Overcoming the Cold Start Problem of CRM using a Probabilistic Machine Learning Approach

Nicolas Padilla
Eva Ascarza[†]

July 2021

[†]Nicolas Padilla is an Assistant Professor of Marketing, London Business School (email: npadilla@london.edu). Eva Ascarza is the Jakurski Family Associate Professor of Business Administration, Harvard Business School (email: eascarza@hbs.edu). The authors are grateful to the Wharton Customer Analytics Initiative (WCAI) for providing the data used in the empirical application. The authors thank Bruce Hardie, Donald Lehmann, Daniel McCarthy, and Oded Netzer for very useful comments and suggestions, the participants of the seminars at Harvard Business School, McCombs School of Business, Rotterdam School of Management, Tilburg University, Tuck School of Business, Questrom School of Business, Rady School of Management at UCSD, The Wharton School, and the audiences of the 2018 Marketing Science conference and the WCAI symposium for their comments. The authors are grateful to Hengyu Kuang for excellent research assistantship.

Abstract

Overcoming the Cold Start Problem of CRM using a Probabilistic Machine Learning Approach

The success of Customer Relationship Management (CRM) programs ultimately depends on the firm's ability to identify and leverage differences across customers — a very difficult task when firms attempt to manage new customers, for whom only the first purchase has been observed. For those customers, the lack of repeated observations poses a structural challenge to inferring unobserved differences across them. This is what we call the "cold start" problem of CRM, whereby companies have difficulties leveraging existing data when they attempt to make inferences about customers at the beginning of their relationship. We propose a solution to the cold start problem by developing a probabilistic machine learning modeling framework that leverages the information collected at the moment of acquisition. The main aspect of the model is that it flexibly captures latent dimensions that govern the behaviors observed at acquisition as well as future propensities to buy and to respond to marketing actions using deep exponential families. The model can be integrated with a variety of demand specifications and is flexible enough to capture a wide range of heterogeneity structures. We validate our approach in a retail context and empirically demonstrate the model's ability at identifying high-value customers as well as those most sensitive to marketing actions, right after their first purchase.

Keywords: Customer Relationship Management (CRM), Deep Exponential Families, Probabilistic Machine Learning, Cold Start Problem.

1 INTRODUCTION

Customers are different, not only in their preferences for products and services, but also in the way they respond to marketing actions. Understanding customer heterogeneity is at the heart of Customer Relationship Management (CRM) programs — from obtaining accurate estimates of the value of current and future customers, to deciding which individual customers should be targeted in the next marketing campaign. Over the last three decades, the marketing literature has provided researchers and analysts with methods to empirically identify unobserved differences across customers using their past history — e.g., customers with higher versus lower expected lifetime value (e.g., Schmittlein et al. 1987; Fader et al. 2005, 2010), those who are less sensitive to a price increase (e.g., Rossi et al. 1996; Allenby and Rossi 1998), or those who are more receptive to marketing communications (e.g., Ansari and Mela 2003). However, when firms attempt to implement CRM programs on customers who have been acquired recently, they only observe these customers' first purchase. This lack of repeated observations presents a structural challenge for estimating unobserved differences across recently-acquired customers, precluding firms from leveraging such heterogeneity.[1] We call this the "cold start" problem of CRM; that is, the challenge that firms face when trying to make inferences about customers at the outset of the relationship, for whom data are limited.

Firms have traditionally relied on demographics (e.g., age, gender) and/or recency metrics (e.g., how many weeks since your last transaction) to target marketing efforts with limited data (Shaffer and Zhang 1995). These approaches, however, face practical limitations: Recency metrics, for example, do not differentiate among recently acquired customers (as they all were acquired at the same time), and relevant personal information is generally hard to collect or poses data privacy challenges. Although, thanks to technological advances, firms can now increasingly observe a wider range of behaviors on each customer touch. What in the past might have been considered simply a transaction added to a customer base is now a collection of behaviors that a customer incurs while making a first purchase (e.g., is that transaction online or offline, did they buy any new products or any old best-sellers in that transaction, did they buy any products on discount). While some of these characteristics may be purely coincidental with the moment in which the customer

[1] In this research we define customer "heterogeneity" as differences in propensities, preferences and sensitivities across customers. This view is very much aligned with the traditional view in Marketing (Allenby and Rossi 1998) of heterogeneity capturing individual differences in the model parameters.

made their first purchase, others may carry important information as they reflect latent customer preferences/attitudes. Thus, whereas firms only observe a just-acquired customer in one occasion, they now have many more cues to form a "first impression" of who this customer is, which can be used to understand heterogeneity across recently acquired customers.

We present a solution to the cold start problem that is flexible, scalable, and general. Specifically, we augment transactional data with information collected when a customer makes their first purchase — information already available in the firm's database — and propose a probabilistic machine learning modeling framework that extracts information relevant to making inferences about the customer's future behavior. The model, which we term the "First Impression Model" (FIM), reflects the premise that behaviors and choices observed in newly-acquired customers can be informative about underlying traits that are, in turn, predictive of their future behavior. We operationalize these customer traits via a finite set of latent factors that enable the model to reduce the dimensionality of, while extracting relevant signals from, the data, and assume those traits to drive, at least partially, customer behaviors observed both at the moment of acquisition and in the future.

In essence, the FIM is a deep probabilistic model of demand (main outcome of interest to the firm) and acquisition characteristics (customer outcomes that are observed to the firm at the moment of acquisition) where the individual-level parameters of each of these sub-models are projected into a lower-dimension space using a two-layered deep exponential family (DEF) component. The lower layer of the DEF component captures the relevant interrelations among the individual-level parameters. We incorporate automatic relevance determination priors (ARD) for this layer, enforcing sparsity and automatically reducing the dimensionality of the individual-level parameters, similarly as in a Bayesian PCA model and modern applications of "supervised" factor models. The model departs from the aforementioned models by allowing non-linear relationships among the factors in the lower layer, through the upper layer.

First among four notable aspects of the proposed modeling approach is that the model is able to capture a wide range of relationships between observed behaviors and variables of interest, for example, the interaction effects between two (or more) acquisition variables and the outcomes of interest. As the model will recover them from the data, those (linear or non-linear) relationships do not need to be pre-specified. Second, unlike traditional dimensionality reduction methods, the number of latent factors do not need to be specified a priori. The model infers the number of

relevant dimensions from the data through automatic relevance determination. Third, the model is scalable, being applicable to datasets with large numbers of customers and many acquisition characteristics, some of which might contain missing observations. When present, these missing observations are easily handled by the FIM, which models them as outcomes using a Bayesian estimation framework. Lastly, the proposed modeling framework is general in the sense that can be integrated with any demand specification, from simple linear specifications to more complex model structures that incorporate a latent attrition component (a.k.a., "buy-till-you-die" models) or other forms of customer dynamics (e.g., hidden Markov models). This desirable feature implies that marketers across business settings, contractual and non-contractual, can use this framework by making minor adjustments to the demand/transactional model.

Using a set of simulation analyses, we demonstrate the FIM inferences for newly-acquired customers' to be more accurate than those generated by multiple tested benchmarks. Unlike other models, our approach accommodates flexible relationships among relevant behaviors, enabling the model to make accurate inferences about newly-acquired customers when the relationships between acquisition characteristics and demand parameters are unknown to the firm or researcher.

