain time of the year. The obtained insights can be even more fine grained, such as concerning the day-of-the-week or time-of-day. Dynamic, autoregressive-like dependencies across shopping trips are modeled as well, where the products bought in one trip may be informative about the products that will be purchased next. In the end, the insights generated by such a comprehensive model allow for more informed and better targeted marketing actions, connecting a customer to the relevant products, at the right time. To deal with the increased complexity of the model, we replace the inference methodology of Jacobs et al. (2016) by a custom-made variational inference algorithm that achieves computationally and statistically efficient estimates.

Industry demand for such a comprehensive and scalable method is highlighted in a research opportunity set up by Wharton Customer Analytics.⁴ In this research opportunity a Fortune 500 Specialty Retailer calls for the development of tools to identify so-called “projects” from purchase history data. In this setting, a project is described as “...*gathering ingredients for a*

⁴ “Using Purchase History to Identify Customer “Projects””, Wharton Customer Analytics (<https://wca.wharton.upenn.edu/research/using-purchase-history-to-identify-customer-projects/>)

special meal or assembling the tools and supplies needed for a craft project, customers frequently purchase a collection of products that they need to complete a specific project". This aligns with the conceptualization of a purchase motivation. In addition, it is mentioned that *"...today's marketers have few tools to help identify collections of products that are associated with projects, or customers who seem to be engaged in such an activity"* and *"...customer segmentations are typically static and basket analysis seldom straddles multiple purchase occasions that might be associated with the same project"*. The method we propose in this paper is able to identify such projects from purchase history data, and to determine the dynamics in relevance of these projects across customers and shopping trips.

The remainder of the paper is structured as follows: Section 2 introduces the conceptual and technical details of our model, and positions it in the existing literature. We describe our scalable variational inference algorithm in Section 3. The data and results of our empirical application are described in Sections 4 and 5. Managerial implications are discussed in Section 6 and we wrap up with conclusions and avenues for further research in Section 7.

2. Modeling large-scale purchase behavior

In this section we introduce our model for large-scale customer purchase behavior that relies on the topic modeling framework. To show the versatility of this framework, we start with a review of topic modeling applications in marketing. We then describe purchase motivations, i.e. topics in a shopping context, at a conceptual level. After that, we introduce the formal statistical model and highlight the essential improvements compared to the LDA-X model introduced in Jacobs et al. (2016). Finally, we discuss alternative approaches to modeling dynamic purchase behavior. Throughout this section we illustrate parts of the model using examples for a hardware store which is the context of our empirical application, but naturally our model extends to other contexts as well.

2.1. Marketing applications of topic models

To develop our large-scale purchase behavior model we build on the machine learning literature, more specifically the research on probabilistic topic models in text analysis (Blei 2012). Several articles in the recent marketing literature apply and adapt methods based on topic models, most notably latent Dirichlet allocation (LDA) (Blei et al. 2003), to provide insight in marketing problems. Most of these papers involve analyzing textual data (Tirunillai and Tellis 2014, Büschken and Allenby 2016, Rutz et al. 2017, Puranam et al. 2017, Liu and Toubia 2018, Büschken and Allenby 2020), which is the traditional application of a topic model. Other papers use methods based on LDA that do not directly model text, but instead leverage the fact that LDA can model data that consists of sets of discrete data points.

Purchase history data was first modeled using a topic model by Jacobs et al. (2016), where purchase behavior of the customer base is described using a small set of purchase motivations (topics), and the relevance of each of these motivations is heterogeneous across customers. Our model extends this work in several ways and overcomes two of its major limitations. First, the time dimension is excluded in Jacobs et al. (2016), as the shopping trips of a customer are aggregated into a "single-basket" purchase history. This prohibits the inclusion of any time-specific effects, identification of dynamics present in purchase behavior, and shopping-trip specific idiosyncrasies. Second, the way in which customer-level heterogeneity in motivation relevance is modeled in Jacobs et al. (2016) is very restrictive, assuming all correlations between the activation of motivations to be negative. Capturing a richer correlation structure not only

provides additional insights, but is also particularly useful in a context with sparse information. We emphasize that lifting these limitations, while conceptually attractive, results in a large increase in the computational costs of traditional inference algorithms. We resolve this issue by developing a custom-made variational inference algorithm, outlined in Section 3.

Trusov et al. (2016) provide another application of LDA in marketing. They model website browsing behavior using an adaptation of LDA, where each household’s browsing history is divided into smaller time periods. For each of these periods, the relevance over the topics is a function of both observed and unobserved heterogeneity and a lagged effect of the previous browsing period. Although the modeling approach in Trusov et al. (2016) is conceptually similar to ours, the scale of their application is several orders of magnitude smaller. They aggregate the browsing data to 29 website categories, where we consider an assortment that contains over 4,000 products in our application. Similar to us, they consider correlations between topic relevance at the customer level. However, their estimation procedure does not scale to a large number of topics, reflected by the presence of only 7 topics in their application. This is in stark contrast to the 100 topics we consider in our application. The need for computationally efficient algorithms to estimate models involving large-scale data was also recently highlighted in Wedel and Kannan (2016).

Dew et al. (2019) provides another example of modeling heterogeneity using LDA. To study the evolution of product reviews, they extend LDA with a dynamic heterogeneity structure that is modeled using Gaussian processes. Although Dew et al. (2019) provide a flexible approach to capturing dynamic heterogeneity, they mention in their conclusion that their proposed method does not scale to large applications. In contrast, we introduce an inference algorithm that enables fast Bayesian inference in large retailing settings.

2.2. Connecting topic models to purchase behavior

In our model we analyze and describe purchase behavior using a relatively small set of latent dimensions. Each of these dimensions describes a specific pattern in the purchase data. Following the nomenclature introduced in Jacobs et al. (2016), these dimensions correspond to latent *purchase motivations* that drive the observed purchase behavior in the customer base. Purchase motivations enable a marketing manager to reason about purchase behavior at a higher abstraction level, which can generate more insights than analyzing individual products separately. Not only do these motivations shed light on relationships between products spanning different product categories, they also serve as input for marketing actions, e.g. through personalization of such actions (Ansari and Mela 2003). In addition, a managerial model-based dashboard (Dew and Ansari 2018) can be constructed to help answer specific customer-behavior questions, e.g. why customers visit the store during certain time periods. Targeted advertising can also be improved, based on insights derived from the purchase motivations. A final example is improvements in store layout and how products are positioned relative to each other in the store, either online or offline, depending on the motivations the products are connected with.

To infer the set of purchase motivations from purchase history data, we build a new model based on the framework of probabilistic topic models (Blei et al. 2003, Blei 2012). Topic models are typically used to identify and learn about latent topics that are present in written documents. The high-level analogy between modeling text and purchase behavior is as follows: a document contains words, while a customer’s purchase history contains products. Each word stems from a predefined dictionary, while each product is purchased from a predefined assortment. A collection of documents can be summarized using a small set of topics, where each topic describes some latent theme in the text corpus; the purchase history for all customers can be summarized using a small set of motivations, where each motivation describes some preference for products.

A document can be succinctly described as a mixture of topics; a customer’s purchase history can be succinctly described as a mixture of purchase motivations. Such a mix of motivations enables a low-dimensional representation of a customer’s preferences over the products in the assortment.

Strictly following this analogy, purchase history data would be analyzed at the customer level with the time dimension being ignored (Jacobs et al. 2016). This implies that purchases made by a customer are exchangeable across purchase trips, which is unrealistic. Variation in a customer’s purchase behavior over time is to be expected and should be accounted for. For example, a customer could first visit the store for a bathroom renovation, while the next trip is for pool maintenance. We capture this systematic variation by modelling a customer’s purchases at each shopping trip, accounting for the interdependencies across a customer’s trips.

2.3. Modeling purchase behavior using motivations

Throughout the paper we use the following notation for the data. Products are indexed by $j = 1, \dots, J$, where J is the assortment size. Customers are indexed by $i = 1, \dots, I$, where I is the number of customers. Customer i makes B_i shopping trips. During shopping trip b customer i purchases N_{ib} products, collected in the set \mathbf{y}_{ib} . Each element in \mathbf{y}_{ib} corresponds to the index of a product: $y_{ibn} \in \{1, \dots, J\}$ for $n = 1, \dots, N_{ib}$. We purposefully ignore the purchase quantity of a product, as it is a measure that is difficult to meaningfully compare across different products and will (unintentionally) overemphasize products with high purchase quantities, e.g. consider that only a single hammer is needed versus many nails. However, in case purchase quantity contains relevant information, the model can trivially be extended by allowing the same product to occur repeatedly in \mathbf{y}_{ib} . Characteristics that are specific to the b th shopping trip made by customer i , such as time-of-day and day-of-the-week of the trip, are captured in the K_X -dimensional vector \mathbf{x}_{ib} . Similarly, variables specific to customer i , like demographics or customer profile information, are captured in the K_W -dimensional vector \mathbf{w}_i .

Conceptually, purchase motivations drive – either intrinsically or by extraneous factors – the preference of a customer for a certain subset of products in the assortment. In this sense the motivation drives a customer to the store. Examples of such motivations are the plan to organize a barbecue, resulting in a preference for products related to barbecuing, or an ongoing home renovation project, resulting in a need for paint and drywall material. The underlying idea of our model is to identify a set of M purchase motivations to describe the common purchase patterns present in the customers’ purchase histories (Jacobs et al. 2016). A given motivation induces the same type of purchase behavior across all customers. However, the relevance of each of the motivations naturally varies across shopping trips and customers. In practice, only very few motivations will be relevant at a single shopping trip, while most other motivations will not be active at all at that moment. Differences in purchase behavior across customers and shopping trips then result from variation in activated motivations.

An important feature of motivations is that they generally do not match with traditional product categorizations, like product groups or product classes. Many retailers use, or at least have used, such product categorizations, usually in the form of a product hierarchy tree. Motivations however, often span multiple product groups, as complementary products are jointly needed to achieve a goal, for example a pen and paper, or a hammer and nails. At the same time, a single product can be linked to more than one motivation. For example, working gloves can be used for gardening and for construction work.

This conceptualization of a motivation is captured in the model through a motivation-specific vector of purchase probabilities for the complete product assortment. Products that are strongly linked to the motivation will receive high purchase probabilities, while the other, less relevant

products will have probabilities close to zero. That is, each motivation $m = 1, \dots, M$ is characterized by ϕ_m , a J -dimensional probability vector, where $\phi_{mj} \geq 0$ denotes the probability that product j will be purchased if motivation m is activated, and $\sum_j \phi_{mj} = 1$.

Customers do not always shop with a single motivation in mind. Instead, they might go shopping for multiple motivations, e.g. the kitchen and bathroom could be renovated simultaneously. The set of products a customer buys in a single trip will then be driven by a mix of motivations. The presence of multiple motivations in a single shopping trip is captured by assuming that each shopping trip is driven by a mixture of motivations. The mixture weights that govern the importance of each of the M motivations for the b th shopping trip of customer i are given by a vector $\theta_{ib} = [\theta_{ib1}, \dots, \theta_{ibM}]$, where $\theta_{ibm} \geq 0$ and $\sum_m \theta_{ibm} = 1$. Motivations that are irrelevant receive a weight close to zero. Variation in the motivation mixture weights across customers and shopping trips creates heterogeneity in purchase behavior.

In sum, customer i selects a purchase motivation for each purchase decision $n = 1, \dots, N_{ib}$ in her b th shopping trip, denoted by $z_{ibn} \in \{1, \dots, M\}$. The selection of these motivations follows the motivation mixture weights for this shopping trip – i.e. the reasons for being in the store – such that:

$$\Pr [z_{ibn} = m | \theta_{ib}] = \theta_{ibm}. \quad (1)$$

Subsequently customer i buys a product based on the purchase probability vector that characterizes the motivation that drives this purchase decision, z_{ibn} . The probability that product j is purchased, given that the underlying motivation for this purchase decision is motivation m , therefore equals:

$$\Pr [y_{ibn} = j | z_{ibn} = m, \phi] = \phi_{mj}. \quad (2)$$

The probability vector over the assortment (ϕ_m) that is connected to a motivation is unknown to the researcher and needs to be inferred from the data. Central to the identification of these probabilities are the observed co-purchases of products within shopping trips. If a certain set of products tends to be purchased together in a shopping trip, and this co-occurrence is present in many different shopping trips, then these products are likely to align with some particular motivation to shop. When motivations are active across multiple shopping trips of a specific customer, the co-purchases of products across trips of this customer also help in the identification of such motivations. This is especially relevant for retail contexts that have many shopping trips, each consisting of very few purchases.

The above exposition connects the observed purchase behavior \mathbf{y} to the M purchase motivations. It extends the LDA-X model introduced in Jacobs et al. (2016), which only considers purchases made at the customer level and does not retain separate shopping trips. As a result, the b dimension related to the shopping trips is not present in the LDA-X model and information on the impact of shopping-trip specific variables \mathbf{x}_{ib} on purchase behavior is lost.

2.4. Modeling activation of purchase motivations

With purchase motivations driving purchase behavior, the next step is to model the relevance of purchase motivations across customers and shopping trips. In the standard LDA model (Blei et al. 2003), the mixture weights θ_{ib} are draws from a Dirichlet distribution: $\theta_{ib} \sim \text{Dirichlet}(\boldsymbol{\alpha})$. However, when modelling customer purchase behavior this has two major disadvantages.

First, the Dirichlet distribution specifies a very restrictive correlation structure. It imposes negative correlations between all pairs of motivations and these correlations are completely determined by the mean of the distribution. This stems from the fact that the Dirichlet distribution is characterized by a single parameter vector $\boldsymbol{\alpha}$, unlike, for example, a multivariate

Normal distribution that has separate parameters for its mean and covariance. This restrictive correlation structure is unlikely to reflect that of the motivation mixture weights. Some motivations are expected to be positively correlated, e.g. a gardening motivation and a swimming pool motivation, while others might be negatively correlated.

The second drawback relates to the complexity of estimating α . If plenty of information is available, i.e. if \mathbf{y}_{ib} contains many elements on average, the exact value of α is less important. Hence, in many traditional (text) applications of LDA this parameter is either fixed to some predefined value or set using heuristics (Wallach et al. 2009). In the context of purchase data, the number of products purchased in a given shopping trip is very small. Therefore, when analyzing purchases at the shopping trip level, the parameter vector of the Dirichlet distribution plays an important role and should be estimated (Jacobs et al. 2016). From a computational perspective, however, this estimation does not scale to applications that involve a large number of motivations, as the density of the Dirichlet involves a product of multiple gamma functions. This is further exacerbated when a customer-specific α_i is specified, for example as in the LDA-X model.

Instead, we opt for an alternative approach to model θ_{ib} that circumvents these drawbacks. In particular, we are inspired by the correlated topic model (CTM), which replaces the restrictive Dirichlet distribution on θ_{ib} by a more flexible logistic Normal distribution that allows for correlations between the motivations (Blei and Lafferty 2007). To be more precise, θ_{ib} is the softmax⁵ of an unrestricted stochastic parameter vector $\alpha_{ib} \in \mathbb{R}^M$:

$$\theta_{ib} \equiv \text{softmax}(\alpha_{ib}) = \frac{\exp(\alpha_{ib})}{\sum_m \exp(\alpha_{ibm})}. \quad (3)$$

The softmax function is a natural choice here as it outputs a probability vector given any input vector of real numbers. The scalar parameter α_{ibm} can be interpreted as a measure of the (latent) relevance of motivation m in the b th shopping trip of customer i , similar to latent utilities in market share attraction models (Bronnenberg et al. 2000) or the utility-based specification of the multinomial logit model (Train 2009).

Many factors relate to the relevance of each of the M motivations in a given shopping trip. Some of these factors can be attributed to a customer’s innate preferences and observed characteristics, while others may be driven by a shopping trip’s contextual factors such as the time the trip takes place. However, after accounting for these factors, some variation in the motivation relevance remains unexplained. To account for all this, we specify a linear model for α_{ibm} that includes a predictable component μ_{ibm} , and a random, unpredictable component ϵ_{ibm} :

$$\begin{aligned} \alpha_{ibm} &= \mu_{ibm} + \epsilon_{ibm} \\ &= \kappa_{im} + \alpha_{ib-1}^\top \boldsymbol{\rho}_m + \mathbf{x}_{ib}^\top \boldsymbol{\beta}_m + \mathbf{w}_i^\top \boldsymbol{\gamma}_m + \epsilon_{ibm}. \end{aligned} \quad (4)$$

The customer-specific intercept κ_{im} captures the innate preferences, i.e. the baseline relevance of motivation m for customer i across all shopping trips. The M intercepts for customer i are collected in the vector $\boldsymbol{\kappa}_i = [\kappa_{i1}, \dots, \kappa_{iM}]$. For $\boldsymbol{\kappa}_i$ we specify a multivariate Normal distribution with mean $\boldsymbol{\mu}_\kappa$ and covariance $\boldsymbol{\Sigma}_\kappa$. The vector $\boldsymbol{\mu}_\kappa$ describes the prevalence of each of the M motivations in the customer base, while the motivation correlations are captured in $\boldsymbol{\Sigma}_\kappa$.

The inclusion of these motivation correlations at the customer level is a major improvement over the commonly used Dirichlet distribution, as it enables the identification of complementarity across motivations. Knowledge of such correlations is particularly relevant for customers with

⁵In the marketing literature the softmax function is better known as the multinomial logit *function*, but we want to avoid confusion with the equivalent named multinomial logit *model*.

only a few observed shopping trips, for whom little information is available to determine their preferences. After observing a single shopping trip, we are able to identify the most likely motivation(s) from that trip, but also the motivations that are most (cor)related to them. Let us illustrate this by foreshadowing our empirical findings: we discover multiple distinct motivations related to gardening, e.g. buying plants during springtime and trash bags in fall. Intuitively these gardening motivations are correlated at the customer level and observing purchases in spring can be used to improve marketing actions by focusing on the correlated motivations relevant in fall.

Explanatory variables that are specific to the shopping trip, such as seasonality dummies, are included in the \mathbf{x}_{ib} vector, while customer-specific variables that are time invariant, e.g. gender, are described in \mathbf{w}_i . Naturally the relevance of each motivation will be affected differently by these explanatory variables, and therefore the corresponding parameter vectors β_m and γ_m are motivation specific. For example, the relevance of some motivations may have strong seasonal fluctuations, whereas other motivations will hardly vary by season.

Dynamics in the model are captured by the first-order vector autoregressive, VAR(1), term $\alpha_{ib-1}^\top \rho_m$, where the relevance of each of the M motivations in the previous shopping trip directly affects the relevance of motivation m in the current shopping trip. The persistence of motivation m is described by ρ_{mm} . The cross effect of motivation m' on m , with $m' \neq m$, is described by $\rho_{mm'}$ and in general this effect is not symmetric, i.e. $\rho_{mm'} \neq \rho_{m'm}$. Hence the model contains a full $M \times M$ matrix with VAR(1)-effects. Note that for a customer's first shopping trip, the lagged α_{ib-1} term is not available. We specify an alternative specification for μ_{i1m} which can be found in Appendix A.