We then apply the FIM to a retail context and demonstrate how the focal firm can overcome the cold start problem by augmenting the (thin) historical data using their transactional database and employing the proposed modeling framework that extracts the relevant information from the augmented customer data. First, we use the transactional data to extract the characteristics of every customer's first purchase (namely price paid, number of products purchased, etc.) as well as observed product characteristics such as category purchased, package size, etc. Second, we leverage the transactional data from customers outside our sample to create a continuous multidimensional representation of products (or product embeddings). Specifically, we use the word2vec algorithm — a machine learning approach originally developed to analyze textual data — to model the co-occurrence of products in customer baskets. This yields a set of product embeddings that can be used to augment data on customers' first transactions based on the specific products they bought. We then estimate the FIM to the augmented cold start data and make individual-level predictions for newly-acquired customers outside the calibration sample.

We empirically demonstrate the superiority of the FIM at distinguishing, immediately after they make their first purchase, heavy spenders from those expected to yield less value. The model can also be used to highlight the set of acquisition characteristics most predictive of future behavior.

For example, we find the predicted Top 10% heavy spenders to be less likely to be acquired during the holiday period and more likely to be acquired offline, and their first purchases to tend to include expensive and discounted products. The model also captures differences in customer responsiveness to marketing actions, enabling firms to identify and characterize those most (or least) sensitive to specific marketing communications. For example, we find that customers most sensitive to email marketing are more likely to be acquired online and buy less expensive products, and their first purchases to include fewer units. We also find non-linear relationships between acquisition characteristics and customer responsiveness to marketing actions. For example, the differences in email sensitivities across customers that received discounts on their first purchase only exist for those who also purchased a recently introduced product.

The present research develops a modeling framework that overcomes the cold start problem by linking customers' early observed behaviors and choices with future purchase behavior, enabling firms to make meaningful predictions about customers just acquired. Methodologically, our paper contributes to the CRM literature by being the first to incorporate in a general, flexible, and scalable way information obtained at the moment of acquisition (generally discarded due to an inability to use it effectively). Substantively, our research is relevant to marketers faced with the challenge of managing customers soon after acquisition. We show how the proposed modeling framework enables firms to identify and characterize, from information collected at the moment of acquisition, high-value customers and those most sensitive to marketing communications. From a practical perspective, our research guides firms in the use of cold start data to augment information already in their databases. To that end, we employ developments in machine learning and natural language processing to create a matrix of product "embeddings" that enable firms to characterize (even recently acquired) customers based on the products they purchase. We believe this approach to customer segmentation to be highly promising, enabling firms to obtain rich information about individual customers without recourse to customer-provided data or external sources that might pose privacy concerns.

The remainder of the paper is organized as follows. Following a brief review of the literature related to our work, we introduce the cold start problem and illustrate the main challenges to solving it in practice. We next present our modeling framework, discuss its components, and evaluate its performance vis-à-vis existing approaches that could be used to solve the cold start problem. We

then apply our model in the context of an international beauty and cosmetic retailer. We conclude with a discussion of the implications, managerial relevance, and future directions of our research.

2 PREVIOUS LITERATURE

Our research relates to the broad literature on customer-base analysis that has provided managers and analysts with tools for understanding, forecasting, and managing the (heterogeneous) behavior of customers. It relates particularly to work that has incorporated the effect of marketing variables or, more generally, time-varying covariates in customer lifetime value (CLV) models. Notable work in this area includes Schweidel and Knox (2013) and Schweidel et al. (2014) who, building on the foundations of the Beta-Geometric/Beta-Binomial (BG/BB) model (Fader et al. 2010), incorporate the effect of direct marketing activity and past customer activity on the latent attrition process and the customer's purchase propensity while alive, and Knox and van Oest (2014) and Braun et al. (2015) , who incorporate the effect of the customer service experience and customer complaints on the latent attrition process of the Beta-Geometric/NBD (BG/NBD) model (Fader et al. 2005). Our research and methodological objectives differ in two main ways. Whereas the main purpose of the aforementioned studies is to capture the effect of time-varying marketing variables (e.g., direct marketing activities, customer complaints) on customer behavior, we extract as much information as possible from cold start data. The referenced models, although they could be used to incorporate a handful of pre-specified acquisition variables, are not well suited to extract relevant information from noisy and redundant variables, the case with cold start data. Second, we do not build on a specific demand specification tied to a business context, but rather provide a modeling framework that can incorporate any of the models of behavior presented in the foregoing papers.

On a substantive level, our work relates to Gopalakrishnan et al. (2016), who propose a framework for multi-cohort data able to predict the behavior of new cohorts of customers for whom little transactional data are available. Gopalakrishnan and colleagues build a model that allows customers to be inherently different depending on when they were acquired (i.e., *which cohort* they belong to), while capturing the underlying dynamics across cohorts. We posit that such inherent heterogeneity can be explained (at least partially) by individual-level observed characteristics collected when customers make their first purchase. This is consistent with Anderson et al. (2020) who document the existence of "harbinger products." These are products that, when purchased by a customer in their first transaction, are an indicator of the customer being less likely to purchase again, and hence, provide less value to the firm. Our work also relates to Loupos et al. (2019), who

use social network data for recently acquired customers to explain heterogeneity in their future value to the firm. To the best of our knowledge, our approach is the first to integrate several types of information collected at the moment of acquisition, and to differentiate responsiveness to marketing actions — not only individual propensity to transact — on the basis of customers' first purchases. The latter aspect is crucial in cases in which targeting occurs soon after the customer is acquired or when securing a second purchase is challenging.

The premise that behaviors observed at the moment of acquisition can help firms explain heterogeneity in future behavior is consistent with empirical findings in the CRM literature (e.g., Fader et al. 2007; Voigt and Hinz 2016), specifically, work on customer acquisition that has investigated the relationship between acquisition-related information — e.g., channel of acquisition — and subsequent customer lifetime value (e.g., Verhoef and Donkers 2005; Lewis 2006; Villanueva et al. 2008; Chan et al. 2011; Steffes et al. 2011; Schmitt et al. 2011; Uncles et al. 2013; Datta et al. 2015). Our work, although it investigates relationships between acquisition-related variables and subsequent customer behavior, differs in two important ways. First, our end goal is to inform decisions related to the management of already acquired customers (e.g., whom to target in the next campaign) rather than the design of optimal strategies for customer acquisition (e.g., free trials to increase customer acquisition). The goal of our modeling framework is to extract as much observed heterogeneity as possible from initial behaviors while controlling for firms' acquisition activities rather than estimate the casual impact of these acquisition variables on future behavior. Second, this literature suggests that customers are inherently different depending on how they have been acquired. We broaden the range of acquisition-related behaviors by looking not only at *how* a customer was acquired (e.g., online vs. offline, trial vs. regular), but also *what* they did when they were acquired (e.g., what kind of product did they buy? how much did they pay?), hence extracting more information from the initial transaction. The latter is especially relevant for managers and analysts in large retail and hospitality businesses, among others, such information not only being easily observed, but typically already residing in their databases.

From a methodological perspective, we contribute to the literature on applying probabilistic machine learning methods to marketing (Jacobs et al. 2016; Dew and Ansari 2018; Dew et al. 2020). More specifically, our work relates to the literature on applying deep exponential families (Ranganath et al. 2015) as building blocks of more complex models (Ranganath et al. 2016; Wang and

Blei 2019), and other generative models such as Bayesian Principal Component Analysis (Bishop 1999; Mohamed et al. 2008).