Any variation in the relevance of motivation m that is left after accounting for all factors mentioned above is absorbed in the stochastic component ϵ_{ibm} . We make the following assumptions about ϵ_{ibm} : i) Across customers, ϵ_{ibm} is independent. ii) For a given customer, the specification of μ_{ibm} captures state dependence and variation across time. Therefore there is no autocorrelation between ϵ_{ibm} for $b = 1, \dots, B_i$. iii) The relevant motivation correlations are captured by Σ_κ in the model. After controlling for these correlations, and given the small number of purchases per shopping trip in our application, any remaining motivation correlation at the shopping trip level can be ignored. As a result, there is no correlation across motivations ϵ_{ibm} for $m = 1, \dots, M$. Combining these assumptions, we specify for ϵ_{ibm} a Normal distribution with zero-mean and motivation-specific variance $\sigma_{\alpha_m}^2$. This allows for heteroscedasticity, as the relevance of some motivations may be more variable than others. The model specification is completed with prior distributions for all population-level parameters, details are provided in Appendix A.

Our model nests the LDA-X model (Jacobs et al. 2016), where only the Dirichlet distribution is replaced with a logistic Normal distribution, cf. Blei and Lafferty (2007). Specifically, LDA-X assumes the motivation mixture weights to depend only on a population-level motivation activation parameter κ_m and customer-specific variables \mathbf{w}_i . This restricted specification results from our model specification in Equation (4) when we assume $\kappa_{im} = \kappa_m = \mu_{\kappa,m}$, $\Sigma_\kappa = \mathbf{0}$, $\rho_m = \mathbf{0}$, and $\beta_m = \mathbf{0}$. This shows that, apart from the distributional assumption for θ_{ib} , the LDA-X model is nested within – and a much simpler version of – our model.

2.5. Alternative approaches to modeling dynamic purchase behavior

In the product recommendation literature, matrix factorization techniques are often used to study dependencies in preferences among large sets of products (Koren et al. 2009). Such techniques however do not allow for the inclusion of customer or shopping-trip characteristics.

One must also choose to either aggregate over shopping trips or to treat each shopping trip completely independent of all other trips by the same customer. Ideas from matrix factorization are also present in dynamic factor models. Here the factorization is used inside a statistical or econometric model. For example Bruce et al. (2012) use such methodology to study the dynamic effects of advertising on sales. However, extending such ideas to a large scale and to discrete dependent variables is not trivial.

Ruiz et al. (2020) present a probabilistic model of consumer choice that also contains matrix factorization ideas. Similar to our model, it uses purchase history data, relies on variational inference, and allows for customer heterogeneity and dependencies across products. One important difference is that it allows for product-specific characteristics such as a product’s price. To model the dependencies across products, Ruiz et al. (2020) rely on the order in which products are chosen in a shopping trip. Interdependencies then arise with the current product choice depending on the products already bought in a trip.

This approach brings several challenges. First, the order of purchases is often not observed. Ruiz et al. (2020) solve this by integrating over all possible permutations using a simulation approach. Second, the dependence between products may not directly follow the sequence in time, as the order of purchases is to a large extent driven by the store layout. Ruiz et al. (2020) acknowledge this and allow the customer to “think one-step ahead”. However, this substantially adds to the complexity of the model and is only illustrated using a small-scale example in the paper. Third, our model accounts for dynamics across shopping trips, which is not possible in the approach of Ruiz et al. (2020).

Our model also differs substantially from Ruiz et al. (2020) in terms of the interpretation of the latent spaces. We build on the idea of motivations and specify probabilities for customers to have a certain motivation and for a product purchase to be driven by a motivation. Ruiz et al. (2020) use matrix factorization inside a multinomial logit model. Their model represents products and customers as vectors in a latent space, with purchase probabilities depending on the inner products of these vectors. Therefore, a customer will have a preference for products that appear in a certain region of the latent space, but a customer cannot prefer two regions of the latent space that are far apart. Under the concept of motivations such behavior is possible as a customer can have a high probability for two different motivations. In addition, our motivations have a direct interpretation, whereas matrix factorization requires post-processing to facilitate interpretation.

In sum, Ruiz et al. (2020) provide a strong and useful model, especially when marketers want to intervene *during* a shopping trip. It allows for characteristics such as price and gives insights in terms of complements and substitutes. We believe that our model provides easier interpretation, is more valuable when targeting customers who are not yet in the store, and is computationally more efficient.

3. Inference

The proposed model contains many latent variables and parameters that need to be estimated, we refer to this set of unknown components as Ω . The information that is available to infer Ω are the product purchases \mathbf{y} .⁶ We apply Bayesian statistical inference and the goal is to examine the posterior distribution of Ω : $p(\Omega|\mathbf{y})$. As in most models, it is not tractable to directly evaluate this distribution. Traditionally, this problem has been circumvented by using Markov chain Monte Carlo (MCMC) methods (Rossi and Allenby 2003), in which the posterior

⁶In our notation we implicitly condition on other exogenous information that is available, i.e. the parameters describing the prior distribution $p(\Omega)$ and the explanatory variables \mathbf{x} and \mathbf{w} .

is approximated by sampling from a Markov chain of which the stationary distribution is the posterior of interest. Asymptotically it is guaranteed that the Markov chain produces samples from the target posterior distribution. For practical purposes however, convergence of the chain to the posterior distribution can be too slow, especially for model structures that result in complex posteriors (Carpenter et al. 2017). Given the hierarchical structure of our model and the large number of customer-specific parameters, using Hamiltonian Monte Carlo (HMC) (Neal 2011) is also ineffective. The complexity of HMC scales with the size of the parameter space and this complexity cannot be simplified using conditional independence assumptions present in hierarchical models.

Variational inference (VI) is an alternative inference technique that is fast and scalable. VI stems from the machine learning literature and works particularly well for large-scale models (Blei et al. 2017). The general idea is to transform posterior inference into an optimization problem. The objective here is to find the distribution that best approximates the true posterior, from a prespecified class of distributions. By limiting the class of distributions to those that are computationally convenient, one can closely approximate the posterior in a computationally efficient way (Jordan et al. 1999). Typically, the output of VI accurately describes posterior means, but underestimates posterior variances (Blei et al. 2017).

The fact that VI yields an approximation of the true posterior distribution is a theoretical disadvantage compared to the asymptotically exact sampling methods. However, there are several practical advantages that justify its use. First, convergence in VI is typically fast – much faster compared to sampling-based methods – and it is possible to monitor this convergence reliably (Ormerod and Wand 2010). Second, the output of VI is a distribution that is parameterized by a set of parameters. This is in contrast to the sampling-based methods, where for each unknown model component a long chain of samples is needed to accurately approximate the posterior, creating a large memory burden for models with many unknowns. Third, VI can be sped up using advances from the optimization literature. Examples are stochastic subsampling, stochastic gradients, and parallelization of computations (Hoffman et al. 2013, Kucukelbir et al. 2017). These optimization techniques lend themselves well for the estimation of large hierarchical Bayesian models.

To date, VI has been adopted in few marketing papers (Braun and McAuliffe 2010, Dzyabura and Hauser 2011, Ansari et al. 2018, Xia et al. 2019). Given the scale of our application and the complexity of our model, we turn to VI as well. We derive an inference algorithm that is customized to our model structure and is computationally highly efficient. In Section 3.1 we discuss VI in general and describe our implementation of this inference technique. Compared to a straightforward application of standard VI techniques our implementation yields a better approximation to the posterior distribution, but potentially comes at a high computational cost. The need for a repeated (numerical) calculation of the inverse of a large matrix lies at the root of this. We derive an analytical solution for the inverse that completely alleviates this computational burden in Section 3.2. Section 3.3 discusses further ways to increase the computational efficiency in the context of very large data sets. In Section 3.4 we describe how to interpret the effect sizes of the explanatory variables in our model.

3.1. Estimation using variational inference

In variational inference, the posterior inference problem is cast as an optimization problem where the search space is constructed of probability distributions. We define \mathcal{Q} as a set of joint probability distributions over the unknown model components Ω . A specific distribution – and corresponding density – is denoted by q , i.e. $q(\Omega) \in \mathcal{Q}$. We refer to q as a *variational distribution*. The objective is to find the distribution q in \mathcal{Q} that best matches the posterior

distribution $p(\boldsymbol{\Omega}|\mathbf{y})$. In VI the match between two probability distributions is typically measured using the Kullback-Leibler (KL) divergence. The KL divergence from some distribution $q(x)$ to another distribution $p(x)$ is given by:

$$\text{KL}\{q(x)||p(x)\} \equiv \int_x \log \frac{q(x)}{p(x)} q(x) dx = \tilde{\mathbb{E}}\{\log q(x) - \log p(x)\}, \quad (5)$$

where $\tilde{\mathbb{E}}$ denotes an expectation under the variational distribution $q(x)$. The objective of VI is to find the distribution $q^*(\boldsymbol{\Omega})$ in \mathcal{Q} that minimizes the KL divergence to the posterior distribution $p(\boldsymbol{\Omega}|\mathbf{y})$:

$$q^*(\boldsymbol{\Omega}) \equiv \arg \min_{q(\boldsymbol{\Omega}) \in \mathcal{Q}} \text{KL}\{q(\boldsymbol{\Omega})||p(\boldsymbol{\Omega}|\mathbf{y})\}. \quad (6)$$

The KL divergence can be rewritten as follows:

$$\log p(\mathbf{y}) - \text{KL}\{q(\boldsymbol{\Omega})||p(\boldsymbol{\Omega}|\mathbf{y})\} = \tilde{\mathbb{E}}\{\log p(\mathbf{y}, \boldsymbol{\Omega})\} + \text{ENTROPY}\{q(\boldsymbol{\Omega})\}. \quad (7)$$

The machine learning literature refers to the right-hand side of Equation (7) as the *evidence lower bound*, or ELBO (Blei et al. 2017). This is a lower bound for $\log p(\mathbf{y})$, as the KL divergence is non-negative by definition. Furthermore, Equation (7) shows that minimizing $\text{KL}\{q(\boldsymbol{\Omega})||p(\boldsymbol{\Omega}|\mathbf{y})\}$ is equivalent to maximizing the ELBO, because $\log p(\mathbf{y})$ is constant with respect to $q(\boldsymbol{\Omega})$. Hence, the objective of VI is to find the variational distribution $q(\boldsymbol{\Omega})$ that maximizes the variational expectation of the model’s log joint density, $\tilde{\mathbb{E}}\{\log p(\mathbf{y}, \boldsymbol{\Omega})\}$, while simultaneously spreading its density across different parameter configurations, as the entropy term $\text{ENTROPY}\{q(\boldsymbol{\Omega})\}$ penalizes for a variational distribution that centers all its probability mass around a single configuration for $\boldsymbol{\Omega}$, i.e. if $q(\boldsymbol{\Omega})$ resembles a point mass.

In case no constraints are placed on \mathcal{Q} , the solution $q^*(\boldsymbol{\Omega})$ is equivalent to the posterior distribution $p(\boldsymbol{\Omega}|\mathbf{y})$. However, this is not a computationally feasible solution as $p(\boldsymbol{\Omega}|\mathbf{y})$ cannot be evaluated directly. Instead, constraints should be imposed on \mathcal{Q} such that the resulting set of distributions is tractable and yet, still results in a good fit to the posterior distribution of interest. A popular restriction in the machine learning literature is to rely on the mean-field assumption (Bishop 2006), which states that each joint distribution in \mathcal{Q} factorizes over the unknown parameters $\boldsymbol{\Omega}$ according to some partitioning $F(\boldsymbol{\Omega})$, with $\bigcup_{\omega \in F(\boldsymbol{\Omega})} \omega = \boldsymbol{\Omega}$, such that:

$$q(\boldsymbol{\Omega}) = \prod_{\omega \in F(\boldsymbol{\Omega})} q(\omega). \quad (8)$$

Given the partitioning $F(\boldsymbol{\Omega})$, the variational distribution that maximizes the ELBO can be found by employing a coordinate ascent algorithm that iterates over the subsets of parameters $\omega \in F(\boldsymbol{\Omega})$, and updates each corresponding variational distribution $q(\omega)$ (Bishop 2006). This algorithm is guaranteed to reach at least a local optimum (Boyd and Vandenberghe 2004).

In general, by imposing this mean-field assumption, \mathcal{Q} no longer contains the true posterior distribution. As a result, $q(\omega)$ should be interpreted as the variational approximation of the marginal posterior distribution of ω : $p(\omega|\mathbf{y}) = \int_{\boldsymbol{\Omega}_{\setminus\omega}} p(\omega, \boldsymbol{\Omega}_{\setminus\omega}|\mathbf{y}) d\boldsymbol{\Omega}_{\setminus\omega}$ (Jordan et al. 1999), where $\boldsymbol{\Omega}_{\setminus\omega}$ refers to all elements of $\boldsymbol{\Omega}$ except ω . Note that the mean-field assumption does not imply that the solution for the optimal variational distribution for ω , $q^*(\omega)$, is independent from the variational distributions for $\boldsymbol{\Omega}_{\setminus\omega}$. This follows from the fact that we are trying to find the variational distribution that best fits the complete posterior distribution $p(\boldsymbol{\Omega}|\mathbf{y})$, in which the parameters depend on each other according to the model specification, irrespective of the partitioning $F(\boldsymbol{\Omega})$.

When selecting the partitioning $F(\Omega)$, a trade-off is required between computational feasibility and the quality of the variational approximation. A finer partitioning is computationally easier but also less accurate, while no partitioning results in the exact posterior distribution as the optimal variational distribution, but this is computationally intractable. The most fine-grained partitioning has each ω as a singleton set containing one (scalar) parameter. This implies that $q(\Omega)$ factorizes across the elements of a multivariate parameter as well, removing any posterior correlation between these elements; an unrealistic assumption in practice. To improve the quality of the variational approximation, the partitioning $F(\Omega)$ should not split the multivariate parameters in the model.

VI has been implemented for the standard LDA model without partitioning the Dirichlet distributed parameters (Blei et al. 2003, Hoffman et al. 2013). In contrast, in the CTM model, Blei and Lafferty (2007) factorize the variational distribution for each multivariate Normal parameter into a set of independent univariate Normal distributions. This simplifying assumption facilitates estimation but reduces the quality of the variational approximation, particularly when there is substantial estimation uncertainty. In our case this is expected for the customer and shopping-trip specific parameters, as most customers tend to buy only a few products per trip.

We therefore specify a partitioning $F(\Omega)$ that retains all elements of a multivariate parameter within a single subset, ω , for *all* multivariate parameters in the model, including the customer-specific multivariate Normal parameter κ_i . Details of the resulting variational inference algorithm are provided in Appendix B and Online Appendix D.

3.2. Computational efficiency

The partitioning $F(\Omega)$ in our implementation of variational inference is guaranteed to improve the quality of the variational solution, but is computationally costly. Specifically, it involves the optimization of many multivariate Normal variational distributions. For example, each κ_i has a customer-specific variational distribution $q(\kappa_i) = \text{MVN}(\tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}_i)$. Determining the optimal value for $\tilde{\boldsymbol{\Sigma}}_i$ in each iteration of the optimization routine involves computing the inverse of an $M \times M$ matrix that is specific to customer i of the following form (cf. Equations (D.6) and (D.7)):

$$\tilde{\boldsymbol{\Sigma}}_i = (\tilde{\mathbb{E}}\{\boldsymbol{\Lambda}_\kappa\} + d(\tilde{\mathbb{E}}\{\boldsymbol{\tau}_\alpha\})s_i)^{-1}, \quad (9)$$

where s_i is a customer-specific scalar, $\boldsymbol{\Lambda}_\kappa \equiv \boldsymbol{\Sigma}_\kappa^{-1}$, and $\boldsymbol{\tau}_\alpha \equiv [\sigma_{\alpha_1}^{-2}, \dots, \sigma_{\alpha_M}^{-2}]$. $\tilde{\mathbb{E}}$ denotes an expectation under the marginal variational distribution for a parameter, e.g. $\tilde{\mathbb{E}}\{\boldsymbol{\Lambda}_\kappa\} \equiv \mathbb{E}_{q(\boldsymbol{\Lambda}_\kappa)}\{\boldsymbol{\Lambda}_\kappa\}$, and $d(\cdot)$ is a function that outputs a diagonal matrix based on an input vector.

Therefore, in a naive implementation, each iteration of the optimization algorithm has a computational complexity of at least $\mathcal{O}(I \times M^3)$, which clearly does not scale well with the number of customers I nor the number of motivations M . As a result, our improved partitioning would render this approach computationally infeasible in large applications of the model.

We completely remove this computational burden by exploiting a special mathematical structure in the optimal value for $\tilde{\boldsymbol{\Sigma}}_i$. Note that $\tilde{\mathbb{E}}\{\boldsymbol{\Lambda}_\kappa\}$ and $d(\tilde{\mathbb{E}}\{\boldsymbol{\tau}_\alpha\})$ are non-singular precision matrices that are specified at the population level and hence, shared between all customers $i = 1, \dots, I$. The only customer-specific part is s_i , the scaling factor for $d(\tilde{\mathbb{E}}\{\boldsymbol{\tau}_\alpha\})$. Our computational shortcut is obtained from the fact that $\tilde{\boldsymbol{\Sigma}}_i$ can be rewritten as:

$$\tilde{\boldsymbol{\Sigma}}_i = (\mathbf{L}^{-1})^\top \mathbf{U}d((\mathbf{v} + s_i)^{-1})\mathbf{U}^\top \mathbf{L}^{-1}, \quad (10)$$

where \mathbf{L} is the lower triangular (in this case diagonal) Cholesky root of $d(\tilde{\mathbb{E}}\{\boldsymbol{\tau}_\alpha\})$ and $\mathbf{U}d(\mathbf{v})\mathbf{U}^\top$ is the singular value decomposition of $\mathbf{L}^{-1}\tilde{\mathbb{E}}\{\boldsymbol{\Lambda}_\kappa\}(\mathbf{L}^{-1})^\top$. Appendix C provides a proof of this

result. This equivalence shows that $\tilde{\Sigma}_i$ can be computed for any customer i without directly taking the inverse of a customer-specific $M \times M$ matrix. Similar results can be obtained for the covariance matrices of $q(\boldsymbol{\rho}_m)$, $q(\boldsymbol{\beta}_m)$, and $q(\boldsymbol{\gamma}_m)$. This result enables us to obtain a better approximation for the posterior distribution, without incurring insurmountable computational costs when many customers and motivations are involved. This result is also an addition to the variational inference literature and can directly be applied to other hierarchical models involving multivariate Normals, such as the CTM.

3.3. Scalability of the inference algorithm to very large data sets

Because of the efficient matrix inverse identity presented in Section 3.2, the vast majority of the computation time of our inference algorithm is spent on the update of the variational distributions that are specific to a product purchase or a shopping trip, i.e. z_{ibn} and $\boldsymbol{\alpha}_{ib}$. For a given number of motivations M , the computation time of a single iteration scales linearly with the number of shopping trips and purchases. In other words, if a large data set doubles in size, the required computational time will approximately double as well.

However, because the customers are conditionally independent in the model, there are multiple ways to improve scalability. First, the customer-specific optimizations can be parallelized over the customers, so the available computing power can be easily increased by using a computing cluster. Second, stochastic optimization techniques can be used, e.g. Hoffman et al. (2013), which reduce the number of epochs needed to achieve convergence, especially for large datasets. This lets our inference algorithm scale to data sets of virtually any size, which is an important advantage of estimation using variational inference. Note that even without such extensions the required computational time of our model is relatively low. For the empirical application in this paper estimation is completed in a matter of hours on a standard laptop. In this time we complete 1,000 iterations of the algorithm for $M = 100$, while model convergence was already achieved within the first few hundred iterations.

3.4. Quantifying the effects of explanatory variables

Our model contains several explanatory variables – both latent ($\boldsymbol{\alpha}_{ib-1}$) and observed ($\mathbf{x}_{ib}, \mathbf{w}_i$) – that affect the relevance of motivations at the shopping-trip level. To interpret the model results, the effect sizes of these explanatory variables should be judged. The corresponding coefficients ($\boldsymbol{\rho}_m, \boldsymbol{\beta}_m, \boldsymbol{\gamma}_m$), could be evaluated directly to learn about the linear effects of the explanatory variables on the relevance of motivation m : α_{ibm} . However, motivation relevance is an abstract latent construct that is not straightforward to interpret. A more tangible model component is the vector of motivation-activation probabilities $\boldsymbol{\theta}_{ib}$, defined as the softmax of $\boldsymbol{\alpha}_{ib}$, cf. Equation (3). Understanding how these probabilities are affected by the explanatory variables enables us to answer managerially relevant questions such as: “How does gender affect motivation activation?” or “How does time of the day shift the likelihood of various motivations?”. To this end we calculate odds ratios for the motivation-activation probabilities, where we contrast the motivation probabilities corresponding to a certain value of the characteristics to those of the “average” shopping trip.

Because θ_{ibm} is a non-linear function of $\boldsymbol{\alpha}_{ib}$, the effect of a focal explanatory variable depends on $\mu_{ibm} = \kappa_{im} + \boldsymbol{\alpha}_{ib-1}^\top \boldsymbol{\rho}_m + \mathbf{x}_{ib}^\top \boldsymbol{\beta}_m + \mathbf{w}_i^\top \boldsymbol{\gamma}_m$ and the disturbance term ϵ_{ib} , cf. Equation (4). Hence, to provide an interpretation to the effect of a focal explanatory variable, sensible baseline values for all explanatory variables need to be specified and integration over the distribution of ϵ_{ib} is needed. We use the “average” shopping trip as a natural baseline. For this baseline we set the exogenous explanatory variables \mathbf{x}_{ib} and \mathbf{w}_i to their sample means, the motivation intercepts

in κ_i to their population-level mean (μ_κ), and the lagged α_{ib-1} to the average posterior mean over all α_{ib} . We use the posterior means for the population-level parameters, as the information available for those parameters is large, and therefore their posterior variance is small. Together, this set of baseline values is used to compute μ^B .

Next, we consider a particular characteristic and, ceteris paribus, change its value relative to the baseline. For a continuous characteristic we consider a shock relative to its baseline value, while for a discrete characteristic we simply consider a particular level and set the explanatory variables accordingly. These shifted values are used to compute μ^S .

The odds ratio of a characteristic is then defined by taking the ratio of the probability of motivation m after the corresponding shift in variables (θ_m^S) and the baseline probability (θ_m^B), while integrating out the disturbance, that is:

$$\mathbb{E} \left[\frac{\theta_m^S}{\theta_m^B} \right] = \int_{\boldsymbol{\varepsilon}} \frac{\exp(\mu_m^S + \varepsilon_m) / \sum_{\ell=1}^M \exp(\mu_\ell^S + \varepsilon_\ell)}{\exp(\mu_m^B + \varepsilon_m) / \sum_{\ell=1}^M \exp(\mu_\ell^B + \varepsilon_\ell)} p(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon}, \quad (11)$$

where $p(\boldsymbol{\varepsilon})$ denotes the density of $\boldsymbol{\varepsilon}$, i.e. $\varepsilon_m \sim N(0, \sigma_{\alpha_m}^2)$ for $m = 1, \dots, M$. This odds ratio can be computed for – and is comparable across – all characteristics, as the baseline μ^B is the same for each.

4. Data

We apply our model to in-store purchase history data recorded at the shopping-trip level that is made available to us by a Fortune 500 Specialty Retailer.⁷ The data contains purchases made by customers in one of their retail stores in Florida during a 24-month period that ranges from March 5, 2012 to March 4, 2014. Customers are known to the retailer, such that different shopping trips can be linked to the same customer. In addition to purchase behavior some information on customer demographics is available, including age, gender, and household size. Descriptive statistics for these variables are presented in Online Appendix E.

The raw data contains information on purchase incidence for 29,027 distinct products by 2,259 distinct customers. The majority of these products is purchased very infrequently: 25,726 products are purchased in less than 10 shopping trips during the 24-month time span. In principle our model works for very infrequent 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 between infrequently purchased products. Instead of removing the infrequently purchased products from the data altogether, we aggregate the infrequent products according to the firm’s product taxonomy.⁸

The product taxonomy defines a product as a unique combination of [Group, Class, Subclass, Description]. For example a product in the data is [Group = BUILDING MATERIALS, Class = GYPSUM, Subclass = BOARD, Description = “1/2”X4’X8’ USG MOLDTOUGH DRY-WALL”]. Another product within the same [Group, Class, Subclass]-combination but with a (slightly) different description such as “1/2”X4’X8’ USG ULTRALIGHT DRYWALL” is considered to be a different product in the data. Products that are purchased very infrequently are aggregated according to this product taxonomy as follows. For each [Group, Class, Subclass]

⁷We are grateful to Wharton Customer Analytics (WCA) for setting up the research opportunity that has connected us to this retailer.

⁸We emphasize that the product taxonomy is only used to aggregate infrequent products in the data. More specifically, the product taxonomy is not used in the model and the identification of motivations does not rely on the product taxonomy.

combination, the corresponding infrequent products are merged. If the aggregate product is still infrequently purchased, the same step is repeated for each [Group, Class] combination and, if necessary, subsequently at the [Group] level. At the end of this aggregation process, only 19 infrequent products remain that are purchased less than 10 times, corresponding to 34 purchases in the data. These products have been removed from the data.

The resulting data set contains 139,622 purchases out of an assortment of 4,266 distinct products. The purchases are made by 2,259 unique customers across 47,568 shopping trips. Some descriptive statistics of the purchase data are displayed in Table 1. The statistics illustrate a loyal customer base, with on average 21.06 shopping trips per customer. However, the amount of information per shopping trip is small. The average number of products purchased per shopping trip is only 2.94 and on average a product in the assortment is purchased in 32.73 shopping trips. Such figures are representative for other modern retailing environments, where the cost of holding a large product assortment is low and customers are encouraged to place many small orders with low shipping costs, e.g. consider the subscription service Amazon Prime. At the product level the data is sparse, with more than half of the products being purchased less than 20 times across the almost 50,000 shopping trips.

	Mean	Mode	Min	25% perc.	Median	75% perc.	Max
Purchases per trip	2.94	1	1	1	2	4	50
Purchases per product	32.73	10	10	12	18	31	724
Trips per customer	21.06	1	1	8	16	28	257
Purchases per customer	61.81	1	1	21	46	85	645

Table 1: Descriptive statistics of the purchase history data.

We split the data set in two parts: an in-sample part – used to estimate the model parameters – and an out-of-sample part – used to determine the model’s predictive performance. As out-of-sample data we take the last shopping trip for every customer that has visited the store more than once. Characteristics of the different data sets are displayed in Table 2.

	Products	Customers	Trips	Purchases
Complete	4,266	2,259	47,568	139,622
Estimation	4,266	2,259	45,473	134,049
Out-of-sample	2,396	2,095	2,095	5,573

Table 2: Characteristics of the different sets of purchase history data.

5. Results

The results in this section are organized as follows. We first describe some high-level characteristics of the inferred set of purchase motivations in Section 5.1. Using these motivations, we describe and visualize the customer journey for two customers in Section 5.2. In Section 5.3 we enrich our understanding of the motivations by examining how the relevance of motivations are affected by both the timing of the trip and customer characteristics. The relations between motivations based on the correlations and VAR(1)-effects are discussed in Section 5.4. We conclude in Section 5.5 by comparing the predictive performance of our model against several benchmarks.

The results in this section are based on the inferred variational posterior distribution of the model parameters. Generally, closed-form solutions do not exist for non-linear functions of parameters under the posterior distribution. As an abundance of information is available for the population-level parameters, and hence very little posterior uncertainty, we use the posterior mean value when evaluating functions that involve these parameters. In contrast, for parameters that relate to either a shopping trip or a customer, much less information is available, and we rely on Monte Carlo integration with 250,000 samples to account for the posterior uncertainty.

5.1. Purchase motivations

Based on the large number of customers, products, and shopping trips, we expect substantial heterogeneity in purchase behavior. In the model we can account for this by setting the number of purchase motivations, M , to a large value. At the same time we are interested in gaining substantive insights from the data, i.e. identifying motivations that are relevant from both a managerial and a customer perspective. This implies that M should also not be too large.

One approach to determine M is to use some performance measure, e.g. predictive performance on hold-out data. However, this does not factor in interpretability of the model and one could end up with too many motivations for the model to be of practical use. Instead we set $M = 100$. We anticipate that this configuration allows us to identify the salient purchase patterns in the data, as well as more specific patterns that may only be relevant for a small subset of the shopping trips or customers. Furthermore, Jacobs et al. (2016) have shown that specifying a value for M that exceeds the actual number of motivations does not significantly impede predictive performance, as long as the proportions for the motivations are estimated at the population level, which is the case for our model. In addition, the ability to deal with such a large number of latent motivations also exhibits the scalability of our model.

The size of each of the M purchase motivations is displayed in Figure 1, which reports the expected number of purchases for each motivation under the posterior distribution, ranging from 4021.36 purchases (3.00%) for motivation 10 to 717.48 purchases (0.54%) for motivation 9. This shows that there is variety in motivation size, and that all motivations are relevant.

Each motivation m is characterized by ϕ_m , a vector of purchase probabilities for all products in the assortment. For a motivation to be managerially useful, it should relate to a clear and relatively small subset of the assortment, i.e. ϕ_m should be a sparse probability vector with many values close to zero and a limited number of large purchase probabilities. Figure 2 shows, for each motivation m , the minimum number of distinct products needed to account for at least 50% of the product purchases under that motivation. The vast majority of the motivations is very sparse as more than half of their product purchases are covered by fewer than 10 products, which is a very small fraction of the 4,266 products in the whole assortment. The largest

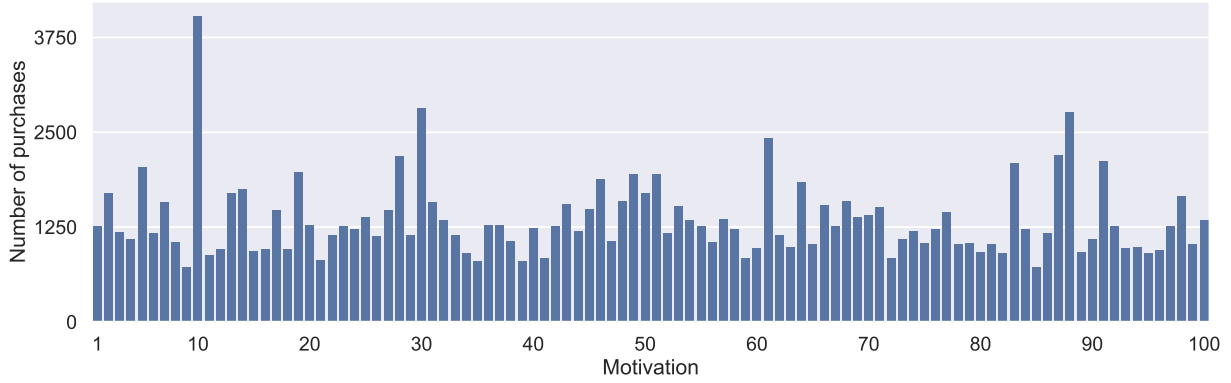


Figure 1: Size of motivations (expected number of purchases assigned to each motivation).

motivation identified ($m = 10$) is also the motivation that is least sparse. Intuitively this makes sense, as a broad motivation is more likely to appeal to a large part of the customer base.

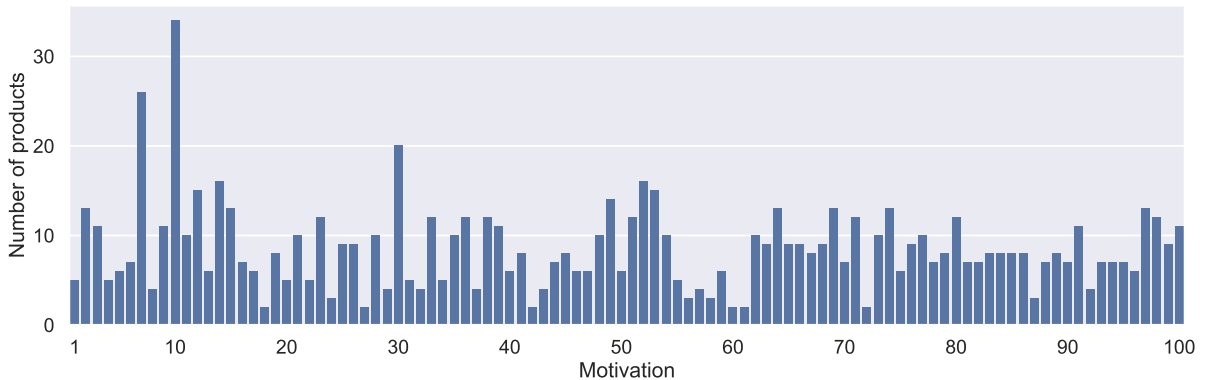


Figure 2: Sparsity of motivations (minimum number of products needed to account for the majority ($\geq 50\%$) of the probability mass in ϕ_m).

Motivations are characterized by the products that are most likely under that particular motivation. The larger the cumulative probability of these products, the better they summarize the whole motivation. However, summarizing and labelling a motivation is subjective. An expert’s opinion, such as that of a product manager at the retailer, will facilitate this task. In absence of such an expert we rely on the available product taxonomy and introspection for labeling the M motivations. We emphasize that the product taxonomy is not used in the model to identify the motivations, but is solely used to facilitate interpretation of the model output.

Table 3 presents the labels we assigned to the 10 largest motivations based on the most likely products for each motivation. It also shows the cumulative purchase probability covered by the 10 most likely products under each motivation, which varies substantially across motivations. For the largest motivation ($m = 10$) it is 23.42%, again indicating that this is a broad motivation with probability mass allocated to a relatively large number of products. On the other hand, for the fourth largest motivation ($m = 60$) the cumulative purchase probability for the 10 most likely products purchased under that motivation is 84.69%, so these 10 products almost completely describe this motivation.

An important message from this paper is that marketers should use motivations instead of – or at least besides – existing product categories, because customer purchase behavior spans across multiple product categories. To illustrate this we display the 10 most likely products with information from the product taxonomy for motivation 51 (rank 12) in Table 4 and motivation

Rank	m	Label	Cum. Prob. Top 10
1	10	Painting: Paint tools and supplies	23.42%
2	30	Cleaning: General cleaning supplies	38.78%
3	88	Gardening: Annuals and perennials	59.63%
4	61	Hardware: Fasteners (screws, nuts, bolts, washers)	84.69%
5	87	Gardening: Landscaping (mulch and top soils)	87.87%
6	28	DIY: Wall plates	50.88%
7	91	DIY: Electrical installations	48.73%
8	83	DIY: Tile installation	59.43%
9	5	Gardening: Seeds, vegetables, herbs	66.56%
10	19	Cleaning: Floors	56.48%

Table 3: Labels for the 10 largest motivations with the cumulative purchase probability of the 10 most likely products under each motivation.

28 (rank 6) in Table 5.⁹ Motivation 51 is related to do-it-yourself projects concerning deck and fence installations. Motivation 28 leads to purchases of wall plates, that is, the installation of light switches and power outlets. Table 4 shows that the 10 most likely products for motivation 51 are spread out across 3 groups, 6 classes, and 8 subclasses. This motivation is an example of a purchase pattern that clearly covers products from multiple product categories. Table 5 for motivation 28, on the other hand, tells a different story: the 10 most likely products are contained in a single group and two classes, although still spread out over 5 different subclasses. Hence, this purchase pattern is more in line with the existing product taxonomy. Our modeling approach has the flexibility to capture both scenarios.

Prob.	Group	Class	Subclass	Description
7.30	Lumber	Pressure trtd wood	PT dimensional lumber	2x4-8ft #2 prime pt weathershield
6.37	Building materials	Metal products	Metal prod/simpson	–
4.93	Hardware	Fasteners	Deck screws	–
4.88	Hardware	Builder’s hardware	Gate hardware	–
4.47	Lumber	Fencing	Pressure treated pickets	5/8”x5-1/2”x6’ pt pine dog ear pckt
4.32	Lumber	Pressure trtd wood	PT timbers	4x4-8ft #2 pt
3.55	Lumber	Pressure trtd wood	PT dimensional lumber	2x4-8ft #2 pt
3.44	Building materials	Concrete	Concrete mixes	50lb sakrete fast-set concrete
3.01	Hardware	Fasteners	Construction/frming nail	–
2.88	Lumber	Pressure trtd wood	PT dimensional lumber	–

Table 4: 10 most likely products (cum. prob. 45.15%) for $m = 51$ (“DIY: Deck and fence installation”).

Prob.	Group	Class	Subclass	Description
11.58	Electrical	Wiring devices	Wall plates (commodity)	–
10.02	Electrical	Wiring devices	Receptacles	–
7.65	Electrical	Wiring devices	Switches	–
6.32	Electrical	Wiring devices	Wall plates (decorative)	–
3.89	Electrical	Wiring devices	Wall plates (commodity)	1g wht nyl midway outlet wallplt
2.66	Electrical	Wiring devices	Wall plates (commodity)	1g wht duplex wallplt
2.47	Electrical	Wiring devices	Wall plates (commodity)	1g wht decora wallplt
2.38	Electrical	Wiring devices	Wall plates (commodity)	1g wht nyl midway decora wallplt
2.15	Electrical	Wiring devices	Switches	20/10a nkl decor on/offsp tggf swtch
1.76	Electrical	Conduit/boxes/fittings	PVC box/covers/access	Old work 1g 14cu

Table 5: 10 most likely products (cum. prob. 50.88%) for $m = 28$ (“DIY: Wall plates”).

⁹Missing information in the product taxonomy, e.g. a description for an aggregated product, is indicated by “–”.

Some of the motivations highlight products that are not very frequently purchased, while other motivations are of course driven by high volume products. To illustrate this, we consider the purchase-frequency ranks of the most likely products under each motivation. The product with rank = 1 has the highest purchase volume in the data, which is a bag of fasteners (e.g. screws, nuts, bolts). It is the most likely product under motivation 61, which indeed relates to fastener products. More interestingly, motivation 7, which relates to exterior paint jobs and waterproofing, places the highest purchase probability on an exterior paint product, which only has a rank of 588 in the data. Descriptive statistics of the ranks for the five most likely products under each motivation are given in Table 6. For the most likely product, the average rank of 116.49 indicates that other motivations also highlight products that are relatively infrequently purchased. If we look beyond the single most likely product under each motivation, we notice that even more products in the tail of the assortment are identified as highly relevant for a motivation. This shows that by using the motivations, our model is able to highlight purchase patterns that also involve low-volume products.

	Mean	Min	25% perc.	Median	75% perc.	Max
Rank of most likely product	116.49	1	31	78	166	588
Rank of second most likely product	206.03	3	78	177	277	626
Rank of third most likely product	296.56	4	156	277	382	755
Rank of fourth most likely product	413.28	61	235	382	571	1128
Rank of fifth most likely product	528.71	25	301	464	713	1676

Table 6: Descriptive statistics of the purchase-frequency rank of the five most likely products under each motivation. The statistics are computed across the M motivations.