3 THE "COLD START" PROBLEM: AN EXAMPLE FROM A RETAIL SETTING

We turn to a retail context to illustrate the cold start problem, and to motivate and validate our modeling framework. Retail is a good context to examine this phenomenon for several reasons. First, firms in this sector increasingly collect transactional data and rely on analytics to better manage their customers (Forbes 2015). Second, retail represents a large proportion of the total economy, with revenues accounting for 31% for the global GDP (Research and Markets 2016). Finally, the data structure in most retail settings — in particular, the one used in this research — resembles that in many other industries such as hospitality, entertainment business, or nonprofit organizations, that face similar data challenges when implementing CRM programs.

3.1 The "cold start" problem

Consider a retailer that sells cosmetic/beauty products both via online and offline channels.[2] Like most other companies, it records the transactions of all individual customers since the moment they were acquired, including the time of purchase, the products purchased in each particular transaction, their price and discounts (if any), along with information about the CRM activities that the company engaged with, such as email marketing activities. With these transactional data at hand, the focal company could apply some of the aforementioned models and be able to predict, with a good degree of accuracy, the number of transactions that customers with different transaction patterns would make in future periods (e.g., Fader et al. 2010). The marketer can also incorporate the historical marketing actions to capture how those variables affected transaction propensities and customer value (e.g., Schweidel and Knox 2013; Schweidel et al. 2014). However, when making these types of inferences for recently acquired customers, for whom the firm has no transactional history nor past marketing interventions, the "best guess" that the marketer can get is the population average. This is what we call the "cold start problem of CRM" whereby firms cannot make individual-level inferences about newly-acquired customers that differentiates them, therefore diminishing the effectiveness of future CRM activities.

[2]This will be the specific context of our empirical application. The full set of details about the focal firm and the data will be presented in Section 5; in this section we only present the relevant information to motivate the business problem and the modeling challenges.

The premise of this research is that, while it is the lack of (historical) data that causes the cold start problem, firms nowadays have access to other data sources that, properly leveraged, can help them overcome the cold start problem. Granted, if firms only observed that the customer made "a transaction" it would be very difficult to overcome the cold start problem. However, most firms not only know when a customer made their first transaction but also record the details such as the channel/store used, the exact product the customer purchased, the price paid, whether they bought in discount, the time of the day, and so forth.[3] We propose leveraging those (already existing) data and extract what we call "acquisition characteristics" from each customer's first transaction.[4] We contend that these acquisition characteristics/choices can be informative about underlying customer differences which can be predictive of customers behavior in the future. Because these data are also available for customers with longer tenure with the company, the firm would be able to uncover the (subtle) relationships between the choices observed at the moment of acquisition and the customer behavior down the road.

3.2 Augmenting cold start data with acquisition characteristics

Considering the retailer introduced above, who is trying to make inferences about its customers right after they have been acquired. A natural first step for the analyst would be to select a handful of variables collected at the acquisition moment (e.g., channel of acquisition) and use existing models to relate those characteristics to future demand (e.g., Chan et al. 2011). The caveat of doing so is that merely few variables might not fully capture the richness of the acquisition data, and the level of personalization would likely be limited as these few variables only capture a coarse representation of customers' heterogeneity. We propose to fully augment the acquisition data to broaden the amount of information that would (potentially) be linked to future behavior, therefore increasing the chance to solve the cold start problem.

[3]Note that the amount of data collected by firms also include data *prior* to the moment of acquisition. For example, e-retailers collect information via cookies, which could identify which customers have visited the website previously (yet, not making a purchase). When available, those data can be included in the exact same fashion as the acquisition characteristics. For simplicity, we denote "acquisition" data to all information available to the firm at the moment of acquisition, acknowledging that such data could also incorporate actions the customer performed before their first transaction.

[4]In theory, the data could also be augmented with characteristics of the second, or third transaction, for customers who are repeat buyers. However, we only use the first transaction because that is the data that *every* customer — just acquired and existing users — have in common, which will be the key to make inferences about recently-acquired customers. Adding information about each later transactions might add precision to the individual-level inferences of repeat users, but not necessarily to the inferences of recently-acquired customers, which is the main focus of this paper.

Specifically, using the (existing) data from each first transaction, we propose to augment cold start data with three types of acquisition variables: *transaction characteristics* (e.g., channel, price paid, holiday season) and *product characteristics*[5] (e.g., product category, package size), which are easily extracted from the transactional database, and *shopping basket (latent) representation*. The latter type of data aims to capture the "nature" of products that the customer purchased, above and beyond what the standard (observed) product categories represent. Our premise is that the nature of products purchased can signal the type of customer who purchases those. For example, in the market of cosmetics, certain ingredients or aroma characterize lines of products. It is possible that customers who discover the brand by buying products of certain "nature" are similar in they way they behave in the future. Because such information is not readily available from the firm's database, we need a method to encode the information embedded in each product, to then aggregate it at the basket level.[6]

Previous literature has used different methods to encode such information, from human coding based on full description of the product, to machine learning approaches that apply textual analyses to the description of products, or that leverage co-occurrence of products in basket data to create measures of similarity across products (e.g., Jacobs et al. 2016; Ruiz et al. 2017; Kumar et al. 2020; Chen et al. 2020). We take the latter approach and leverage the transaction data from anonymous customers to create continuous multidimensional representations of products, called product embeddings, that capture the nature of the product. Specifically, we create a co-occurrence matrix based on the composition of shopping baskets — i.e., which SKUs are purchased together — and implement *word2vec* (Mikolov et al. 2013), a machine learning approach widely used for natural language processing, to map each item to a multi-dimensional vector that captures similarities across products. This exercise is similar to creating a perceptual map from association data (Netzer et al. 2012) in which the co-occurrence of products in a basket is used as proxy of association between two products. (See Appendix A for all the details about how we process the transaction data and create the product embeddings using the *word2vec* algorithm.) Once we represent each product by

[5] Acquisition variables are constructed from the whole first transaction, which might include one or multiple products. That is, the *product characteristics* are summary statistics from the collection of products purchased on the first transaction.

[6] One alternative to this solution would be to include a dummy variable per (available) SKU. This approach would be straight forward in business contexts where the product space is small. However, when the firm offers a large selection of items or SKUs — as it is the case for most retailers — the vector of dummy variables would be too sparse to capture similarities among baskets and thus would prevent any model to learn across customers. For those cases, we recommend using a lower-dimensional vector representing the product space, as we do in this research.

a continuous vector, we can easily characterize the first purchase of any customer by computing moments of the product vectors in that basket.

In sum, using the transactional data already collected by the firm, one can easily augment each customer's data with a high-dimensional vector that captures a wide variety of acquisition characteristics including details about the first transaction as well as the type of products purchased.[7]

3.3 Predictive power of augmented data

A natural question to ask is: Do acquisition characteristics carry information about future behavior? While this is an empirical question, we present preliminary evidence from our empirical application that these augmented acquisition characteristics in turn explain differences in subsequent demand behavior across customers. To do so, we select customers who have been with the company for at least 15 months and relate their total number of repeat purchases during those 15 months with their (augmented) acquisition characteristics. We explore the relationship between individual acquisition characteristics and future transactions (Figure 1), as well as possible interactions among acquisition variables in their association with future demand (Figure 2).

– Insert Figures 1 and 2 here –

Indeed, acquisition characteristics are predictive of customers future transactions. Consistent with common belief in the industry (e.g., Artun 2014; RJMetrics 2016), customers that were acquired during the holiday season are less valuable to the firm, as we find that they are less likely to transact in the future. On the other hand, customers who bought using discounts on their first transaction generally buy more during the next 15 months than customers who did not. A similar pattern exists for customers who bought a recently-introduced product on their first transaction, and those who purchased products from the hair care category. Interestingly, this model-free analysis also suggest that some of these relationships are likely to be non-linear. For example, looking at average price paid per item, customers in the lowest quartile (Q1) tend to buy less frequently in their first 15 periods than all other customers. Similar non-linear relationships appear for the number of units and the total amount of the ticket.