Another advantage of our method is that it results in soft clusters of products, that is, the same product can be relevant for more than one purchase motivation. For example, we identified several motivations related to plumbing, each involving pipes of different widths. For each of these motivations the “1/2” X260” PTFE THRD SEAL TAPE” (PTFE Thread Seal Tape) product receives a relatively high purchase probability, which intuitively makes sense as it is needed in different plumbing projects. In a hard clustering approach this product could only have been assigned to a single motivation. Similar examples in our empirical application are identified for multi-use products such as paint brushes, caulks, and cement mixes.

5.2. Customer journey

A customer’s journey at the retailer can be succinctly described and visualized using the identified motivations. In Figures 3 and 4 we illustrate the journey for two customers at the retailer that each have made 10 shopping trips. For each shopping trip the motivation-activation probabilities are displayed as a stacked bar plot, where we focus on motivations that have a substantial probability in at least one of the shopping trips for a customer. From these figures several distinct patterns can be identified.

Customer 144 in Figure 3 is primarily interested in gardening activities, combined with a DIY project related to kitchen renovation. Note that the kitchen renovation motivation is only active for three purchase trips in a row in June 2013. In contrast, the gardening motivations are more persistent across the shopping trips.

Similarly, customer 211 in Figure 4 is interested in gardening as well, but has different needs compared to customer 144. This is reflected by the activation of different gardening motivations. Customer 211 seems to be more focused on landscaping and is not interested in DIY projects.

On the other hand, motivations related to Christmas, cleaning, and rust preventative spray paints are relevant to this customer as well.

The journeys of these two customers also highlight the necessity for modeling purchase behavior at the shopping trip level, instead of the customer level. Each customer could be described by some general customer profile, or a mix of those, for example a “gardener” or a “do-it-yourselfer”. However, within these profiles there is substantial variation across shopping trips and between customers. By modeling purchase behavior at the shopping trip level, we are able to capture this variation, which in turn enables us to find more nuanced motivations that describe a customer’s profile more accurately.

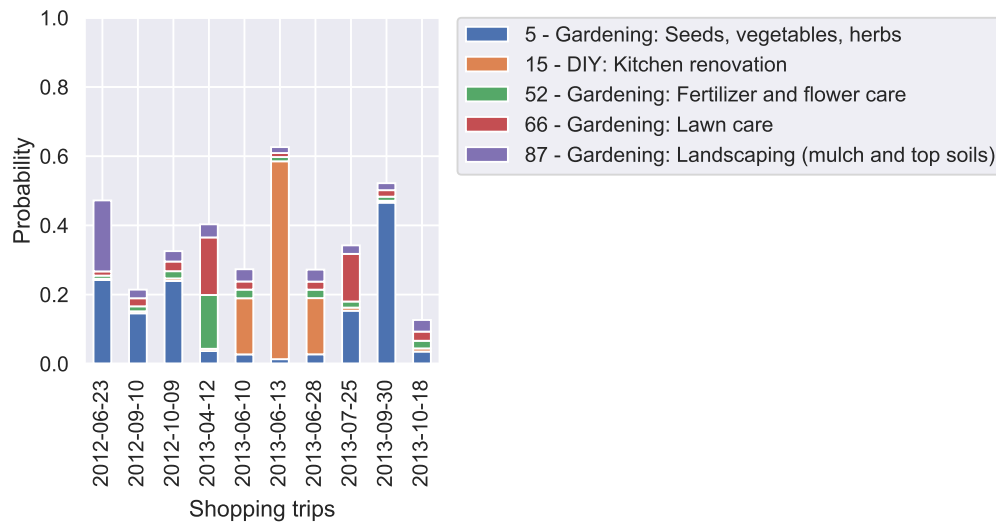


Figure 3: Journey for customer with ID 144.

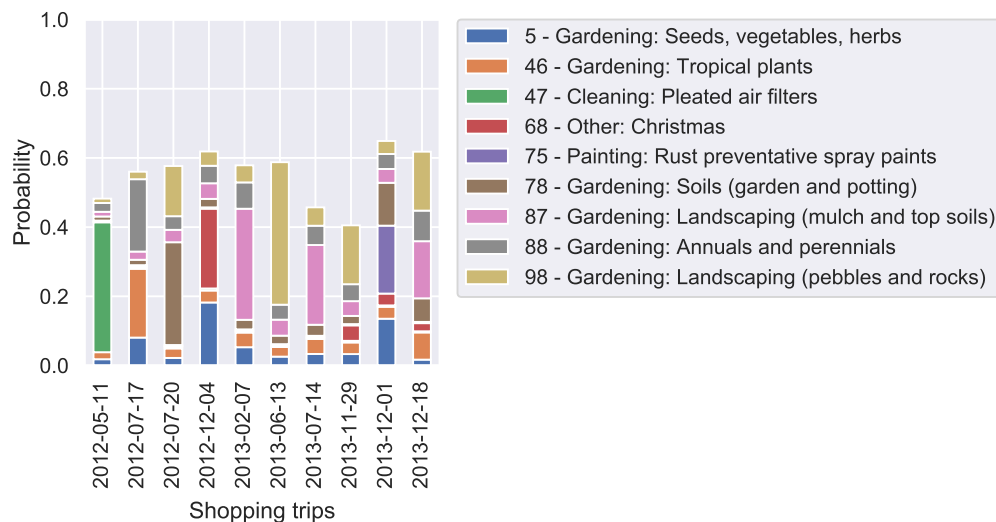


Figure 4: Journey for customer with ID 211.

5.3. Effects of explanatory variables

The relevance of motivation m in shopping trip b of customer i is, among other things, a function of explanatory variables specific to the shopping trip, \mathbf{x}_{ib} , and the observed customer characteristics, \mathbf{w}_i . In our application, the variables in \mathbf{x}_{ib} relate to the timing of the trip, e.g. month, weekday vs. weekend, daytime vs. evening. The available customer characteristics \mathbf{w}_i are age brackets, gender, and a proxy for household size. For these explanatory variables we measure the effect on motivation relevance by considering odds ratios as defined in Section 3.4, where the baseline is the average shopping trip.

We report descriptive statistics across the different motivations for the odds ratios related to \mathbf{x}_{ib} in Table 7. These statistics show that the shopping-trip specific variables have substantial effects on motivation relevance. For example, the weekend dummy increases the likelihood for motivation 79 by 26.6% (odds ratio 1.266). Further inspection shows that this motivation relates to grills and patio furniture. On the other hand, motivation 42, which is about key and lock replacements, becomes 17.8% less likely during the weekend (odds ratio 0.822). At the same time, this motivation has a large positive shift of 23.9% for shopping trips after work hours, which seems intuitively plausible. The largest odds ratios are obtained for November (29.449) and December (34.194), both for motivation 68. These extreme odds ratios make sense, as motivation 68 is about Christmas and holiday decorations.

Characteristic	Mean	Min	5% perc.	Median	95% perc.	Max
Year 1	1.002	0.791	0.883	0.994	1.115	1.319
Year 2	1.000	0.792	0.911	1.002	1.106	1.212
January	0.999	0.383	0.598	1.022	1.377	1.880
February	1.001	0.480	0.706	1.007	1.295	1.659
March	1.009	0.533	0.723	0.978	1.490	1.968
April	1.021	0.298	0.690	0.923	1.749	2.081
May	1.011	0.357	0.715	0.947	1.459	1.770
June	1.008	0.319	0.736	0.997	1.403	1.864
July	1.011	0.286	0.703	0.988	1.518	1.912
August	0.994	0.321	0.683	0.960	1.393	1.818
September	1.001	0.345	0.718	1.011	1.238	1.518
October	1.005	0.650	0.752	1.000	1.263	1.614
November	1.230	0.571	0.625	0.940	1.260	29.449
December	1.274	0.405	0.588	0.864	1.523	34.194
Weekday	1.000	0.878	0.905	1.010	1.085	1.125
Weekend	1.003	0.799	0.855	0.977	1.196	1.266
Daytime (before 5pm)	1.001	0.902	0.939	0.999	1.063	1.089
Evening (after 5pm)	1.001	0.723	0.789	0.995	1.250	1.451

Table 7: Descriptive statistics of the posterior means of the odds ratios for each shopping trip characteristic. The statistics are computed across the M motivations.

A similar analysis can be made for the customer characteristics in \mathbf{w}_i using Table 8. Again we see a substantial spread of effect sizes for the different characteristics, indicating that demographics are informative of motivation relevance. For example, for customers that are 65 years or older, the largest positive shift in motivation relevance is an 74.7% increase for a gardening motivation that involves fertilizers and flower care ($m = 52$). An interesting find is that large odds ratios are obtained for the dummy variables that correspond to missing values for a customer’s age and gender. This shows that customers who do not provide this information, differ a lot from those who do.

In Figure 5 we zoom in on the odds ratios corresponding to a specific customer characteristic.

Characteristic	Mean	Min	5% perc.	Median	95% perc.	Max
Gender: Female	0.997	0.660	0.820	1.004	1.222	1.334
Gender: Male	1.003	0.804	0.883	0.999	1.150	1.236
Gender: Unknown	1.038	0.416	0.572	0.991	1.677	2.108
Age [1, 35)	1.005	0.517	0.707	0.993	1.410	1.836
Age [35, 45)	1.007	0.662	0.781	0.980	1.288	1.490
Age [45, 55)	1.009	0.650	0.822	1.004	1.207	1.458
Age [55, 65)	1.003	0.717	0.821	0.998	1.184	1.283
Age [65, 93)	1.011	0.597	0.701	1.015	1.405	1.747
Age: Unknown	0.990	0.391	0.615	0.993	1.457	1.608
Household: Small	0.999	0.763	0.855	0.996	1.131	1.242
Household: Large	1.002	0.857	0.915	1.000	1.114	1.207

Table 8: Descriptive statistics of the posterior means of the odds ratios for each customer characteristic. The statistics are computed across the M motivations.

The odds ratios for the male dummy variable show that some of the technical motivations related to DIY and plumbing are more relevant for males. This approach of finding (ir)relevant motivations for a given characteristic, can be applied to all explanatory variables. In the same vein the model results could be used to determine the likely motivations for a specific demographic profile, e.g. female customers in the age of 35–45 living in a large household. Such data-driven insights can be useful to further understand the customer base and to help better target specific groups of customers. In addition, the demographic profile could be extended by considering a particular time frame, for example a weekend in October. Such model-based odds ratios can be integrated into model-based dashboards that support marketing managers in identifying relevant motivations for targeted customer segments.

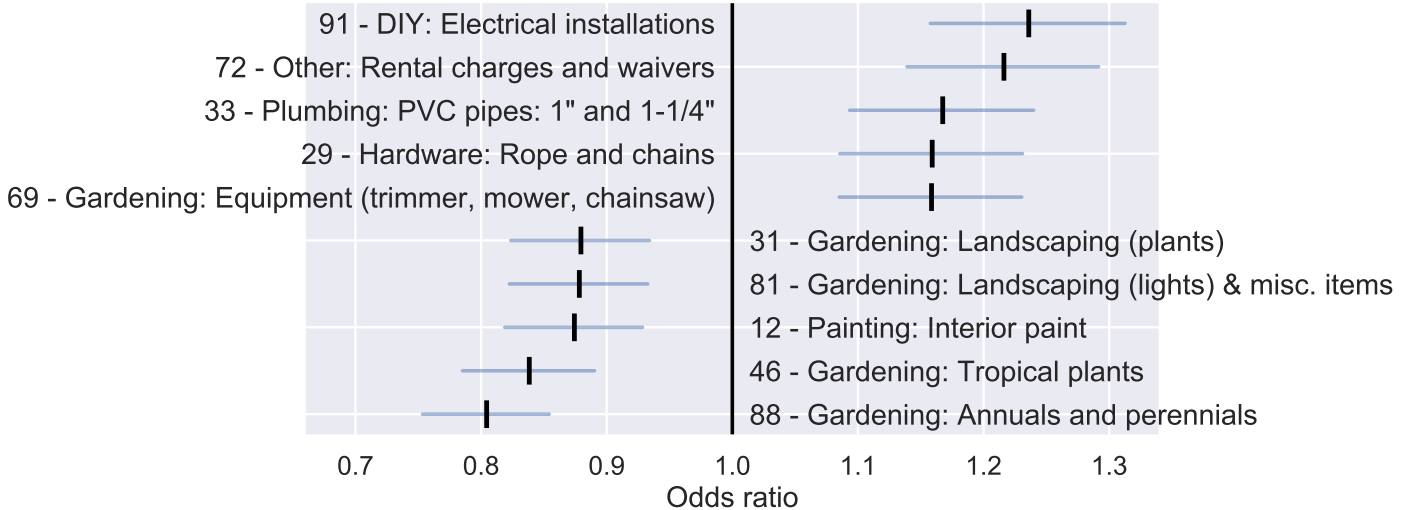
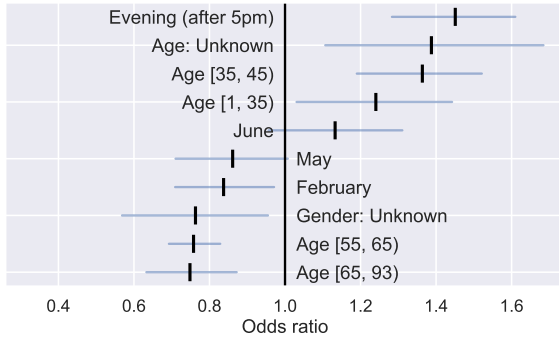


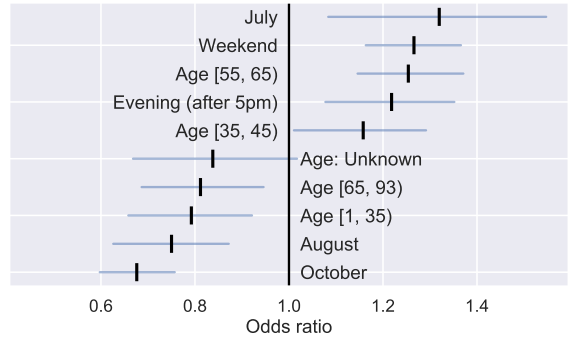
Figure 5: Odds ratios corresponding to the “Gender: Male” customer characteristic. The posterior mean and 95% highest posterior density interval is displayed for the motivations with the 5 largest and 5 smallest odds ratios for this variable.

Up to this point we have provided interpretations from the perspective of the explanatory variables describing a customer and shopping trip, in order to identify the most relevant motivations. Turning this idea around makes a motivation the unit of analysis. This supports managers in identifying the likely demographics and time periods for a motivation of interest. There are two advantages to this approach. First, this enables marketers to identify the segments in the customer base for which the focal motivation is relevant, and to identify the time period during

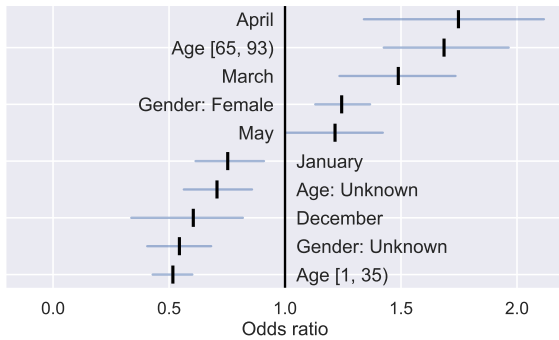
which its relevance peaks. Second, the explanatory variables can aid in the task of describing and interpreting a motivation, in addition to the most likely products for a motivation. This is especially useful when no detailed product taxonomy is available.



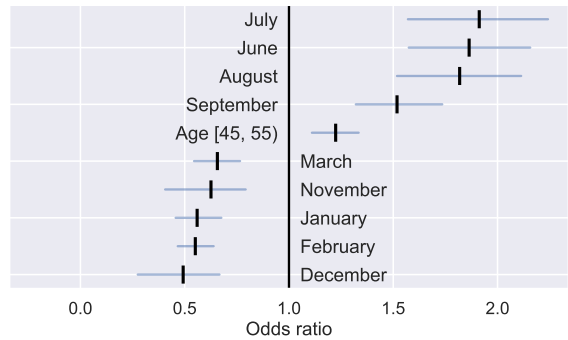
(a) Motivation 50 - Food products



(b) Motivation 79 - Grills and patio furniture



(c) Motivation 17 - Gardening: Mulch and soils



(d) Motivation 69 - Gardening: Large equipment

Figure 6: Odds ratios corresponding to the explanatory variables for four selected motivations. The posterior mean and 95% highest posterior density interval is displayed for the 5 largest and 5 smallest odds ratios.

In Figure 6 we display the odds ratios of a set of explanatory variables for four selected motivations. Motivation 50 in Figure 6a is about food products, like sodas and snacks. We observe a large positive shift for this motivation for shopping trips that take place during the evening and for younger customers. Motivation 79 is about grills and patio furniture and has a strong seasonality component where the motivation is more likely in July and during weekend shopping trips, see Figure 6b. Motivations 17 and 69 in Figures 6c and 6d both relate to gardening. The seasonality patterns however, are different. Motivation 17 is about soils and mulch and this motivation is more likely in the period from March – May and for older customers. On the other hand motivation 69, which is about landscaping equipment such as trimmers, mowers, and chainsaws, is more likely later in the year from June – September. This shows that even if one would have identified that a customer is interested in gardening, the relevance of specific motivations related to gardening is different across time and individuals. In contrast, if our model for purchase behavior would have aggregated shopping trips to the customer level, these separate motivations would have been aggregated into a single overarching gardening motivation and as a result, these seasonality effects would have been lost. By accounting for dynamics in shopping behavior, i.e. by not aggregating across the shopping trips of a customer in the model, we have enabled the identification of such sub-goals within a broader motivation such as gardening.

5.4. Relations between motivations

The model captures correlations across the relevance of motivations at the customer level through Σ_κ with corresponding correlation matrix $\text{Corr}(\Sigma_\kappa)$. Not only do these correlations lead to better predictive distributions (Blei and Lafferty 2007), they also provide additional insight in the relations between motivations. Such insights can help identify cross-selling opportunities, or assist in targeting customers that have not yet engaged in a motivation that is likely for them, based on their activities in (cor)related motivations.

The results reveal strong positive correlations of up to 0.844 for a pair of motivations that are both related to gardening ($m = 17$ and $m = 88$). The most negative correlation is -0.758 between again motivation 17 and motivation 36, which is about drywall materials. A histogram of the elements in the lower-triangular part of the correlation matrix $\text{Corr}(\Sigma_\kappa)$ is displayed in Figure 7. A sizable portion of the correlations is relatively large, for example about 3.1% of the pairwise correlations exceeds 0.50 and about 3.3% falls below -0.50. This shows that substantial correlations exist between the motivations.

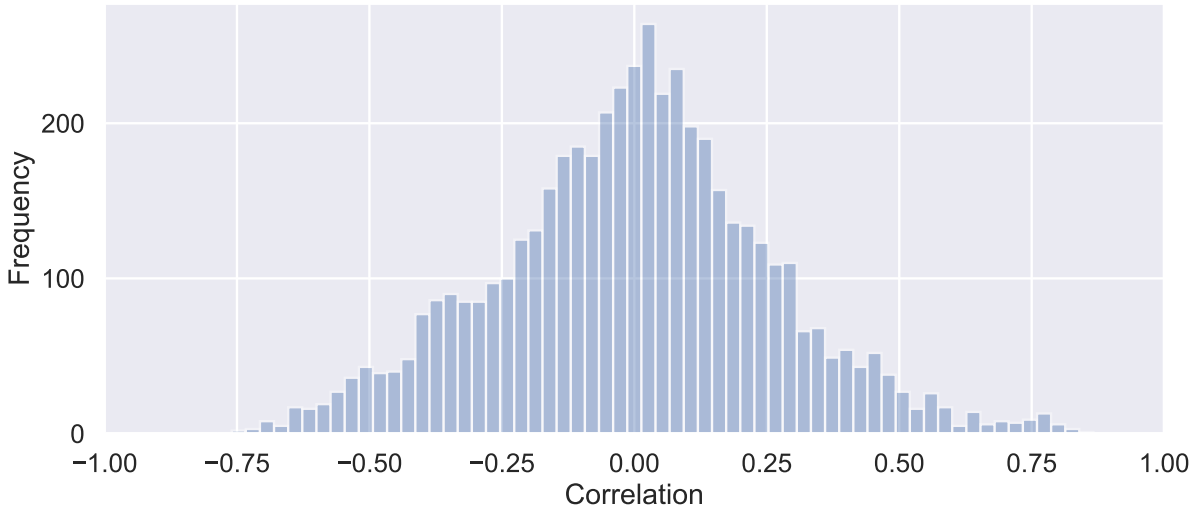


Figure 7: Histogram of the correlations in the lower-triangular part of $\text{Corr}(\Sigma_\kappa)$.

To interpret the correlation structure among motivations, one can zoom in on a focal motivation and analyze which motivations correlate with that motivation. As an example, Table 9 lists the motivations that correlate most strongly with motivation 23, which is about concrete mixes. Again the results show interesting insights that have high face validity. A motivation for concrete mixes positively correlates with certain do-it-yourself projects and hardware, but negatively correlates with motivations related to cleaning products.

m	Motivation label	Correlation
51	DIY: Deck and fence installation	0.579
49	DIY: Lumber (studs, strips, sheathing)	0.564
54	Plumbing: PVC pipes: 1-1/2" and 2"	0.559
⋮	⋮	⋮
1	Cleaning: Chemicals	-0.424
19	Cleaning: Floors	-0.428
30	Cleaning: General cleaning supplies	-0.538

Table 9: Most extreme correlations for motivation 23 (“concrete mixes”).

The previous analysis examines a one-to-many relationship for each motivation in isolation. An alternative approach is to directly visualize the relationships between all M motivations as captured by the correlation matrix. Multidimensional scaling (MDS) is a technique that maps objects to a lower-dimensional space based on their pairwise dissimilarities. The correlation matrix corresponding to Σ_κ can be converted to an $M \times M$ dissimilarity matrix D with elements: $D_{mm'} = \sqrt{2 - 2\text{Corr}(\Sigma_\kappa)_{mm'}}$ (cf. Borg and Groenen (2005)). With dissimilarity defined in this way, motivations that have a strong positive correlation will be placed close together in the MDS space, while motivations with small or negative correlations will be positioned further apart.

Figure 8 displays a two-dimensional MDS solution based on the dissimilarity matrix D . Most of the motivations can be roughly categorized in broad overarching themes such as “Gardening”, “Plumbing”, or “Painting”. Each of the six subfigures in Figure 8 accentuates such a theme. This reveals that most of the themes are concentrated in different regions in the graph, suggesting positive correlations between the corresponding motivations. For example, the gardening motivations can be found at the top-left of the MDS plot (Figure 8a), while the south/south-east covers the more technical do-it-yourself projects (Figure 8b), e.g. those that involve electrical installations, drywalls, and woodworking.

Note that “Plumbing” and “Hardware” overlap in location with the “DIY projects”. This could be expected, as plumbing can be seen as a special case of a do-it-yourself project, and hardware is needed to complete a DIY project. The two DIY motivations that are close to the gardening space are about the installation of an irrigation sprinkler system ($m = 48$) and the installation of screens and rodent control measures ($m = 90$). Indeed, such activities are expected to be related to gardening. The outlying DIY project ($m = 94$) relates to small bathroom repairs to the sink, drain, or toilet. Such a small-scale project is indeed less similar to the other more involved and technical DIY projects.

An interesting result is shown in Figure 8e for the motivations related to painting. All but one of these motivations are placed close together. Inspecting the outlying motivation 7 more closely reveals that it is a paint motivation dealing with exterior painting and waterproofing structures. Such a project differs in many aspects from an ordinary paint job like painting an indoor wall, explaining its distinctive position.

In summary, if one has a particular motivation in mind it can suffice to analyze the motivations that strongly correlate with that motivation. On the other hand, if one is interested in gaining insights about groups of motivations and how these relate, a visualization method such as MDS can be more appropriate as it reduces the high-dimensionality of the pairwise correlations to an interpretable graphical representation. The common denominator is that these results have high face validity and are mostly according to intuition. Even in case the position of a motivation is surprising at first, it typically can be placed in context by examining the motivation in more detail, as we have illustrated above. However, this does not mean that one would be able to identify and quantify these relations without the aid of our model.

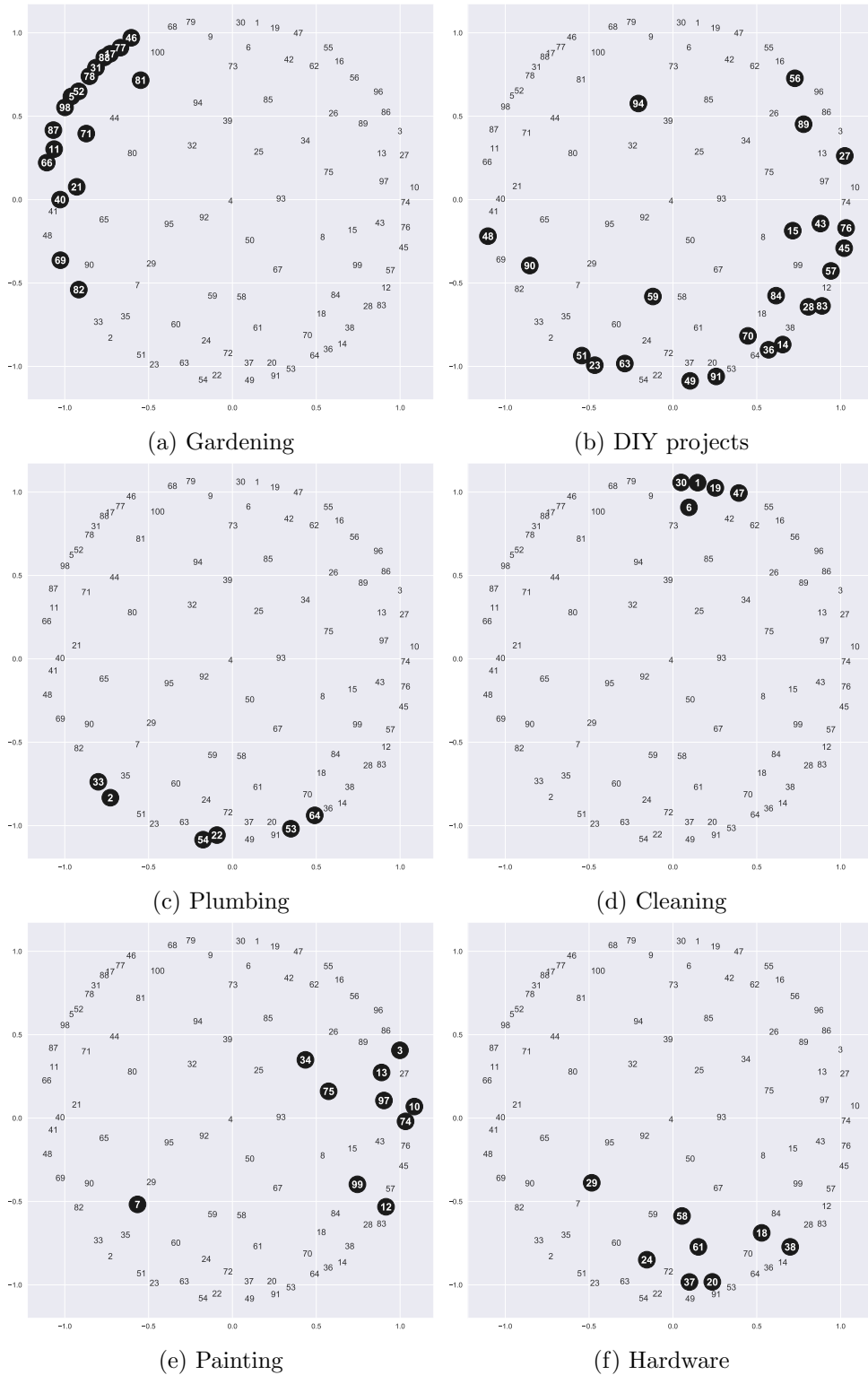


Figure 8: Two-dimensional MDS solution based on the motivation correlation matrix $\text{Corr}(\Sigma_{\kappa})$. Each subplot accentuates a different group of motivations from the same MDS solution.

In addition to the covariance matrix Σ_{κ} , relations between motivations are modeled by the VAR(1)-effects in the specification of motivation relevance α_{ib} . The VAR(1)-coefficient vector ρ_m gives the effect of the lagged motivation relevance vector (α_{ib-1}) on the relevance of motivation m : α_{ibm} . Similar to the coefficients for the other explanatory variables, the impact of the VAR(1)-coefficients can be assessed using the odds ratio as defined in Equation (11). The effect for a single lagged motivation m on the odds ratios of all M motivation activation probabilities is measured by increasing the lagged relevance of m ($\alpha_{i,b-1,m}$) by two standard deviations ($2\sigma_{\alpha_m}$), ceteris paribus. We can split the resulting odds ratios in an “own” AR-effect, for motivation m on itself, and “cross” AR-effects, for motivation m on the other $M - 1$ motivations.

The posterior mean and 95% highest posterior density interval for each own AR-effect is displayed in Figure 9. The posterior means range from 1.019 to 1.186, with an average of 1.057. We note that each of these effects is larger than 1, indicating at least some persistence for all motivations. A number of motivations have a relatively large own AR-effect. Examples are motivation 10 (effect 1.186), related to painting supplies, motivation 83 (effect 1.175), related to tile setting, and motivation 91 (effect 1.144) which is related to electrical installations. These motivations seem to correspond to large or persistent projects, as they are more likely to be activated across consecutive shopping trips. Other motivations have smaller own AR-effects, suggesting that these motivations are less persistent.

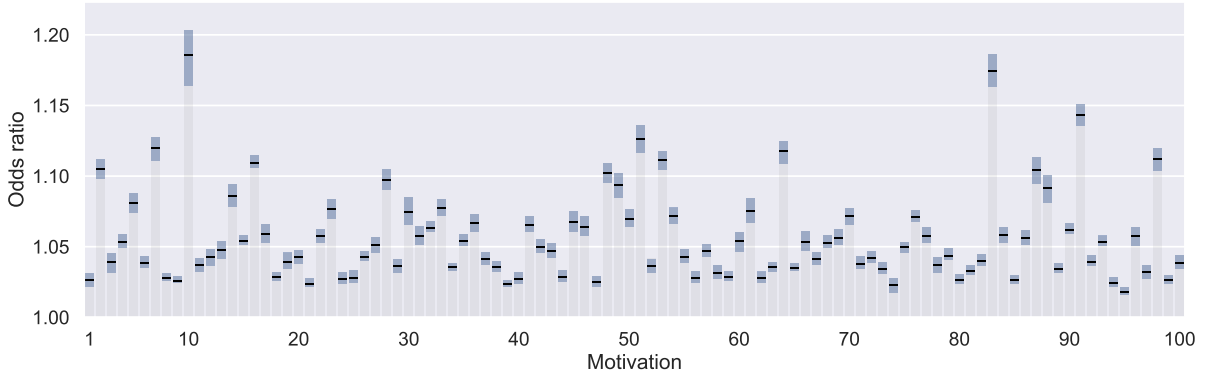


Figure 9: Posterior means and 95% highest posterior density interval for the odds ratios of the “own” AR-effects of the motivations.

There are a few cross AR-effects that have a reasonable impact: the largest odds ratio of 1.254 is from motivation 60 on 72. Motivation 72 relates to rental charges and waivers, while the most likely product in motivation 60 is an e-deposit product with purchase probability 0.499. It is very plausible that if motivation 60 is activated it becomes more likely that motivation 72 will become active in the next shopping trip. Note that this does not hold the other way around, as the asymmetric cross AR-effect from motivation 72 on 60 is much closer to unity (1.035). The next largest cross AR-effect that we find is much smaller, 1.048, and relates to motivation 2 (Plumbing - PVC pipes: 1/2” and 3/4”) and motivation 48 (installation of irrigation systems). Again, it seems reasonable that these two motivations are related. Furthermore, this relationship is more symmetric as the cross AR-effect from motivation 48 on motivation 2 is 1.045.

Even though there is some variation among the cross AR-effects, the majority of the odds ratios is around unity. Hence, in our application we do not find many cases where one motivation starkly increases – or decreases – the probability of another motivation at the next shopping trip. A potential explanation is that in our empirical application, relationships between motivations mainly exist at the customer level, and not over time within a customer. Such relationships are absorbed in the heterogeneity covariance matrix Σ_{κ} . In other empirical applications, however, path dependence in motivation activation could be more important.

5.5. Predictive performance

To determine the predictive performance of the model we use the hold-out data set that contains the last shopping trip for every customer that has visited the store more than once. This data set is denoted by \mathbf{y}^{OOS} . We measure predictive performance of a model on this hold-out data by evaluating $\log p(\mathbf{y}^{\text{OOS}}|\mathbf{y})$: the log predictive likelihood of \mathbf{y}^{OOS} conditional on the estimation data \mathbf{y} .

We contrast the predictive performance of our model against several competing methods. First, to better assess the added value of the VAR(1) effects, we consider our own model without these effects. Second, we consider a version of the LDA-X model (Jacobs et al. 2016) that is nested within our model, see Section 2.4, which aggregates over trips and therefore does not include any trip-specific information. It also ignores correlations between the motivations. Finally, to place the performance of these models into perspective, we consider two model-free benchmarks without customer heterogeneity: a marginal probability vector where the probabilities are based on the relative purchase frequencies in the estimation data, and a naive flat probability vector where each product has a purchase probability of J^{-1} .

The predictive performance results are displayed in Table 10. Our model clearly outperforms the alternatives. The LDA-X model is outperformed by a margin of over 1,000 log-likelihood points. A large driver of this performance increase are the time effects and motivation correlations that result in a more realistic model. Furthermore, the VAR(1) effects in our model provide a positive contribution to the model’s predictive performance as well, which validates their inclusion in the model. In addition, the results clearly show that all models outperform the model-free benchmarks.

Model	\mathbf{w}_i	\mathbf{x}_{ib}	Correlations	VAR(1)	$\log p(\mathbf{y}^{\text{OOS}} \mathbf{y})$
Complete model	x	x	x	x	-40,517.69
Model without VAR effects	x	x	x	–	-40,541.65
LDA-X	x	–	–	–	-41,545.51
Marginal probability vector	–	–	–	–	-43,814.22
Flat probability vector	–	–	–	–	-46,581.54

Table 10: The log predictive likelihood of the hold-out data for different methods.

6. Managerial implications

One of the key strengths of our model is that it identifies purchase motivations from purchase history data alone. These motivations provide a novel view on purchase behavior, exposing the salient purchase patterns and how these differ across customers and shopping trips. This is especially useful for retailers with large product assortments, in which such purchase patterns can be obfuscated due to the volume of purchases and variety of the product assortment. In such cases, considering customer behavior expressed through purchase motivations is simpler, more intuitive, and may lead to better insights. In this section we discuss how the model results, and in particular the motivations, can serve as input to managerial decisions.

Managerial dashboards: all the insights we generated in the previous section are helpful in managerial decision making. The model results provide company experts detailed, quantified knowledge on the customer base and product assortment. To facilitate the discovery of relevant insights one could easily build a managerial model-based dashboard around the model results. The focal perspective of such a dashboard may be based on a product, a motivation, a time

period, or even an individual customer. A dashboard for a specific product may contain a ranking of motivations, e.g. displaying the five most likely motivations for that product.

A dashboard can also be constructed for a specific motivation. Naturally, the most likely products under that motivation can be displayed, which allows for the creation of intuitive labels for the motivations. In addition, the motivations that correlate most strongly with the focal motivation can be listed. Yet another insight in a specific motivation is given by seasonality and the effects of the explanatory variables such as customer age and gender. Such effects provide a detailed picture on when – and for whom – a motivation is most likely active. Conversely, it is possible to list the most relevant motivations for a given customer segment at a certain point in time.

Explainable product recommendations: the model structure can serve as the basis for explainable product recommendations at a retailer. At the highest level, an explanation for the recommended products can be provided through a link with the activated motivations that drive the product recommendation. At a more detailed level, the model output can be used to explain the predicted relevance of the motivations for the customer’s shopping trip. Specifically, a selection of the most important explanatory variables such as a customer’s innate preferences, demographics, the context of the shopping trip, or the contents of the previous shopping trip, can be used to explain why these motivations are predicted to be active. This provides a detailed, yet understandable explanation why certain motivations are active for a customer and hence, why certain products are expected to be relevant for that customer. Common recommendation systems, such as matrix factorization techniques, cannot provide such insights.

Targeted advertising: the inferred motivations also provide opportunities for targeted advertising and the specific timing of such actions. Targeted advertising can even be applied within the store. As an example of such a real-time action, consider a “smart shopping cart”. This shopping cart may contain a device that registers when a customer places a product in her shopping cart. The relevant motivations corresponding to this product can be identified and this information can be directly communicated to the customer, e.g. on the shopping cart or on their mobile phone.

In the context of a hardware store, such information can be an instructional video that explains how to complete a do-it-yourself project corresponding to the motivation. More generally, a list of relevant products, based on the motivation, or a promotion tailored to the motivation can be shown. Opportunities for personalized communication continue after a shopping trip has ended. For example, consider timely follow-up (e)mails that promote motivations that have become more relevant after a shopping trip. Such motivations can be determined using the inferred motivation-correlations and the VAR(1)-effects.

Knowing which motivations are (ir)relevant for a customer can also improve communication with the customer. If a customer is likely to engage in do-it-yourself home improvement projects but has shown little interest in gardening motivations, marketing campaigns directed at this customer can be adjusted accordingly. Reaching out to customers at the level of their motivations is preferred over communication at the product level for at least two reasons. First, given the large size of the product assortment, it is quite unlikely to accurately predict the specific product the customer wants to buy. Predicting the correct motivation is much more realistic and hence this might be more effective in engaging the customer. Second, communicating at a higher level in terms of the purchase motivations – instead of at the detailed product level – is more appreciated by customers, especially when the communication is not immediately linked to a purchase decision (da Costa Hernandez et al. 2015).

Improving store layout: insights about the relevant motivations can be used to improve store layouts, both for online and offline stores. First consider a brick-and-mortar store: small

products that are relevant across multiple motivations can be positioned strategically at multiple places within the store to improve the shopping experience. As an example, work gloves are relevant for gardening activities, but also for renovation jobs. By placing the work gloves in both locations in the store, they become easier to locate for customers. Additionally, the correlations between motivations could be used to position product categories in a store relative to each other. For example, if certain PVC pipes are most often used in a bathroom renovation, it may make sense to position those PVC pipes relatively close to products such as toilets, bath tubs, and showers.