Interesting patterns also emerge in Figure 2. On the left, we group customers on whether they were acquired during the winter holiday season, coupled with whether they purchased travel-

[7] In our empirical application this vector has 31 dimensions. Further details are presented in Section 5.

size products. We find that purchasing travel-size products moderates the relationship between being acquired during the holidays and the future number of transactions. Turning to the figure on the right, we observe that purchasing a discounted product on the first transaction signals lower value *only* if such a purchase did not include a new product. Taken together, these results present evidence of a relationship between acquisition characteristics and future transactions, confirming that augmenting cold start data with acquisition characteristics incorporates relevant information to infer customers' differences.

Nevertheless, this simple analysis is insufficient for solving the cold start problem of CRM as we would likely miss useful information from the data. First, it can only be performed for sub-sample of customers — those we observe for relatively long period of time (e.g., 15 months) — in order to have a fair comparison across customers over the same number of periods. Second, this type of analysis examines each variable independently (Figures 1), at most allowing for single interactions (Figure 2). Given that the goal is to extract relevant interrelations in high-dimension cold start data, it will be more effective (and efficient) to examine these interrelations collectively, while allowing for flexible relationships among the variables. Furthermore, the model-free analysis does not shed any light about customers' response to marketing actions. These results indicate that "holiday" customers are less likely to transact again. However, are they more/less sensitive to the firm's communication? How strongly will they react product introductions? A model would be certainly necessary to effectively extract the information from the acquisition characteristics to predict differences in transaction propensities *as well as* in responsiveness to marketing actions. Before presenting our modeling framework, we describe the methodological challenges that such a model should overcome.

3.4 Modeling challenges

Our solution to overcome the cold start problem ultimately depends on the ability of the model to extract the information hidden in the augmented data that is predictive of future behavior. Naturally, increasing the dimensionality of the acquisition data increases the chances of adding (at least potentially) information that will be relevant to infer customer differences down the road. However, expanding the dimensionality of the acquisition data also adds methodological challenges.

First, several of those augmented variables are likely to be irrelevant. Many of the behaviors observed in the first purchase are likely to be random and not systematically related with how customers will behave in the future. Second, some of these augmented data are multiple signals

from the same underlying behaviors, implying that much of those data would be redundant. For example, a price-conscious customer may purchase a set of travel-sized, cheap products that are discounted. Although, the variables price and discount capture different types of information (e.g., a discounted product may still be an expensive one), these variables are clearly correlated as they are both signals of this customer's preferences for inexpensive products. Moreover, if one also were to include latent representations of the products bought, these representations would likely correlate with prices and with how frequently they are discounted; adding to the redundancy already present among augmented variables. Taken together, these characteristic suggest that it is likely that cold start data would have low "signal-to-noise" ratio, increasing the difficulty of recovering the relationships between acquisition characteristics and future behavior.

Importantly, the underlying relationships between acquisition variables and future demand is unknown. As indicated by the early exploration of the data (Figures 1 and 2), those relationships are unlikely to be linear. It is unrealistic to recommend that a firm would explore all possible interactions and non-linear specifications among their augmented acquisition characteristics, and is especially cumbersome when also interested in customers' response to marketing actions. Moreover, increasing the dimensionality of the augmented data only emphasizes this challenge as it would increase the number of potential non-linear relationships and interactions among acquisition variables. Another potential limitation of increasing the dimensionality of the acquisition variables is that some variables might be missing for some customers. Missing observations present challenges to estimate models that use those missing variables as covariates as they require imputation methods — cumbersome for high-dimensional spaces — or deletion of customers (or variables) from the data — which directly reduces the amount of information, defeating the purpose of the data augmentation step.

In this research, we propose a modeling framework that overcomes all these issues at once. We combine a flexible demand specification (such that can be applicable to a wide rage of marketing contexts) with state-of-the-art machine learning methods (addressing nonlinearities and data redundancy) within a Bayesian framework (that extract signals from the acquisition characteristics while handling missing data). The resulting modeling framework is a flexible probabilistic machine learning model that links the individual-level parameters governing customer's future behavior (e.g., transaction propensities, sensitivity to marketing actions) with a latent representation of the behaviors/choices observed at the moment of acquisition. This modeling approach seamlessly captures

flexible relationships among variables (linear and non-linear) without the need to pre-specify those relationships a priori. Moreover, the model explicitly accounts for interrelations among acquisition data which helps regularize the flexible model avoiding overfitting.

These benefits will become clear as we build and validate the model in the next section, where we also show how this approach dominates existing alternatives that addressed some (but not all) modeling challenges. For example, we compare it with a standard hierarchical Bayesian model with acquisition characteristics are included as covariates; a fully hierarchical model where acquisition characteristics and demand are jointly correlated using a multivariate Gaussian distribution; or a (supervised) Bayesian PCA that aims to reduce dimensionality of acquisition characteristics as well as demand parameters.

Finally, as we show in our empirical application that, if we simplify the task and only consider the model's ability to predict future transactions, our modeling approach performs at the level of traditional machine learning (ML) approaches such as a random forest and a deep neural network (proven to capture non-linear relationships very well). Our model stands out in comparison with these ML benchmarks in two main ways. Methodologically, it can be easily be combined with multiple demand specifications, as well as allows for missing observations in acquisition characteristics without relying on data imputation. Practically, our model provides inferences beyond predictions of future transactions, enabling marketers to get insights about customer heterogeneity in preferences and in sensitivity to marketing actions.

4 MODELING FRAMEWORK

4.1 Model development

Our modeling framework — which we call "First Impression Model" (FIM) — comprises three main components: (1) the *demand model*, main outcome of interest to the firm, which could include customers transactions, purchase volume, etc., (2) the *acquisition model*, capturing all customer outcomes that are observed to the firm at the moment of acquisition, and (3) the *probabilistic model* that links the underlying customer parameters influencing these two types of behaviors through hidden traits.

4.1.1 Demand model

We start by assuming a general model for demand, suitable for different specifications, and parametrized using individual-level parameters and population-level parameters. Specifically, for

customer i at period t, we denote

$$p(y_{it}|\widetilde{\mathbf{x}}_{it}^{y}, \boldsymbol{\beta}_i^{y}, \boldsymbol{\sigma}^{y}) = f^{y}(y_{it}|\widetilde{\mathbf{x}}_{it}^{y}, \boldsymbol{\beta}_i^{y}, \boldsymbol{\sigma}^{y}) \qquad i \in \{1, \ldots, I\},\ t \in \{1, \ldots, T_i\}, \tag{1}$$

where I represents the total number of customers, T_i denotes the number of periods since the customer was acquired, $\boldsymbol{\beta}_i^{y}$ is a vector containing customer i's individual-level parameters, the vector $\boldsymbol{\sigma}^{y}$ contains the parameters that are common across customers, and $\widetilde{\mathbf{x}}_{it}^{y}$ includes the observed covariates for customer i at period t. Finally, $f^{y}(\cdot)$ is the pdf/pmf for outcome y_{it}; for example, if the outcome of interest is purchase incidence, we would specify $p(y_{it} = 1) = \text{logit}^{-1}\left[\mathbf{x}_{it}^{y\prime} \cdot \boldsymbol{\beta}_i^{y}\right]$.[8]

4.1.2 Acquisition model

We denote A_i the vector of characteristics that are collected at the moment of acquisition, and a_{ik} the k'th component/behavior (e.g., did the customer purchase a discounted product on their first transaction?). These acquisition characteristics are likely to be influenced by individual-level parameters (e.g., does this customer have the tendency to buy on discount?) but also by the market conditions at the moment of acquisition (e.g., was the company running heavy discounts during that period?). We account for these effects by modeling the acquisition characteristics as a probabilistic outcome, rather than as an input/covariate to the model. Note that we do not model acquisition per se, i.e., *whether* the customer is acquired or not. Rather, we model the characteristics of the first purchase given that the customer was acquired. This approach is adequate in this case because the goal of the model is to allow the firm to manage acquired customers, and not to alter the marketing mix that drive the acquisition process to change the pool of acquired customers.