Similarly, inferred motivations can be used in an online store. For example, for each motivation a landing page can be created, containing both products *and* other motivations which are related to the focal motivation. As an example, consider a landing page specific to a barbecue motivation. Not only would the relevant products be displayed, such as grills and utensils, but relevant motivations could be linked as well, such as motivations related to insect repellents or an outdoor pool. Our model is able to identify the most related motivations. In a similar vein, a product page can refer to motivations under which the product is likely to be purchased. As one of these motivations is likely to be relevant, linking to that motivation can create promising cross-selling opportunities.

Another approach is to link (sponsored) search queries to motivations. For each search query that does not exactly match a specific product, one can identify the best matching motivation. A landing page related to a motivation will likely outperform a landing page that only lists a single product. In this way, one would directly connect the customer to the full set of products the customer is interested in, which again provides clear opportunities for cross-selling.

There are other benefits to identifying purchase motivations that are not directly connected to marketing actions. As an example, consider inventory management. A promotion for a certain product may create spillover effects to other products. Using motivations, these spillover effects can be anticipated. This example, and those presented above, all seem to benefit from a succinct representation of purchase behavior, expressed using a small set of motivations. We consider the actual application of motivation-assisted marketing decision making a very exciting and fruitful avenue for further research.

7. Conclusions and further research

In this paper we have introduced a novel model that efficiently analyzes large-scale purchase history data, consisting of product purchase decisions made by many customers – in an even larger number of shopping trips – out of a product assortment consisting of thousands of products. Using the model we extract insights relevant for marketing decision makers and help determine the salient purchase patterns to improve understanding of purchase behavior in the customer base. The fundamental idea is to represent the high-dimensional purchase data in a lower-dimensional space, spanned by latent purchase motivations. The motivations are incorporated in the model as probability distributions over the product assortment, which are identified using only observational purchase history data – data that is ubiquitous in modern retail applications. The model results can be used to formalize and quantify how probabilities for motivation activation and product purchases are linked to and affected by explanatory variables, such as seasonality or customer demographics.

The model is designed such that it scales well to realistic retail settings, ensuring that it can be used to generate valuable insights in practice. We have achieved this in several ways. First, the model is built on the class of topic models, which is known to scale well to large applications. Second, to estimate the model parameters we rely on an estimation technique from the machine

learning literature called variational inference. This technique has proven successful for estimating large and complex models in a reasonable time frame. Third, we extend results from the variational inference literature, by deriving a novel algorithm that can be used to efficiently estimate a large motivation-covariance matrix for applications with many motivations.

In the current model we only rely on off-the-shelf purchase history data to identify the purchase motivation structure. While this modeling decision makes the model easy to adopt in practice, it also has some limitations. Purchases made by customers are partially affected by factors under the control of a retailer, e.g. product prices, promotion strategies, and store layout. Controlling for these factors in the inference of the motivations is a promising avenue for further research.

Another direction for further research are product recommendations based on the inferred motivations. In this paper we have shown that our model is able to accurately predict purchases in the next shopping trip. A next step would be to incorporate these predictions in a product recommendation system. The benefit of such a recommendation system is that it leads to product recommendations that can be explained in terms of motivations and explanatory variables. However, a product that is likely to be purchased by a customer is not necessarily an effective recommendation (Bodapati 2008). The effectiveness of recommendations based on purchase motivation is an empirical question.

Developing and implementing marketing actions based on the inferred motivations provides another direction for further research. In this paper we have shown that our model is able to capture meaningful motivations at the customer level and across the shopping trips of a single customer. The inferred motivation structure also results in accurate predictions of a customer’s purchases in the next shopping trip. Future research could implement and study the effectiveness of marketing actions that build on the insights generated by our model, such as landing pages tailored to the most likely motivation, or a recommendation system that is based on accurate, explainable predictions. In sum, we see our model as an important tool for the analysis of purchase behavior in high-dimensional retail settings and as a stepping stone for motivation-based marketing actions at retailers with large assortments.

A. Additional model details

This appendix describes additional model details: a specification of the motivation relevance for the first shopping trip and the prior distributions for all population-level parameters in the model.

The motivation relevance for shopping trip b made by customer i , α_{ib} , is modeled as a function of the previous shopping trip: α_{ib-1} , cf. Equation (4). This lagged value is not available if b refers to the first (observed) shopping trip of a customer. For these shopping trips we specify an adjusted model for the expected motivation relevance:

$$\mu_{i1m} = \delta_m + \delta_\kappa \kappa_{im} + \delta_\beta \mathbf{x}_{i1}^\top \boldsymbol{\beta}_m + \delta_\gamma \mathbf{w}_i^\top \boldsymbol{\gamma}_m, \quad (\text{A.1})$$

where δ_m is a motivation-specific intercept. In addition, δ_κ , δ_β and δ_γ are scalar factors that allow for shifts in effect sizes that are specific to the first shopping trip of a customer. Note that as these three parameters reflect proportional shifts, their priors are centered around unity instead of zero.

For the specification of the prior, we transform variance parameters to their inverse, i.e. precision parameters. More specifically, the variance for ϵ_{ibm} , $\sigma_{\alpha_m}^2$, is transformed to a precision $\tau_{\alpha_m} \equiv \sigma_{\alpha_m}^{-2}$, with $\boldsymbol{\tau}_\alpha = [\tau_{\alpha_1}, \dots, \tau_{\alpha_M}]$. The covariance matrix for $\boldsymbol{\kappa}_i$, $\boldsymbol{\Sigma}_\kappa$, is transformed to a precision matrix $\boldsymbol{\Lambda}_\kappa \equiv \boldsymbol{\Sigma}_\kappa^{-1}$. For each population parameter the prior is chosen such that it is conjugate to the parameter's full-conditional distribution. The parameters of the prior distributions are set such that they represent relatively uninformative distributions. Note that large parameters inside a softmax function tend to result in implausible outcomes, see Table 1 in Pachali et al. (2020) for a related conclusion, so a prior variance of 1 is already substantial. \mathbf{I} denotes an identity matrix, and $\mathbf{0}$ and $\mathbf{1}$ are vectors of zeros and ones respectively. The priors for the population-level parameters are:

For $m = 1, \dots, M$:

$$\boldsymbol{\phi}_m \sim \text{Dirichlet}_J(\boldsymbol{\alpha} = \mathbf{1}J^{-1})$$

$$\boldsymbol{\rho}_m \sim \text{MVN}_M(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$$

$$\boldsymbol{\beta}_m \sim \text{MVN}_{K_X}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$$

$$\boldsymbol{\gamma}_m \sim \text{MVN}_{K_W}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$$

$$\delta_m \sim \text{Normal}(\mu = 0, \sigma^2 = 1)$$

$$\tau_{\alpha_m} \sim \text{Gamma}(\alpha = 1, \beta = 1)$$

$$\delta_\kappa, \delta_\beta, \delta_\gamma \sim \text{Normal}(\mu = 1, \sigma^2 = 1)$$

$$\boldsymbol{\mu}_\kappa \sim \text{MVN}_M(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$$

$$\boldsymbol{\Lambda}_\kappa \sim \text{Wishart}_M(n = 2M, \mathbf{V} = \mathbf{I}(2M)^{-1})$$

B. Pseudocode for estimation routine

This appendix contains the pseudocode for our estimation routine that uses variational inference (VI) to estimate the unknown model components.

Algorithm 1 Pseudocode for VI coordinate ascent optimization algorithm

```

1: Initialize  $q(\Omega)$ 
2: while ELBO has not converged do
3:   for  $i = 1, \dots, I$  do
4:     for  $\ell = 1, \dots, L$  do
5:       for  $b = 1, \dots, N_{ib}$  do
6:         Update  $q(\mathbf{z}_{ib}), q(\alpha_{ib})$ 
7:       end for
8:       Update  $q(\kappa_i)$ 
9:     end for
10:  end for
11:  Update  $q(\phi), q(\beta), q(\gamma), q(\rho), q(\delta), q(\delta_\kappa), q(\delta_\beta), q(\delta_\gamma), q(\tau_\alpha), q(\mu_\kappa), q(\Lambda_\kappa)$ 
12: end while

```

Algorithm 2 Pseudocode for variational update for $q(\alpha_{ib})$

```

1: Input:  $\tilde{\mu}_{ib}, \tilde{\sigma}_{ib}^2, ss_{ib}^{\tilde{\mu}}, ss_{ib}^{\log \tilde{\sigma}}, ss_{\text{factor}}, ss_{\text{min}}, ss_{\text{max}}, q(\text{MB}_{\alpha_{ib}})$ 
2: Compute current ELBO contribution:  $\text{ELBO}_{ib} = \text{ELBO}(\tilde{\mu}_{ib}, \tilde{\sigma}_{ib}^2, q(\text{MB}_{\alpha_{ib}}))$ 
3: Create candidate for  $\tilde{\mu}_{ib}$ :  $\tilde{\mu}_{ib}^\dagger = \tilde{\mu}_{ib} + ss_{ib}^{\tilde{\mu}} \nabla_{\tilde{\mu}_{ib}} \text{ELBO}(\tilde{\mu}_{ib}, \tilde{\sigma}_{ib}^2, q(\text{MB}_{\alpha_{ib}}))$ 
4: Compute candidate ELBO contribution:  $\text{ELBO}_{ib}^\dagger = \text{ELBO}(\tilde{\mu}_{ib}^\dagger, \tilde{\sigma}_{ib}^2, q(\text{MB}_{\alpha_{ib}}))$ 
5: if  $\text{ELBO}_{ib}^\dagger > \text{ELBO}_{ib}$  then
6:    $\tilde{\mu}_{ib} = \tilde{\mu}_{ib}^\dagger$  and  $\text{ELBO}_{ib} = \text{ELBO}_{ib}^\dagger$ 
7:   Increase  $ss_{ib}^{\tilde{\mu}}$ :  $ss_{ib}^{\tilde{\mu}} = \min(ss_{\text{factor}} ss_{ib}^{\tilde{\mu}}, ss_{\text{max}})$ 
8: else
9:   Decrease  $ss_{ib}^{\tilde{\mu}}$ :  $ss_{ib}^{\tilde{\mu}} = \max(ss_{\text{factor}}^{-1} ss_{ib}^{\tilde{\mu}}, ss_{\text{min}})$ 
10: end if
11: Create candidate for  $\log \tilde{\sigma}_{ib}$ :  $\log \tilde{\sigma}_{ib}^\dagger = \log \tilde{\sigma}_{ib} + ss_{ib}^{\log \tilde{\sigma}} 2\tilde{\sigma}_{ib} \cdot \nabla_{\tilde{\sigma}_{ib}^2} \text{ELBO}(\tilde{\mu}_{ib}, \tilde{\sigma}_{ib}^2, q(\text{MB}_{\alpha_{ib}}))$ 
12: Create candidate for  $\tilde{\sigma}_{ib}^2$ :  $(\tilde{\sigma}_{ib}^2)^\dagger = \exp(2 \log \tilde{\sigma}_{ib}^\dagger)$ 
13: Compute candidate ELBO contribution:  $\text{ELBO}_{ib}^\dagger = \text{ELBO}(\tilde{\mu}_{ib}, (\tilde{\sigma}_{ib}^2)^\dagger, q(\text{MB}_{\alpha_{ib}}))$ 
14: if  $\text{ELBO}_{ib}^\dagger > \text{ELBO}_{ib}$  then
15:    $\tilde{\sigma}_{ib}^2 = (\tilde{\sigma}_{ib}^2)^\dagger$ 
16:   Increase  $ss_{ib}^{\log \tilde{\sigma}}$ :  $ss_{ib}^{\log \tilde{\sigma}} = \min(ss_{\text{factor}} ss_{ib}^{\log \tilde{\sigma}}, ss_{\text{max}})$ 
17: else
18:   Decrease  $ss_{ib}^{\log \tilde{\sigma}}$ :  $ss_{ib}^{\log \tilde{\sigma}} = \max(ss_{\text{factor}}^{-1} ss_{ib}^{\log \tilde{\sigma}}, ss_{\text{min}})$ 
19: end if
20: Output:  $\tilde{\mu}_{ib}, \tilde{\sigma}_{ib}^2, ss_{ib}^{\tilde{\mu}}, ss_{ib}^{\log \tilde{\sigma}}$ 

```

The objective of VI is to minimize the KL divergence from $q(\Omega)$ to $p(\Omega|\mathbf{y})$ with respect to the parameters of the variational distribution. This is equivalent to maximizing the corresponding ELBO. The pseudocode to maximize the ELBO of our model is given in Algorithm 1. After each update step in the algorithm, the ELBO is guaranteed to not decrease. Per iteration and for each customer i we take $L = 25$ subiterations to jointly optimize the variational distributions specific to customer i . The optimization is completed once the ELBO has converged.

The pseudocode for the update for $q(\alpha_{ib}) = \text{Normal}_M(\mu = \tilde{\mu}_{ib}, \sigma^2 = \tilde{\sigma}_{ib}^2)$, is given in Algorithm 2, where $q(\text{MB}_{\alpha_{ib}})$ refers to the variational distributions for the parameters in the Markov blanket¹⁰ of α_{ib} . The algorithm uses adaptive step sizes $ss_{ib}^{\tilde{\mu}}$ and $ss_{ib}^{\log \tilde{\sigma}}$. If a step in the direction

¹⁰The Markov blanket for a parameter is defined as its parents, children, and the co-parents of its children in the directed acyclical graph (DAG) representation of the model (Pearl 2009).

of the gradient results in an increase (decrease) of the ELBO, the step size is multiplied (divided) by ss_{factor} . Furthermore, the adaptive step sizes are bounded between ss_{min} and ss_{max} . In our implementation we take $ss_{\text{factor}} = 1.125$, $ss_{\text{min}} = 10^{-6}$, $ss_{\text{max}} = 1$.

Details for the update steps in Algorithms 1 and 2 are provided in Online Appendix D. The closed-form analytical solutions for the updates of all variational distributions except $q(\boldsymbol{\alpha}_{ib})$ are given in Appendix D.1. The ELBO function and gradients used in the update step for $q(\boldsymbol{\alpha}_{ib})$ are described in Appendix D.2. Details on the initialization of $q(\boldsymbol{\Omega})$ are provided in Appendix D.3.

C. Proof for efficient covariance matrix inverse and (log) determinant

Let \mathbf{M}_i be a $K \times K$ covariance matrix that is composed of two non-singular $K \times K$ precision matrices \mathbf{P} and \mathbf{C} , where \mathbf{C} is multiplied by a non-negative scalar s_i :

$$\mathbf{M}_i \equiv (\mathbf{P} + \mathbf{C}s_i)^{-1}. \quad (\text{C.1})$$

Both \mathbf{M}_i and its determinant can be computed without inverting a $K \times K$ matrix that depends on s_i .

Proof \mathbf{M}_i

Step 1: Compute a Cholesky decomposition of \mathbf{C} : $\mathbf{C} = \mathbf{L}\mathbf{L}^\top$ and rewrite \mathbf{M}_i as:

$$\mathbf{M}_i = \left(\mathbf{P} + \mathbf{L}\mathbf{L}^\top s_i \right)^{-1} = \left(\mathbf{L} \left(\mathbf{L}^{-1}\mathbf{P}(\mathbf{L}^{-1})^\top + \mathbf{I}s_i \right) \mathbf{L}^\top \right)^{-1}. \quad (\text{C.2})$$

Step 2: Compute a singular value decomposition of $\mathbf{L}^{-1}\mathbf{P}(\mathbf{L}^{-1})^\top$: $\mathbf{L}^{-1}\mathbf{P}(\mathbf{L}^{-1})^\top = \mathbf{U}d(\mathbf{v})\mathbf{U}^\top$, where $d(\mathbf{v})$ creates a diagonal matrix out of the vector \mathbf{v} . By design, $\mathbf{U}^\top = \mathbf{U}^{-1}$ and $\mathbf{U}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^{-1} = \mathbf{I}$. This enables us to further rewrite \mathbf{M}_i :

$$\begin{aligned} \mathbf{M}_i &= \left(\mathbf{L} \left(\mathbf{U}d(\mathbf{v})\mathbf{U}^\top + \mathbf{U}\mathbf{U}^\top s_i \right) \mathbf{L}^\top \right)^{-1} = \left(\mathbf{L}\mathbf{U}d(\mathbf{v} + s_i)\mathbf{U}^\top \mathbf{L}^\top \right)^{-1} \\ &= (\mathbf{L}^{-1})^\top \mathbf{U}d((\mathbf{v} + s_i)^{-1})\mathbf{U}^\top \mathbf{L}^{-1}. \end{aligned} \quad (\text{C.3})$$

■

Proof determinant

Using the general properties of the determinant and the identity that $\mathbf{U}\mathbf{U}^\top = \mathbf{I}$:

$$\begin{aligned} \det(\mathbf{M}_i) &= \det \left((\mathbf{L}^{-1})^\top \mathbf{U}d((\mathbf{v} + s_i)^{-1})\mathbf{U}^\top \mathbf{L}^{-1} \right) \\ &= \det \left((\mathbf{L}^{-1})^\top \right) \det(\mathbf{U}) \det \left(d((\mathbf{v} + s_i)^{-1}) \right) \det(\mathbf{U}^\top) \det(\mathbf{L}^{-1}) \\ &= \det(\mathbf{U}\mathbf{U}^\top) \det \left(\mathbf{L}^{-1}(\mathbf{L}^{-1})^\top \right) \det \left(d((\mathbf{v} + s_i)^{-1}) \right) \\ &= \det(\mathbf{C})^{-1} \prod_{k=1}^K (v_k + s_i)^{-1}. \end{aligned} \quad (\text{C.4})$$

The corresponding log determinant is given by:

$$\log \det(\mathbf{M}_i) = -\log \det(\mathbf{C}) - \sum_{k=1}^K \log(v_k + s_i). \quad (\text{C.5})$$

■

D. Online Appendix: Implementation details estimation routine

This appendix contains the implementation details for the variational inference coordinate ascent optimization algorithm, which is used to determine the parameters of the optimal variational distribution $q^*(\Omega)$.