Modeling the acquisition characteristics as an output not only allows us to control for the time-varying factors that shift demand at the moment of acquisition, but also allows for a flexible modeling specification of the latent traits that overcome challenges such as redundancy, irrelevance of variables, and missing data commonly encountered in the firm's database. (We discuss these challenges in Section 4.1.3). Specifically, we denote

$$p(a_{ip}|\beta_{ip}^{a}, \boldsymbol{\sigma}_p^{a}, \mathbf{x}_{m(i)\tau(i)}^{a}) = f_p^{a}(a_{ip}|\beta_{ip}^{a}, \boldsymbol{\sigma}_p^{a}, \mathbf{x}_{m(i)\tau(i)}^{a}) \qquad i \in \{1, \ldots, I\},\ p \in \{1, \ldots, P\}, \tag{2}$$

[8] The model can easily be adapted to other forms of demand (e.g., continuous demand, count) and extended to dynamic specifications such as latent attrition models. For the latter, one could define (1) as a state-space model (e.g., a hidden Markov model) with state variable s_{it} and $p(y_{it}, s_{it}|y_{i1:t-1}, s_{i1:t-1}) = p(y_{it}|s_{it}) \cdot p(s_{it}|s_{it-1})$. We would implement such a model by having two individual level vectors, $\boldsymbol{\beta}_i^{yq}$ and $\boldsymbol{\beta}_i^{ye}$, as well as two population level vectors, $\boldsymbol{\sigma}^{yq}$ and $\boldsymbol{\sigma}^{ye}$, that would govern transitions among the hidden states and emissions in a state, respectively. We would substitute (11) for $p(y_{it}, s_{it}|y_{i1:t-1}, s_{i1:t-1}, \mathbf{x}_{it}^{y}, \boldsymbol{\beta}_i^{y}, \boldsymbol{\sigma}^{y}) = p(y_{it}|s_{it}, \mathbf{x}_{it}^{y}, \boldsymbol{\beta}_i^{yq}, \boldsymbol{\sigma}^{yq}) \cdot p(s_{it}|s_{it-1}, \mathbf{x}_{it}^{y}, \boldsymbol{\beta}_i^{ye}, \boldsymbol{\sigma}^{ye})$, where $\boldsymbol{\beta}_i^{y} = \left[\boldsymbol{\beta}_i^{yq} \quad \boldsymbol{\beta}_i^{ye}\right]$, and $\boldsymbol{\sigma}^{y} = \left[\boldsymbol{\sigma}^{yq} \quad \boldsymbol{\sigma}^{ye}\right]$ be the parameters of the demand model.

where P is the number of different types of behaviors collected at acquisition, β^a_{ip} is an individual level parameter that reflects tendency to observe such a behavior when customer i is acquired, $\boldsymbol{\sigma}^a_p$ denotes a vector of parameters that are common across customers, and $\mathbf{x}^a_{m(i)\tau(i)}$ comprises the set of market-level covariates, with $m(i)$ indicating the market customer i belongs to, and $\tau(i)$ denoting the time period at which the customer was acquired.

The term $f^a_p(\cdot\,|)$ is the pdf/pmf of a distribution to model acquisition behavior p. Note that some of these behaviors will likely be binary[9] (e.g., whether the customer was acquired online), in which case we specify $\boldsymbol{\sigma}^a_p = \left[\mathbf{b}^a_p\right]$ and model p as

$$p(a_{ip} = 1) = \text{logit}^{-1}\left[\beta^a_{ip} + \mathbf{x}^a_{m(i)\tau(i)}{}' \cdot \mathbf{b}^a_p\right]. \tag{3}$$

For continuous acquisition variables (e.g., total amount spent in the first transaction) we define $\boldsymbol{\sigma}^a_p = \left[\mathbf{b}^a_p, \sigma^a_p\right]$ and model p as

$$p(a_{ip}) = \mathcal{N}(\beta^a_{ip} + \mathbf{x}^a_{m(i)\tau(i)}{}' \cdot \mathbf{b}^a_p, \sigma^a_p), \tag{4}$$

specification that can be easily adjusted for multivariate outcomes as we do with some acquisition variables in our empirical application.

All of these types of variables are easily incorporated by adjusting the acquisition model accordingly. We define $\boldsymbol{\beta}^a_i = \left[\beta^a_{i1} \quad \dots \quad \beta^a_{iP}\right]$ and $\boldsymbol{\sigma}^a = \left[\boldsymbol{\sigma}^a_1 \quad \dots \quad \boldsymbol{\sigma}^a_P\right]$ as the full set of individual- and population-level vectors of acquisition parameters, respectively.

Note that we only have one observation per individual and behavior. Hence, in theory, having an individual-level parameter β^a_{ip} could completely capture the residual variance of a_{ip} that is not systematically explained by the market-level factors (as in a regression with individual random effects but only one observation per individual). However, because we model demand and acquisition jointly, our model will balance fitting each acquisition behavior a_{ip} with fitting the other acquisition characteristics, as well as fitting demand, with a reduced set of individual factors or traits. Therefore, the individual level parameters β^a_{ip} will not have full flexibility to accommodate perfectly to the behavior a_{ip}. Rather, these parameters will capture the residual variance that is correlated with the rest of the acquisition variables and with the demand model. This remark

[9] Categorical acquisition behaviors can easily be incorporated using a categorical distribution with a softmax link function.

will become clearer when we specify the relationship between the individual-level demand and acquisition parameters, $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$, as we do in the next section.

Finally, the term $\mathbf{x}^a_{m(i)\tau(i)}$ controls for the overall marketing intensity that a yet-to-be-acquired customer might have been exposed to in a particular market at the moment of acquisition. For example, if there is a strong promotional activity in market m in period t, one would likely observe a higher-than-usual share of discounted products among the acquisition characteristics, not only driven by the customers' propensity to buy on discount, but also by the fact that the majority of products were discounted.[10] Accordingly, we want to capture this systematic shift in the acquisition characteristics as a market-related shift and not as a customer-driven shift, and therefore set $\mathbf{b}_p^a$ common across customers.

4.1.3 Linking acquisition and future demand: Deep probabilistic model

We use a deep exponential family (DEF) component (Ranganath et al. 2015) to relate demand and acquisition parameters hierarchically, through hidden layers. We chose such specification because of its hierarchical nature — allowing the model to identify/extract individual-level traits that affect both acquisition and future demand — and because the presence of multiple layers facilitates the reduction of dimensionality while accommodating a wide range of possible relationships between acquisition and demand variables. Furthermore, one important characteristic of DEFs is that the latent variables are distributed according to distributions that belong to the exponential family (e.g., Gaussian, Poisson, Gamma), making them a good candidate to model the wide range of data types encountered in the firm's database. Finally, DEFs also enjoy the flexibility of probabilistic models, allowing them to be easily incorporated in more complex model structures, as we do in this research. (See Appendix B for more details on DEFs.)

Turning our attention to our modeling challenge, the primary goal of our model is to infer the individual-level parameters $\boldsymbol{\beta}_i^y$. Therefore, we specify the DEF component such that the lowest level captures the individual-level traits that affect both the acquisition characteristics and future demand. Specifically, we define

[10] If the model did not control for these market-level conditions and the firm managed acquisition and retention efforts strategically, the interrelations between acquisition characteristics and demand parameters obtained by the model could be spurious in the sense that they could be driven by the firm's actions and not by customers' underlying preferences.

$$\boldsymbol{\beta}_i^y = \boldsymbol{\mu}^y + \mathbf{W}^y \cdot \mathbf{z}_i^1 \tag{5}$$

$$\boldsymbol{\beta}_i^a = \boldsymbol{\mu}^a + \mathbf{W}^a \cdot \mathbf{z}_i^1 \tag{6}$$

such that the individual level parameters, $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$ are a (deterministic) function of mean parameters, $\boldsymbol{\mu}^y$ and $\boldsymbol{\mu}^a$, and individual deviations from this mean which are a function of the lower layer vector $\mathbf{z}_i^1$, and weight matrices $\mathbf{W}^y$ and $\mathbf{W}^a$. Similarly as in a Bayesian Principal Components Analysis (Bayesian PCA) model (Bishop 1999), the vector $\mathbf{z}_i^1$ captures the individual level traits that explain jointly demand and acquisition behavior. The weight matrices $\mathbf{W}^y$ and $\mathbf{W}^a$ capture how each one of these traits manifests in both demand and acquisition characteristics respectively.

We assume that each component k of the lower layer, z_{ik}^1, is distributed Gaussian with mean $g(-\mathbf{w}_k^{1\prime} \cdot \mathbf{z}_i^2)$, and variance 1,

$$p(z_{i,k}^1 | \mathbf{z}_i^2, \mathbf{W}^1) = \mathcal{N}\left(z_{i,k}^1 | g\left(-\mathbf{w}_k^{1\prime} \cdot \mathbf{z}_i^2\right), 1\right) \qquad k \in \{1, \ldots, N_1\}, \tag{7}$$

where N_1 is the dimension of the lower layer, $\mathbf{z}_i^2$ is the top layer vector (of dimension $N_2 < N_1$)[11], $g(x) = \log\left(\log\left(1 + \exp(x)\right)\right)$ is the log-softplus function (Ranganath et al. 2015),[12], and $\mathbf{W}^1$ is the weight matrix that links the upper and lower layers. The upper layer captures higher-level traits (resembling the structure of neural networks), while allowing for non-linear interrelations between the traits in the lower level $\mathbf{z}_i^1$. The dependence between the top components and the lower layer components is a key aspect of the DEFs that enables the model to capture interrelations among the lower layer components. The dependence between lower layer and higher layer is regularized through sparse gamma priors on $\mathbf{W}^1$ inducing the model to pick up the relevant correlations among those traits (see Appendix C). Moreover, the non-linear relationships are captured by the non-linear link function $g(\cdot)$, which relates the higher-level traits with the lower-level traits that manifest in demand and acquisition. Finally, we model the upper layer using a standard Gaussian distribution,

$$p(z_{i,k}^2) = \mathcal{N}\left(z_{i,k}^2 | 0, 1\right) \qquad k \in \{1, \ldots, N_2\}. \tag{8}$$

[11] In theory, N_2 could be larger than N_1 but such a model would not necessarily reflect patterns in data as information would be lost going from the upper layers of the DEF to the lower layers of the DEF. Ranganath et al. (2015) only estimate models with decreasing dimensions of upper layers.

[12] In Stan, the softplus function, defined as $f(x) = \log(1 + \exp(x))$, can be computed using `log1p_exp(·)`.

To sum, we link the individual-level demand and acquisition parameters using a DEF component of two Gaussian layers, $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$. The model could easily accommodate more layers (e.g., Ranganath et al. 2015, use up to 3 layers, $L \leqslant 3$, in their empirical applications).[13]

4.1.4 Dimensionality of the DEF component

At first glance, the choice of the layers dimensions N_1 and N_2 may seem cumbersome. On the one hand, high values of N_1 and N_2 increase the computational burden of the inference procedure, which is not desirable. On the other hand, a model with low values for N_1 and N_2 may miss relevant associations that are needed to infer customers' parameters. In the extreme, if the number of components of the lower layer, N_1, is set to one, the model would only learn a single trait to describe the variation across all parameters, which will fail to capture the heterogeneity in the demand parameters, and their (potentially non-linear) relationships with acquisition characteristics. Similarly, if the number of components of the higher layer, N_2, is set to zero, the model would be stripped away from the non-linear function $g(\cdot)$ that allows the model to capture non-linear relationships between demand and acquisition parameters.

Similar to other latent-space models, one could test all possible combinations of N_1 and N_2 (increasing in magnitude) and choose the optimal values using cross-validation. Such exercise is certainly required when using Maximum Likelihood Estimation, as more flexibility in a model leads to over-fitting following the classical bias-variance trade-off, and therefore poor performance in holdout samples. However, when using Bayesian inference, this exercise would not only be computationally very costly, but also unnecessary, provided that adequate priors such as spike-and-slab or sparse-gamma (Karaletsos and Rätsch 2015; MacKay 1995; Neal 2012) are used to induce regularization in the parameters governing the weights that activate the traits. Using such priors ensures that a trait only manifests in a particular variable if the improvement in fit is substantial; otherwise, that trait is "shut down" by the prior (Ranganath et al. 2015).

Therefore, our approach to specifying the dimensionality of the model is to set a "large enough" number of traits to ensure that all relevant traits are recovered, while using sparse priors to ensure that the model only activates the relevant traits, thus avoiding overfitting the data.

[13] We follow the specifications from Ranganath et al. (2015), where the model is estimated using, at most, 3 layers ($L \leqslant 3$). In that paper, the model is trained on two large text corpora (5.9K and 8K terms), two matrix factorization tasks on a movie ratings dataset (50K users and 17.7K movies), and a click dataset (18K users and 20K documents). All of these datasets are considerably larger than our data (both in the simulations and in the empirical application). Furthermore, Tables 2 and 3 from Ranganath et al. (2015) do not show consistently whether $L = 3$ is better than $L = 2$. As a result, we use $L = 2$ as it is the smallest configuration that allows for non-linear relationships.

Specifically, we use sparse Gamma priors for $\mathbf{W}^1$ and hierarchical Gaussian automatic relevance determination (ARD) priors for $\mathbf{W}^y$ and $\mathbf{W}^a$, both of which are spike-and-slap-like priors that have shown to perform well on feature selection (e.g., Bishop 2006; Kucukelbir et al. 2017). These priors ensure that once a trait is "shut down," adding more traits (i.e., increasing N_1 or N_2) would just add irrelevant traits with weights all being close to zero, not affecting the performance of the model. (See Appendix C.1 for details about these priors.)

The added benefit of inducing regularization through the priors is that we can look at the posterior estimates of the variances of the weights ($\mathbf{W}^y$, $\mathbf{W}^a$, and $\mathbf{W}^1$) to evaluate whether the number of dimensions (N_1 and N_2) are sufficient to represent the data. Examining N_1 is straightforward as the model parameter α^1 captures the variance of the lower layer traits. Regarding N_2, while there is not one specific parameter capturing the relevance of the upper layer traits, we can compute a pseudo-α^1_m for each upper trait m using the components of the weight matrix $\mathbf{W}^1$ that map to relevant lower level traits (see Appendix D for details). Finally, examining the posterior estimates of α^1 and pseudo-α^1_m — and observing that some traits have been "shut down" by the model — we corroborate whether N_1 and N_2 are "large enough" for any specific dataset.