The specification of $\epsilon_{ibm} \equiv \alpha_{ibm} - \mu_{ibm}$ and $\epsilon_{ib} \equiv [\epsilon_{ib1}, \dots, \epsilon_{ibM}]$ plays a central role in the variational updates for α_{ib} and the parameters in the Markov blanket for α_{ib} . The relevant variational expectations of ϵ_{ibm} are defined as:

$$\begin{aligned}\tilde{\mathbb{E}}\{\epsilon_{ibm}\} &\equiv \tilde{\mathbb{E}}\{\alpha_{ibm}\} - \tilde{\mathbb{E}}\{\mu_{ibm}\}, \\ \tilde{\mathbb{E}}\{\epsilon_{ibm}^2\} &\equiv \tilde{\mathbb{E}}\{\alpha_{ibm}^2\} + \tilde{\mathbb{E}}\{\mu_{ibm}^2\} - 2\tilde{\mathbb{E}}\{\alpha_{ibm}\}\tilde{\mathbb{E}}\{\mu_{ibm}\}.\end{aligned}\tag{D.1}$$

Based on the specification of μ_{ibm} in Equation (4) we obtain:

$$\tilde{\mathbb{E}}\{\mu_{ib}\} = \tilde{\mathbb{E}}\{\boldsymbol{\kappa}_i\} + \tilde{\mathbb{E}}\{\mathbf{R}\}\tilde{\mathbb{E}}\{\boldsymbol{\alpha}_{ib-1}\} + \tilde{\mathbb{E}}\{\mathbf{B}\}\mathbf{x}_{ib} + \tilde{\mathbb{E}}\{\mathbf{G}\}\mathbf{w}_i,\tag{D.2}$$

with $\mu_{ib} \equiv [\mu_{ib1}, \dots, \mu_{ibM}]$ and where the $\boldsymbol{\rho}_m$, $\boldsymbol{\beta}_m$, and $\boldsymbol{\gamma}_m$ vectors have been collected in the matrices \mathbf{R} , \mathbf{B} , and \mathbf{G} that are respectively of dimensions $M \times M$, $M \times K_X$, and $M \times K_W$. In other words, the m th row of each matrix corresponds to parameters $\boldsymbol{\rho}_m$, $\boldsymbol{\beta}_m$, and $\boldsymbol{\gamma}_m$, respectively.

We start by introducing notation to facilitate exposition of our results. We often need to compute α_{ib} cleaned from the explained effect of all but one of the factors in μ_{ib} . To illustrate this with an example, we can clean α_{ib} of the effect of everything except κ_i as follows: $\alpha_{ib} - \mathbf{R}\alpha_{ib-1} - \mathbf{B}\mathbf{x}_{ib} - \mathbf{G}\mathbf{w}_i = \alpha_{ib} - (\mu_{ib} - \kappa_i) = \epsilon_{ib} + \kappa_i \equiv \epsilon_{ib} \setminus (\kappa_i)$. From the left-hand-side of this equality we see that $\epsilon_{ib} \setminus (\kappa_i)$ is not a function of κ_i and is therefore constant with respect to $q(\kappa_i)$. In general, we will denote this ‘‘cleaned disturbance term’’ as $\epsilon_{ib} \setminus (\mathbf{c}_\vartheta \boldsymbol{\vartheta})$, where $\boldsymbol{\vartheta}$ is a placeholder for a focal parameter and \mathbf{c}_ϑ the coefficient (scalar, vector, or matrix) corresponding to this parameter. By construction $\epsilon_{ib} \setminus (\mathbf{c}_\vartheta \boldsymbol{\vartheta})$ is not a function of $\boldsymbol{\vartheta}$ and is hence a constant with respect to $q(\boldsymbol{\vartheta})$.

In addition we introduce notation to stack the \mathbf{x}_{ib} , \mathbf{w}_i , ϵ_{ib} , and α_{ib} vectors into matrices. $\mathbf{X}_{(1)}$ stacks the I \mathbf{x}_{ib} vectors that correspond with a first shopping trip in an $I \times K_X$ matrix. The same notation applies for \mathbf{w}_i ($\mathbf{W}_{(1)}$) and ϵ_{ib} ($\mathbf{E}_{(1)}$). Similarly, $\mathbf{X}_{(2+)}$ stacks the remaining \mathbf{x}_{ib} vectors that do not correspond with a first shopping trip in an $((\sum_{i=1}^I B_i) - I) \times K_X$ matrix. Again, the same notation applies for ϵ_{ib} ($\mathbf{E}_{(2+)}$) and \mathbf{w}_i ($\mathbf{W}_{(2+)}$), where $\mathbf{W}_{(2+)}$ contains $B_i - 1$ duplicated rows \mathbf{w}_i for each customer i . Lastly, $\mathbf{A}_{(-1)}$ stacks all lagged values of α_{ib} , i.e. all the vectors α_{ib} that do not correspond to a last shopping trip, in a $((\sum_{i=1}^I B_i) - I) \times M$ matrix.

Finally, we use the following notation in the remainder of this appendix:

- $[x]$: Iverson Bracket, equal to 1 if x is True, else 0
- $\mathcal{I}[z]$ with $z \in \{1, \dots, M\}$: Generates an M -dimensional ‘‘one-hot’’ vector filled with zeros, except for the z th element which is equal to 1
- $d(\mathbf{v})$: creates a diagonal matrix out of the vector \mathbf{v}
- $\tilde{\mathbb{E}}\{\boldsymbol{\vartheta}\}$: expectation of parameter $\boldsymbol{\vartheta}$ under its variational distribution $q(\boldsymbol{\vartheta})$, i.e. $\tilde{\mathbb{E}}\{\boldsymbol{\vartheta}\} \equiv \mathbb{E}_{q(\boldsymbol{\vartheta})}\{\boldsymbol{\vartheta}\}$
- \mathbf{I} : identity matrix; $\mathbf{1}$: vector of ones; $\mathbf{0}$: vector of zeros

D.1. Closed-form solutions for the variational updates

Our model specification only consists of distributions from the exponential family and we apply mean-field variational inference (VI) with a partitioning $F(\Omega)$ that retains all elements of a multivariate parameter within a single subset, ω , for all multivariate parameters in the model. This allows us to use known results from the VI literature to directly write down a general expression for the optimal variational distribution $q^*(\omega)$ for each conditionally conjugate parameter ω in Ω . In our model this set includes all parameters except α_{ib} .

More specifically, we use the general form of the optimal variational distribution under the mean-field assumption (Bishop 2006), applied to distributions in the exponential family (Blei et al. 2017). That is, given that ω refers to a conditionally conjugate parameter, and that ω is distributed according to a distribution from the exponential family in the model, it holds that the optimal variational distribution for ω is given by:

$$q^*(\omega) \propto h(\omega) \exp\left(t(\omega)^\top \tilde{\mathbb{E}}_{\text{MB}_\omega}\{\eta(\text{MB}_\omega)\}\right). \quad (\text{D.3})$$

Here h and t are functions that match the functional form corresponding to the base measure and the sufficient statistic of the prior distribution for ω in the model, and MB_ω refers to the Markov blanket for ω . $\eta(\text{MB}_\omega)$ is the natural parameter corresponding to the full-conditional distribution $p(\omega|\text{MB}_\omega)$. Its variational expectation, $\tilde{\mathbb{E}}_{\text{MB}_\omega}\{\eta(\text{MB}_\omega)\}$, is the natural parameter corresponding to the optimal variational distribution $q^*(\omega)$. Note that evaluating this expectation only requires the variational distributions of the parameters in the Markov blanket for ω .

Update $q(z_{ibn})$: M -dimensional Categorical with probability vector $\tilde{\mathbf{p}}_{ibn}$. The probability that $z_{ibn} = m$ under the variational distribution is given by:

$$\tilde{p}_{ibnm} = \frac{\exp\left(\tilde{\mathbb{E}}\{\alpha_{ibm}\} + \tilde{\mathbb{E}}\{\log \phi_{m,y_{ibn}}\}\right)}{\sum_{\ell=1}^M \exp\left(\tilde{\mathbb{E}}\{\alpha_{ib\ell}\} + \tilde{\mathbb{E}}\{\log \phi_{\ell,y_{ibn}}\}\right)}. \quad (\text{D.4})$$

Update $q(\phi_m)$: J -dimensional Dirichlet with J -dimensional parameter vector $\tilde{\mathbf{a}}_m$. The j th element of $\tilde{\mathbf{a}}_m$ is defined as:

$$\tilde{a}_{mj} = J^{-1} + \sum_{i=1}^I \sum_{b=1}^{B_i} \sum_{n=1}^{N_{ib}} \tilde{\mathbb{E}}\{[z_{ibn} = m]\}[y_{ibn} = j]. \quad (\text{D.5})$$

In the updates for $q(\kappa_i)$, $q(\rho_m)$, $q(\beta_m)$, and $q(\gamma_m)$, the variational parameters can be derived using a functional form that is related to the normal equations used in multiple regression, i.e. by using terms $\tilde{\mathbb{E}}\{\mathbf{X}^\top \mathbf{X}\}$ and $\tilde{\mathbb{E}}\{\mathbf{X}^\top \mathbf{y}\}$. Here the specification of \mathbf{X} and \mathbf{y} depends on the focal variational parameter ϑ . Given expressions for $\tilde{\mathbb{E}}\{\mathbf{X}^\top \mathbf{X}\}$ and $\tilde{\mathbb{E}}\{\mathbf{X}^\top \mathbf{y}\}$, the solution of the variational distribution can be written down as a multivariate Normal distribution with covariance and mean:

$$\Sigma = \tilde{\mathbb{E}}\{\mathbf{X}^\top \mathbf{X}\}^{-1}, \quad \boldsymbol{\mu} = \Sigma \tilde{\mathbb{E}}\{\mathbf{X}^\top \mathbf{y}\}. \quad (\text{D.6})$$

Update $q(\kappa_i)$: M -dimensional multivariate Normal with:

$$\begin{aligned} \tilde{\mathbb{E}}\{\mathbf{X}_{\kappa_i} \mathbf{y}_{\kappa_i}\} &= \tilde{\mathbb{E}}\{\Lambda_\kappa\} \tilde{\mathbb{E}}\{\boldsymbol{\mu}_\kappa\} + \tilde{\mathbb{E}}\{\boldsymbol{\tau}_\alpha\} \cdot \left(\tilde{\mathbb{E}}\{\delta_\kappa\} \tilde{\mathbb{E}}\{\boldsymbol{\epsilon}_{i1} \setminus (\kappa_i \delta_\kappa)\} + \sum_{b=2}^{B_i} \tilde{\mathbb{E}}\{\boldsymbol{\epsilon}_{ib} \setminus (\kappa_i)\} \right), \\ \tilde{\mathbb{E}}\{\mathbf{X}_{\kappa_i}^\top \mathbf{X}_{\kappa_i}\} &= \tilde{\mathbb{E}}\{\Lambda_\kappa\} + d(\tilde{\mathbb{E}}\{\boldsymbol{\tau}_\alpha\})(\tilde{\mathbb{E}}\{\delta_\kappa^2\} + B_i - 1). \end{aligned} \quad (\text{D.7})$$

Update $q(\boldsymbol{\beta}_m)$: K_X -dimensional multivariate Normal with:

$$\begin{aligned}\tilde{\mathbb{E}}\{\mathbf{X}_{\beta_m} \mathbf{y}_{\beta_m}\} &= \mathbf{0}_{K_X} + \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\delta_\beta\} \mathbf{X}_{(1)}^\top \tilde{\mathbb{E}}\{\mathbf{E}_{(1),m} \setminus (\mathbf{X}_{(1)} \boldsymbol{\beta}_m \delta_\beta)\} + \mathbf{X}_{(2+)}^\top \tilde{\mathbb{E}}\{\mathbf{E}_{(2+),m} \setminus (\mathbf{X}_{(2+)} \boldsymbol{\beta}_m)\}, \\ \tilde{\mathbb{E}}\{\mathbf{X}_{\beta_m}^\top \mathbf{X}_{\beta_m}\} &= \mathbf{I}_{K_X} + \left(\mathbf{X}_{(1)}^\top \mathbf{X}_{(1)} \tilde{\mathbb{E}}\{\delta_\beta^2\} + \mathbf{X}_{(2+)}^\top \mathbf{X}_{(2+)} \right) \tilde{\mathbb{E}}\{\tau_{\alpha_m}\}.\end{aligned}\tag{D.8}$$

Update $q(\boldsymbol{\gamma}_m)$: K_W -dimensional multivariate Normal with:

$$\begin{aligned}\tilde{\mathbb{E}}\{\mathbf{X}_{\gamma_m} \mathbf{y}_{\gamma_m}\} &= \mathbf{0}_{K_W} + \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\delta_\gamma\} \mathbf{W}_{(1)}^\top \tilde{\mathbb{E}}\{\mathbf{E}_{(1),m} \setminus (\mathbf{W}_{(1)} \boldsymbol{\gamma}_m \delta_\gamma)\} + \mathbf{W}_{(2+)}^\top \tilde{\mathbb{E}}\{\mathbf{E}_{(2+),m} \setminus (\mathbf{W}_{(2+)} \boldsymbol{\gamma}_m)\}, \\ \tilde{\mathbb{E}}\{\mathbf{X}_{\gamma_m}^\top \mathbf{X}_{\gamma_m}\} &= \mathbf{I}_{K_W} + \left(\mathbf{W}_{(1)}^\top \mathbf{W}_{(1)} \tilde{\mathbb{E}}\{\delta_\gamma^2\} + \mathbf{W}_{(2+)}^\top \mathbf{W}_{(2+)} \right) \tilde{\mathbb{E}}\{\tau_{\alpha_m}\}.\end{aligned}\tag{D.9}$$

Update $q(\boldsymbol{\rho}_m)$: M -dimensional multivariate Normal with:

$$\begin{aligned}\tilde{\mathbb{E}}\{\mathbf{X}_{\rho_m} \mathbf{y}_{\rho_m}\} &= \mathbf{0}_M + \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\mathbf{A}_{(-1)}\}^\top \tilde{\mathbb{E}}\{\mathbf{E}_{(2+),m} \setminus (\mathbf{A}_{(-1)} \boldsymbol{\rho}_m)\}, \\ \tilde{\mathbb{E}}\{\mathbf{X}_{\rho_m}^\top \mathbf{X}_{\rho_m}\} &= \mathbf{I}_M + \tilde{\mathbb{E}}\{\mathbf{A}_{(-1)}^\top \mathbf{A}_{(-1)}\} \tilde{\mathbb{E}}\{\tau_{\alpha_m}\}.\end{aligned}\tag{D.10}$$

Update $q(\delta_m)$: Normal with parameters:

$$\tilde{\mu}_{\delta_m} = \tilde{\sigma}_{\delta_m}^2 \left(0 + \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \sum_{i=1}^I \tilde{\mathbb{E}}\{\epsilon_{i1m} \setminus (\delta_m)\} \right), \quad \tilde{\sigma}_{\delta_m}^2 = (1 + I \tilde{\mathbb{E}}\{\tau_{\alpha_m}\})^{-1}.\tag{D.11}$$

Update $q(\delta_\kappa)$: Normal with parameters:

$$\begin{aligned}\tilde{\mu}_{\delta_\kappa} &= \tilde{\sigma}_{\delta_\kappa}^2 \left(1 + \sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \left(\sum_{i=1}^I \tilde{\mathbb{E}}\{\kappa_{im}\} \tilde{\mathbb{E}}\{\epsilon_{i1m} \setminus (\kappa_{im} \delta_\kappa)\} \right) \right), \\ \tilde{\sigma}_{\delta_\kappa}^2 &= \left(1 + \sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \left(\sum_{i=1}^I \tilde{\mathbb{E}}\{\kappa_{im}^2\} \right) \right)^{-1}.\end{aligned}\tag{D.12}$$

Update $q(\delta_\beta)$: Normal with parameters:

$$\begin{aligned}\tilde{\mu}_{\delta_\beta} &= \tilde{\sigma}_{\delta_\beta}^2 \left(1 + \sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\boldsymbol{\beta}_m\}^\top \mathbf{X}_{(1)}^\top \tilde{\mathbb{E}}\{\mathbf{E}_{(1),m} \setminus (\mathbf{X}_{(1)} \boldsymbol{\beta}_m \delta_\beta)\} \right), \\ \tilde{\sigma}_{\delta_\beta}^2 &= \left(1 + \text{trace} \left(\sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\boldsymbol{\beta}_m \boldsymbol{\beta}_m^\top\} \mathbf{X}_{(1)}^\top \mathbf{X}_{(1)} \right) \right)^{-1}.\end{aligned}\tag{D.13}$$

Update $q(\delta_\gamma)$: Normal with parameters:

$$\begin{aligned}\tilde{\mu}_{\delta_\gamma} &= \tilde{\sigma}_{\delta_\gamma}^2 \left(1 + \sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\boldsymbol{\gamma}_m\}^\top \mathbf{W}_{(1)}^\top \tilde{\mathbb{E}}\{\mathbf{E}_{(1),m} \setminus (\mathbf{W}_{(1)} \boldsymbol{\gamma}_m \delta_\gamma)\} \right), \\ \tilde{\sigma}_{\delta_\gamma}^2 &= \left(1 + \text{trace} \left(\sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\boldsymbol{\gamma}_m \boldsymbol{\gamma}_m^\top\} \mathbf{W}_{(1)}^\top \mathbf{W}_{(1)} \right) \right)^{-1}.\end{aligned}\tag{D.14}$$

Update $q(\tau_{\alpha_m})$: Gamma with parameters:

$$\tilde{\alpha}_{\tau_{\alpha_m}} = 1 + \frac{1}{2} \sum_{i=1}^I B_i, \quad \tilde{\beta}_{\tau_{\alpha_m}} = 1 + \frac{1}{2} \sum_{i=1}^I \sum_{b=1}^{B_i} \tilde{\mathbb{E}}\{\epsilon_{ibm}^2\}. \quad (\text{D.15})$$

Update $q(\mu_{\kappa})$: M -dimensional multivariate Normal with parameters:

$$\tilde{\boldsymbol{\mu}}_{\mu_{\kappa}} = \tilde{\boldsymbol{\Sigma}}_{\mu_{\kappa}} \left(\mathbf{0}_M + \tilde{\mathbb{E}}\{\boldsymbol{\Lambda}_{\kappa}\} \sum_{i=1}^I \tilde{\mathbb{E}}\{\boldsymbol{\kappa}_i\} \right), \quad \tilde{\boldsymbol{\Sigma}}_{\mu_{\kappa}} = (\mathbf{I}_M + I \tilde{\mathbb{E}}\{\boldsymbol{\Lambda}_{\kappa}\})^{-1}. \quad (\text{D.16})$$

Update $q(\boldsymbol{\Lambda}_{\kappa})$: $M \times M$ -dimensional Wishart with parameters:

$$\tilde{n}_{\boldsymbol{\Lambda}_{\kappa}} = 2M + I, \quad \tilde{\mathbf{V}}_{\boldsymbol{\Lambda}_{\kappa}} = \left((2M)\mathbf{I}_M + \sum_{i=1}^I \left(\tilde{\mathbb{E}}\{\boldsymbol{\mu}_{\kappa} \boldsymbol{\mu}_{\kappa}^{\top}\} + \tilde{\mathbb{E}}\{\boldsymbol{\kappa}_i \boldsymbol{\kappa}_i^{\top}\} - \tilde{\mathbb{E}}\{\boldsymbol{\mu}_{\kappa}\} \tilde{\mathbb{E}}\{\boldsymbol{\kappa}_i^{\top}\} - \tilde{\mathbb{E}}\{\boldsymbol{\kappa}_i\} \tilde{\mathbb{E}}\{\boldsymbol{\mu}_{\kappa}^{\top}\} \right) \right)^{-1}. \quad (\text{D.17})$$

D.2. Variational update for $q(\boldsymbol{\alpha}_{ib})$

Because $\boldsymbol{\alpha}_{ib}$ is not a conditionally conjugate parameter in the model, an analytical solution for its optimal variational distribution does not exist. Instead, we place $q(\boldsymbol{\alpha}_{ib})$ in the prior family for $\boldsymbol{\alpha}_{ib}$, i.e. a set of independent Normal distributions, with variational parameters $\tilde{\boldsymbol{\mu}}_{ib}$ and $\tilde{\boldsymbol{\sigma}}_{ib}^2$. In the optimization algorithm, we rely on gradient information to update these $\tilde{\boldsymbol{\mu}}_{ib}$ and $\tilde{\boldsymbol{\sigma}}_{ib}^2$ parameters.