These insights are further developed in Appendix D.7 where we explore the dimensionality of the DEF component by analyzing the results of estimating the FIM on simulated data, where we know how many traits are needed. There we show how the performance of the model remains largely unchanged by the additional dimensions (on either N_1 or N_2) after the relevant number of traits are accounted for. We also show how the posterior estimates of the variances of the weights (α^1 and pseudo-α^1_m) are diagnostic of relevant and non-relevant traits.[14]

To sum, we take a hybrid approach to model selection in which we make sure that the number of pre-specified dimensions is large enough — phenomenon that can be validated from the model parameters — while we rely on the priors of the model to ensure regularization.

4.1.5 Bringing it all together

We briefly discuss how each part of the model contributes to the desired goals and how the FIM compares with alternative approaches to overcome the cold start problem. In essence, the model comprises a demand and an acquisition model, whose individual-level parameters are projected

[14]The posterior distribution of $\boldsymbol{\alpha}$ and $\mathbf{W}^1$ from real world data sets would not display as clear cut distinction between those traits that are meaningful and those that are not compared to our simulation analyses. We come back to this point when discussing the specification of the FIM for our empirical application.

into a lower-dimensional space through a two-layered DEF component. The lower layer of the DEF captures the relevant associations among the individual-level parameters while reducing the dimensionality of those vectors. An alternative approach to link the acquisition and demand parameters could be through using traditional full hierarchical Bayesian priors (e.g., multivariate Gaussian). Such an approach would assume that all individual-level parameters ($\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$) are distributed jointly according to a flexible multivariate distribution which parameters capture all the potential correlations among the variables. However, this full hierarchical approach would require require the model to estimate a very high-dimensional correlation matrix which can become computationally expensive, especially as the number of acquisition variables increases. On the contrary, because the FIM includes ARD priors for the lower layer of the DEF, the model only allows for "relevant" associations to emerge, automatically reducing the dimensionality of the individual-level parameters. This is a desirable feature not only because the number of acquisition variables could be large, but also because some of the acquisition variables are likely to be correlated among each other.[15]

The upper layer of the DEF, and in particular, the non-linear link function $g(x)$ that relates the higher-level traits with the lower-level traits allows the model to capture a wide range of relationships — linear and non-linear — among the variables of interests. A simpler specification of the FIM would be one that does not incorporate the second layer and therefore imposes linear relationships among the individual parameters. Such a nested version of the FIM would be equivalent to a "supervised" factor analysis or Bayesian PCA where the latent traits are extracted from the acquisition variables as well as from the demand model. The limitation of such a (nested) approach is that the model would lose its accuracy at forming first impressions the moment the assumption of linearity does not hold, either because acquisition variables relate to demand parameters in a non-linear way, or when two (or more) acquisition variables interact in their relationship with the demand parameters. As we show in Section 4.4, our FIM specification (that includes the second layer) captures several forms of relationships (including linear, interaction effects, and maximum function) without the need for specifying those relationships a priori. This is a very desirable prop-

[15] An alternative but similar specification for the model could be a two-step approach that first reduces dimensionality among the acquisition variables (i.e., connecting z_i^1 to $\boldsymbol{\beta}_i^a$) and then connects those factors with future demand. We choose to connect the lower level of the DEF model with both components jointly in order to be robust to the possibility that the residual variance of the acquisition variables not explained by the main factors of the first step is predictive of demand behavior; and to inform the choice of factors that are predictive of demand behavior, as in supervised topic models (Mcauliffe and Blei 2008), and therefore, to overcome redundancy and irrelevance of acquisition variables simultaneously.

erty of the model because managers/reseachers/data scientists generally do not know the exact form of the relationships among the variables of interest.

Finally, a different approach to overcome the cold start problem could be to simply specify the individual-level demand parameters ($\boldsymbol{\beta}_i^y$) as a direct function of the acquisition variables (A_i^y). Such a specification would resemble a typical demand model with interactions, or a multi-level (hierarchical) model in which $\boldsymbol{\beta}_i^y$ are a function of the observed A_i and some population distribution (Rossi et al. 1996; Allenby and Rossi 1998; Ansari and Mela 2003; Chan et al. 2011). While a linear model is attractive for its simplicity and ease of interpretation, if the underlying relationships between the acquisition variables were not linear (or did not follow the specified relationship, due to variable transformation), the model will fail at inferring individual-level demand parameters for newly-acquired customers with certain level of accuracy. While non-linearities could be captured by higher order interactions, such an approach becomes intractable when the parameter space for the acquisition variables increases. In addition, specifying acquisition characteristics as covariates would require data imputation or data augmentation techniques in order to handle missing observations. In contrast, our modeling framework does not require those types of techniques because we model acquisition characteristics as an outcome.

– Insert Figure 3 here –

To conclude, Figure 3 shows the graphical model for the FIM, connecting all the individual components. We propose a model of demand and acquisition characteristics where the individual-level parameters of each of these sub-models are projected into a lower-dimension space via a DEF component. The specification of the demand sub-model is general such that the modeling framework can be applicable to a wide range of business contexts. The sub-model for acquisition characteristics enables the model to control for market conditions or firm-initiated actions that can potentially shift the type of customers that are acquired over time. If these shifts were not captured, the model would not be able to differentiate market conditions from customer underlying preferences. Regarding the DEF component, there are three main benefits of using a two-layered DEF to connect both types of individual-level parameters. First, the model provides dimensionality reduction, avoiding the curse of redundancy and irrelevance of variables among the acquisition variables. Second, the model allows for flexible relationships (e.g., non-linear relationships) among the model components. Third, the model can incorporate acquisition characteristics with missing observations, as these are modeled as outcomes which are easily handled using a Bayesian estimation framework. These

benefits will become clearer in Sections 4.4 through 5, when we compare the predictive accuracy of the FIM with that of several alternative specifications.

4.2 Estimation and Identification

We estimate the model using full Bayesian statistical inference with MCMC sampling. We sample the parameters from the posterior distribution which is proportional to the joint,[16]

$$
\begin{aligned}
p\left(\{\mathbf{z}_i^1, \mathbf{z}_i^2\}_{i=1}^I, \mathbf{W}^y, \mathbf{W}^a, \mathbf{W}^1, \boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}_a, \{y_{i1:T}, A_i\}_i\right) = & \left[\prod_{i=1}^{I}\prod_{t=1}^{T_i} p(y_{it} \mid \mathbf{x}_{it}^y, \mathbf{z}_i^1, \mathbf{W}^y, \boldsymbol{\mu}^y, \boldsymbol{\sigma}^y)\right] \\
& \cdot \left[\prod_{i=1}^{I} p(A_i \mid \mathbf{x}_i^a, \mathbf{z}_i^1, \mathbf{W}^a, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^a, \mathbf{b}_a)\right] \cdot \left[\prod_{i=1}^{I} p(\mathbf{z}_i^1 \mid \mathbf{z}_i^2, \mathbf{W}^1)\right] \cdot \left[\prod_{i=1}^{I} p(\mathbf{z}_i^2)\right] \\
& \cdot p(\mathbf{W}^y, \mathbf{W}^a, \mathbf{W}^1, \boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}_a). \qquad (9)
\end{aligned}
$$