To derive the gradients corresponding to $\tilde{\boldsymbol{\mu}}_{ib}$ and $\tilde{\boldsymbol{\sigma}}_{ib}^2$, we first denote the part of the ELBO that depends on the variational parameters $\tilde{\boldsymbol{\mu}}_{ib}$ and $\tilde{\boldsymbol{\sigma}}_{ib}^2$ of $q(\boldsymbol{\alpha}_{ib})$ as ELBO_{ib} :

$$\begin{aligned} \text{ELBO}_{ib} &= -\frac{1}{2} \tilde{\mathbb{E}}\{\boldsymbol{\tau}_{\alpha}\}^{\top} \left(\tilde{\boldsymbol{\sigma}}_{ib}^2 + \tilde{\boldsymbol{\mu}}_{ib}^2 \right) + (\tilde{\mathbb{E}}\{\boldsymbol{\tau}_{\alpha}\} \cdot \tilde{\mathbb{E}}\{\boldsymbol{\mu}_{ib}\})^{\top} \tilde{\boldsymbol{\mu}}_{ib} \\ &\quad + \left(\sum_{n=1}^{N_{ib}} \tilde{\mathbb{E}}\{\mathcal{I}[z_{ibn}]\} \right)^{\top} \tilde{\boldsymbol{\mu}}_{ib} - N_{ib} \log \sum_{m=1}^M \exp(\tilde{\boldsymbol{\mu}}_{ibm} + \frac{1}{2} \tilde{\boldsymbol{\sigma}}_{ibm}^2) \\ &\quad + [b \neq B_i] \text{trace} \left(-\frac{1}{2} \left(d(\tilde{\boldsymbol{\sigma}}_{ib}^2) + \tilde{\boldsymbol{\mu}}_{ib} \tilde{\boldsymbol{\mu}}_{ib}^{\top} \right) \left(\sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\boldsymbol{\rho}_m \boldsymbol{\rho}_m^{\top}\} \right) \right) \\ &\quad + [b \neq B_i] (\tilde{\mathbb{E}}\{\boldsymbol{\tau}_{\alpha}\} \cdot \tilde{\mathbb{E}}\{\boldsymbol{\epsilon}_{ib+1} \setminus (\mathbf{R} \boldsymbol{\alpha}_{ib})\})^{\top} \tilde{\mathbb{E}}\{\mathbf{R}\} \tilde{\boldsymbol{\mu}}_{ib} + \frac{1}{2} \sum_{m=1}^M \log \tilde{\boldsymbol{\sigma}}_m^2, \end{aligned} \quad (\text{D.18})$$

where we have applied Jensen's Inequality to lowerbound $-\tilde{\mathbb{E}}\{\log \sum_{m=1}^M \exp \alpha_{ibm}\}$, which is the variational expectation of the log of the denominator of the softmax function, with $-\log \sum_{m=1}^M \tilde{\mathbb{E}}\{\exp \alpha_{ibm}\}$ (Braun and McAuliffe 2010).

The gradient of (D.18) with respect to $\tilde{\boldsymbol{\mu}}_{ib}$ is given by:

$$\begin{aligned} \nabla_{\tilde{\boldsymbol{\mu}}_{ib}} \text{ELBO}_{ib} &= \tilde{\mathbb{E}}\{\boldsymbol{\tau}_{\alpha}\} \cdot (\tilde{\mathbb{E}}\{\boldsymbol{\mu}_{ib}\} - \tilde{\boldsymbol{\mu}}_{ib}) \\ &\quad + \sum_{n=1}^{N_{ib}} \tilde{\mathbb{E}}\{\mathcal{I}[z_{ibn}]\} - N_{ib} \frac{\exp(\tilde{\boldsymbol{\mu}}_{ib} + \frac{1}{2} \tilde{\boldsymbol{\sigma}}_{ib}^2)}{\sum_{m=1}^M \exp(\tilde{\boldsymbol{\mu}}_{ibm} + \frac{1}{2} \tilde{\boldsymbol{\sigma}}_{ibm}^2)} \\ &\quad + [b \neq B_i] \left(\sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\boldsymbol{\rho}_m \boldsymbol{\rho}_m^{\top}\} \right) \tilde{\boldsymbol{\mu}}_{ib} \\ &\quad + [b \neq B_i] \tilde{\mathbb{E}}\{\mathbf{R}\}^{\top} (\tilde{\mathbb{E}}\{\boldsymbol{\tau}_{\alpha}\} \cdot \tilde{\mathbb{E}}\{\boldsymbol{\epsilon}_{ib+1} \setminus (\mathbf{R} \boldsymbol{\alpha}_{ib})\}). \end{aligned} \quad (\text{D.19})$$

The gradient of (D.18) with respect to $\tilde{\sigma}_{ib}^2$ is given by:

$$\begin{aligned}
\nabla_{\tilde{\sigma}_{ib}^2} \text{ELBO}_{ib} &= -\frac{1}{2} \tilde{\mathbb{E}}\{\tau_\alpha\} \\
&\quad -\frac{1}{2} N_{ib} \frac{\exp(\tilde{\mu}_{ib} + \frac{1}{2} \tilde{\sigma}_{ib}^2)}{\sum_{m=1}^M \exp(\tilde{\mu}_{ibm} + \frac{1}{2} \tilde{\sigma}_{ibm}^2)} \\
&\quad -\frac{1}{2} [b \neq B_i] \mathbf{1} \left(\sum_{m=1}^M \tilde{\mathbb{E}}\{\tau_{\alpha_m}\} \tilde{\mathbb{E}}\{\rho_{mm}^2\} \right) \\
&\quad + \frac{1}{2} \tilde{\sigma}^2.
\end{aligned} \tag{D.20}$$

D.3. Initialization of $q(\Omega)$

This appendix contains the initialization settings for $q(\Omega)$ used in Algorithm 1. In general, we can divide the Normally distributed parameters in two groups: one group for which the initialization of the variational variance has an effect on the outcome of the optimization; and for the other group this initialization has no effect. This independence of the initialization is due to the order in which the variational updates are performed in the optimization algorithm. For the population parameters for which the initialization of the variational variance matters, we initialize this variance to a small conservative value M^{-2} to allow for a smooth optimization path. Additionally, the posterior distributions of these population parameters are expected to be sharply peaked, as a lot of information is available to estimate parameters at the population level. For the variance of $q(\alpha_{ibm})$ we specify an initial value of 1, as less information is available at the level of an individual shopping trip. Full details of the variational initialization per parameter are provided below.

- $q(\phi_m) = \text{Dirichlet}_J$: the parameter vector of the Dirichlet distribution is initialized according to the pseudocounts extracted from the output of a Collapsed Gibbs LDA algorithm applied to the data
- $q(z_{ibn}) = \text{Categorical}_M(\mathbf{p} = [M^{-1}, \dots, M^{-1}])$: this initialization has no effect as $q(z_{ibn})$ is updated before $q(\phi)$ is updated in the coordinate ascent algorithm
- $q(\alpha_{ibm}) = \text{Normal}(\mu = 0, \sigma^2 = 1)$
- $q(\kappa_i) = \text{MVN}_M(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$: initial covariance has no effect as $q(\kappa_i)$ is updated before $q(\delta_\kappa), q(\tau_\alpha)$
- $q(\rho_m) = \text{MVN}_M(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I}M^{-2})$
- $q(\beta_m) = \text{MVN}_{K_X}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$: initial covariance has no effect as $q(\beta_m)$ is updated before $q(\delta_\beta), q(\tau_\alpha)$
- $q(\gamma_m) = \text{MVN}_{K_W}(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$: initial covariance has no effect as $q(\gamma_m)$ is updated before $q(\delta_\gamma), q(\tau_\alpha)$
- $q(\delta_m) = \text{Normal}(\mu = 0, \sigma^2 = 1)$: initial variance has no effect as $q(\delta_m)$ is updated before $q(\tau_\alpha)$
- $q(\delta_\kappa) = \text{Normal}(\mu = 1, \sigma^2 = M^{-2})$
- $q(\delta_\beta) = \text{Normal}(\mu = 1, \sigma^2 = M^{-2})$
- $q(\delta_\gamma) = \text{Normal}(\mu = 1, \sigma^2 = M^{-2})$
- $q(\tau_{\alpha_m}) = \text{Gamma}(\alpha = 1, \beta = 1)$: only the ratio α/β affects the outcome of the optimization
- $q(\mu_\kappa) = \text{MVN}_M(\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I})$: initial covariance has no effect as $q(\mu_\kappa)$ is updated before $q(\Lambda_\kappa)$
- $q(\Lambda_\kappa) = \text{Wishart}_M(n = 2M, \mathbf{V} = \mathbf{I}(2M)^{-1})$

E. Online Appendix: Description of explanatory variables

This appendix describes the exogenous explanatory variables in the data. The only information available at the shopping trip level besides the IDs for the purchased products and the customer is the date and time of the trip. This information can be used to construct a set of dummy variables that all relate to the timing of the shopping trip. The frequencies of these variables are displayed in Table 11. The levels ‘Year 1’, ‘March’, ‘Weekday’ and ‘Daytime (00:00–17:00)’ are taken as the reference levels, such that 14 dummy variables remain in \mathbf{x}_{ib} for which the effects have to be estimated in the model. For all variables the 0/1 dummy coding is demeaned such that $\sum_{ib} \mathbf{x}_{ib} = \mathbf{0}$.

Dummy level	Frequency	Percentage
Year 1	20,704	43.53%
Year 2	26,864	56.47%
January	3,803	7.99%
February	3,559	7.48%
March	4,140	8.70%
April	4,195	8.82%
May	4,168	8.76%
June	3,825	8.04%
July	3,802	7.99%
August	3,858	8.11%
September	3,700	7.78%
October	3,823	8.04%
November	4,410	9.27%
December	4,285	9.01%
Weekday	30,844	64.84%
Weekend	16,724	35.16%
Daytime (00:00–17:00)	37,474	78.78%
Evening (17:00–00:00)	10,094	21.22%

Table 11: Frequencies of the explanatory variables at the shopping trip level.

The information available at the customer level is related to age, gender, and household size of the customer and contains missing values. Age is summarized in five age brackets. Gender is a binary variable indicating whether a customer is male or female. A missing value for age or gender is included as a separate dummy variable. A proxy for household size of a customer is available as well, from which a dummy is created indicating if this a small household (proxy < 2) or a large household (proxy ≥ 2). The frequencies of these variables are displayed in Table 12. The levels ‘Age 45–55’, ‘Gender: Male’, and ‘Household: Small’ are taken as the reference levels, such that 8 dummy variables remain in \mathbf{w}_i for which the effects have to be estimated in the model. For all variables the 0/1 dummy coding is demeaned such that $\sum_i \mathbf{w}_i = \mathbf{0}$.

Dummy level	Count	Percentage
Age: Unknown	228	10.09%
Age: 0–35	234	10.36%
Age: 35–45	325	14.39%
Age: 45–55	458	20.27%
Age: 55–65	495	21.91%
Age: 65+	519	22.97%
Gender: Unknown	168	7.44%
Gender: Female	857	37.94%
Gender: Male	1,234	54.63%
Household: Small	932	41.26%
Household: Large	1,327	58.74%

Table 12: Frequencies of the explanatory variables at the customer level.

References

- Asim Ansari and Carl F. Mela. E-customization. *Journal of Marketing Research*, 40(2):131–145, 2003. doi: 10.1509/jmkr.40.2.131.19224. URL <https://doi.org/10.1509/jmkr.40.2.131.19224>.
- Asim Ansari, Yang Li, and Jonathan Z. Zhang. Probabilistic topic model for hybrid recommender systems: A stochastic Variational Bayesian approach. *Marketing Science*, 37(6):987–1008, 2018. doi: 10.1287/mksc.2018.1113. URL <https://doi.org/10.1287/mksc.2018.1113>.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55:77–84, 2012.
- David M. Blei and John D. Lafferty. A correlated topic model of science. *Ann. Appl. Stat.*, 1(1):17–35, 06 2007. doi: 10.1214/07-AOAS114. URL <https://doi.org/10.1214/07-AOAS114>.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL <https://doi.org/10.1080/01621459.2017.1285773>.
- Anand V. Bodapati. Recommendation systems with purchase data. *Journal of Marketing Research*, 45(1):77–93, 2008. doi: 10.1509/jmkr.45.1.077. URL <https://doi.org/10.1509/jmkr.45.1.077>.
- Ingwer Borg and Patrick J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.
- Stephen Boyd and Lieven Vandenbergh. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.
- Michael Braun and Jon McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010. doi: 10.1198/jasa.2009.tm08030. URL <https://doi.org/10.1198/jasa.2009.tm08030>.
- Bart J. Bronnenberg, Vijay Mahajan, and Wilfried R. Vanhonacker. The emergence of market structure in new repeat-purchase categories: The interplay of market share and retailer distribution. *Journal of Marketing Research*, 37(1):16–31, 2000. doi: 10.1509/jmkr.37.1.16.18723. URL <https://doi.org/10.1509/jmkr.37.1.16.18723>.
- Norris I. Bruce, Kay Peters, and Prasad A. Naik. Discovering how advertising grows sales and builds brands. *Journal of Marketing Research*, 49(6):793–806, 2012. doi: 10.1509/jmr.11.0060. URL <https://doi.org/10.1509/jmr.11.0060>.
- Joachim Büschken and Greg M. Allenby. Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6):953–975, 2016. doi: 10.1287/mksc.2016.0993. URL <https://doi.org/10.1287/mksc.2016.0993>.
- Joachim Büschken and Greg M. Allenby. Improving text analysis using sentence conjunctions and punctuation. *Marketing Science*, 39(4):727–742, 2020. doi: 10.1287/mksc.2019.1214. URL <https://doi.org/10.1287/mksc.2019.1214>.
- Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL <https://www.jstatsoft.org/v076/i01>.
- José Mauro da Costa Hernandez, Scott A. Wright, and Filipe Ferminiano Rodrigues. Attributes versus benefits: The role of construal levels and appeal type on the persuasiveness of marketing messages. *Journal of Advertising*, 44(3):243–253, 2015. doi: 10.1080/00913367.2014.967425. URL <https://doi.org/10.1080/00913367.2014.967425>.
- Ryan Dew and Asim Ansari. Bayesian nonparametric customer base analysis with model-based visualizations. *Marketing Science*, 37(2):216–235, 2018. doi: 10.1287/mksc.2017.1050. URL <https://doi.org/10.1287/mksc.2017.1050>.
- Ryan Dew, Asim Ansari, and Yang Li. Modeling dynamic heterogeneity using Gaussian processes. *Journal of Marketing Research*, forthcoming, 2019. doi: 10.1177/0022243719874047. URL <https://doi.org/10.1177/0022243719874047>.
- Daria Dzyabura and John R. Hauser. Active machine learning for consideration heuristics. *Marketing Science*, 30(5):801–819, 2011. doi: 10.1287/mksc.1110.0660. URL <https://doi.org/10.1287/mksc.1110.0660>.

- Peter M. Guadagni and John D. C. Little. A logit model of brand choice calibrated on scanner data. *Marketing Science*, 2(3):203–238, 1983. doi: 10.1287/mksc.2.3.203. URL <https://doi.org/10.1287/mksc.2.3.203>.
- Sunil Gupta. Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research*, 25(4):342–355, 1988. doi: 10.1177/002224378802500402. URL <https://doi.org/10.1177/002224378802500402>.
- Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013. URL <http://jmlr.org/papers/v14/hoffman13a.html>.
- Bruno J. D. Jacobs, Bas Donkers, and Dennis Fok. Model-based purchase predictions for large assortments. *Marketing Science*, 35(3):389–404, 2016. doi: 10.1287/mksc.2016.0985. URL <https://doi.org/10.1287/mksc.2016.0985>.
- Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov 1999. ISSN 1573-0565. doi: 10.1023/A:1007665907178. URL <https://doi.org/10.1023/A:1007665907178>.
- Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
- Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017. URL <http://jmlr.org/papers/v18/16-107.html>.
- Jia Liu and Olivier Toubia. A semantic approach for estimating consumer content preferences from online search queries. *Marketing Science*, 37(6):930–952, 2018. doi: 10.1287/mksc.2018.1112. URL <https://doi.org/10.1287/mksc.2018.1112>.
- Puneet Manchanda, Asim Ansari, and Sunil Gupta. The “shopping basket”: A model for multicategory purchase incidence decisions. *Marketing Science*, 18(2):95–114, 1999. doi: 10.1287/mksc.18.2.95. URL <https://doi.org/10.1287/mksc.18.2.95>.
- Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. CRC press, 2011.
- J. T. Ormerod and M. P. Wand. Explaining variational approximations. *The American Statistician*, 64(2):140–153, 2010. doi: 10.1198/tast.2010.09058. URL <https://doi.org/10.1198/tast.2010.09058>.
- Max J. Pachali, Peter Kurz, and Thomas Otter. How to generalize from a hierarchical model? *Quantitative Marketing and Economics*, forthcoming, 2020.
- Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.
- Dinesh Puranam, Vishal Narayan, and Vrinda Kadiyali. The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors. *Marketing Science*, 36(5):726–746, 2017. doi: 10.1287/mksc.2017.1048. URL <https://doi.org/10.1287/mksc.2017.1048>.
- Peter E. Rossi and Greg M. Allenby. Bayesian statistics and marketing. *Marketing Science*, 22(3):304–328, 2003. doi: 10.1287/mksc.22.3.304.17739. URL <https://doi.org/10.1287/mksc.22.3.304.17739>.
- Peter E. Rossi, Robert E. McCulloch, and Greg M. Allenby. The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340, 1996. doi: 10.1287/mksc.15.4.321. URL <https://doi.org/10.1287/mksc.15.4.321>.
- Francisco J. R. Ruiz, Susan Athey, and David M. Blei. Shopper: A probabilistic model of consumer choice with substitutes and complements. *Ann. Appl. Stat.*, 14(1):1–27, 03 2020. doi: 10.1214/19-AOAS1265. URL <https://doi.org/10.1214/19-AOAS1265>.
- Oliver J. Rutz, Garrett P. Sonnier, and Michael Trusov. A new method to aid copy testing of paid search text advertisements. *Journal of Marketing Research*, 54(6):885–900, 2017. doi: 10.1509/jmr.14.0186. URL <https://doi.org/10.1509/jmr.14.0186>.
- Seshadri Tirunillai and Gerard J. Tellis. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, 51(4):463–479, 2014. doi: 10.1509/jmr.12.0106. URL <https://doi.org/10.1509/jmr.12.0106>.
- Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

- Michael Trusov, Liye Ma, and Zainab Jamal. Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science*, 35(3):405–426, 2016. doi: 10.1287/mksc.2015.0956. URL <https://doi.org/10.1287/mksc.2015.0956>.
- Hanna M. Wallach, David M. Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. Curran Associates, Inc., 2009. URL <http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf>.
- Michel Wedel and P.K. Kannan. Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6):97–121, 2016. doi: 10.1509/jm.15.0413. URL <https://doi.org/10.1509/jm.15.0413>.
- Feihong Xia, Rabikar Chatterjee, and Jerrold H. May. Using conditional restricted Boltzmann machines to model complex consumer shopping patterns. *Marketing Science*, 38(4):711–727, 2019. doi: 10.1287/mksc.2019.1162. URL <https://doi.org/10.1287/mksc.2019.1162>.