In particular, we use the No U-Turn Sampling Hamiltonian Monte Carlo algorithm, implemented in the Stan probabilistic programming language (Carpenter et al. 2016; Hoffman and Gelman 2014), which is freely available, and facilitates the use of this model among researchers and practitioners.[17]

Regarding the identification of the model parameters, the demand and acquisition parameters ($\boldsymbol{\beta}_i^y$, σ^y, $\boldsymbol{\beta}_i^a$ and σ^a) are identified, provided the functional forms described in (1) and (2) are well specified. On the contrary, not every single parameter of the DEF component is fully identified. [*Lower layer*] The parameters that link the lower layer of the DEF with $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$ are identified up to a rotation, similar to a traditional factor analysis model. Specifically, the scales of the lower layer trait ($\mathbf{z}_i^1$) and weights ($\mathbf{w}^y$ and $\mathbf{w}^a$) are identified through the priors scales. Small rotations are identified by the sparsity of the ARD priors (see Appendix C for details) —these priors favor the activation of fewer traits, avoiding the rotation of a large trait into smaller ones. Orthogonal rotations are not fully identified due to possible sign change in traits and label switching.[18] However, we can obtain behavioral insights from the lower layer of model — e.g., what trait(s) are most predictive of specific behaviors — by carefully rotating the lower layer traits and weights parameters across draws to maintain a consistent interpretation of these parameters (see Appendix E for details). [*Top layer*] The top layer of the DEF and the parameters that link the top and lower layer are not identified. This is similar to deep neural networks, in which the lower layer is a combination

[16] All details about the prior distribution $p(\mathbf{W}^y, \mathbf{W}^a, \mathbf{W}^1, \boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}_a)$ are presented in Appendix C.2.

[17] The code is available from the authors.

[18] Note that the lower traits themselves are not orthogonal by design, as they are related through the upper layer.

of the values of the upper layer and the weights linking them. In our model specification, this translates to the value of the top layer ($\mathbf{z}_i^2$) not being identified as different combinations of $\mathbf{z}_i^2$ and $\mathbf{w}^1$ could generate the same value for $\mathbf{z}_i^1$. Most importantly, this lack of identification in the DEF component does not preclude the model from uniquely identifying the individual-level demand parameters $\boldsymbol{\beta}_i^y$ (as corroborated in Sections 4.4 and 5), which is the main goal when overcoming the cold start problem.

4.3 Model inferences for newly acquired customers

Recall that the main purpose of the model is to assist firms in the task of making inferences about how individual customers will behave in the future (e.g., how they will respond to marketing interventions), based on the observed behaviors at the moment of acquisition. Intuitively, that process would works as follows: A new customer is acquired and the firm observes their behaviors at the moment of acquisition. At that point, and given the firms' prior knowledge of the market (i.e., the model parameters and market conditions), the firm makes an inference about that particular customer's latent traits, which are then used to infer the individual-level parameters that will determine their demand (e.g., how likely is it that the customer will purchase in the future, their responsiveness to marketing interventions).

More formally, we want to infer $p(\boldsymbol{\beta}_j^y|A_j, \mathcal{D})$ for customer j who was not in the training sample, for whom we observe acquisition characteristics A_j, and where $\mathcal{D} = \{y_{i1:T_i}, A_i\}_{i=1}^I$ comprises the calibration data. Denoting $\Theta = \{\boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \mathbf{W}^y, \mathbf{W}^a, \mathbf{W}^1, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}^a\}$ the population parameters and $\mathbf{Z}_j = \{\mathbf{z}_j^1, \mathbf{z}_j^2\}$, we can write $p(\boldsymbol{\beta}_j^y|A_j, \mathcal{D})$ by both integrating out over the parameters Θ and $\mathbf{Z}_j$, and using the factorization of the joint distribution provided in (9). That is,

$$
\begin{aligned}
p(\boldsymbol{\beta}_j^y|A_j, \mathcal{D}) &= \int p(\boldsymbol{\beta}_j^y, \mathbf{Z}_j, \Theta|A_j, \mathcal{D}) \cdot d\mathbf{Z}_j \cdot d\Theta \\
&= \int p(\boldsymbol{\beta}_j^y|\mathbf{Z}_j, \Theta, A_j) \cdot p(\mathbf{Z}_j|\Theta, A_j) \cdot p(\Theta|A_j, \mathcal{D}) \cdot d\mathbf{Z}_j \cdot d\Theta \\
&= \int_\theta \Big[\int_{\mathbf{Z}_j} p(\boldsymbol{\beta}_j^y|\mathbf{Z}_j, \Theta, A_j) \cdot p(\mathbf{Z}_j|\Theta, A_j) \cdot d\mathbf{Z}_j \Big] \cdot p(\Theta|A_j, \mathcal{D}) \cdot d\Theta \\
&\approx \int_\theta \Big[\int_{\mathbf{Z}_j} p(\boldsymbol{\beta}_j^y|\mathbf{Z}_j, \Theta, A_j) \cdot p(\mathbf{Z}_j|\Theta, A_j) \cdot d\mathbf{Z}_j \Big] \cdot p(\Theta|\mathcal{D}) \cdot d\Theta. \qquad (10)
\end{aligned}
$$

The last approximation suggests that if the number of customers in the calibration data is large, we can proxy the posterior of the population parameter with focal customer j by the posterior distribution obtained without the focal customer j. In other words, adding one more customer

would not significantly change the posterior of the population parameters. This approximation is very useful in practice because it allows us to draw from $p(\Theta|\mathcal{D})$ using the calibration sample, and draw the individual parameters of the focal customer j once this customer has been acquired, without the need to re-estimate the model to incorporate A_j. (See Appendix F for a description of the corresponding algorithm.)

4.4 Model performance

Before applying the new modeling framework to the empirical context, we need to demonstrate the accuracy of the model at inferring the individual-level parameters for newly-acquired customers. Because individual-level parameters are, by definition, unobserved, we perform this task using a simulation analysis in which we know the exact values of $\boldsymbol{\beta}_j^y$ and can therefore evaluate the model's ability at recovering the true parameters using (10). Unlike other simulation exercises, the goal of this analysis is *not* to confirm that the model can recover the (population) parameters. Rather, we use simulations to demonstrate that the proposed model is able to recover customers' individual-level parameters accurately, even when the data generating process for those individual-level parameters is not known, and possibly different from the modeling assumptions. In reality, marketers (and researchers) never know the exact relationship between acquisition characteristics and future demand parameters, therefore, having a flexible model that performs well in a variety of contexts is of critical importance. (We briefly describe the main aspects of the simulation design while including all details in Appendix D.)

We generate three scenarios for the underlying relationship between acquisition variables and demand parameters. In each scenario, customers are "endowed" with a set of demand parameters that follow a specific relationship with their observed acquisition characteristics, namely (1) *linear*, (2) *quadratic/interactions* (allowing the relationship between one acquisition variable and the demand parameters to vary depending on the value of other acquisition characteristics), and (3) *positive-part* (forcing the relationship between acquisition characteristics and demand parameters to be zero for low values of the acquisition characteristic). Given those individual-level demand parameters, customer transaction history is simulated for 2,200 customers. We use 2,000 customers to estimate the model, and the remaining 200 customers to evaluate the accuracy of the model at inferring demand parameters for newly-acquired customers. Specifically, only using the acquisition characteristics for these 200 customers, we use the model to infer their individual-level demand parameters, and compare those estimates with the true values.

We compare the performance of the FIM with that of three other specifications: (i) a HB-linear model, where individual demand parameters are specified as a linear function of the acquisition characteristics (this corresponds to the simulated data under the *linear* scenario), (ii) a full hierarchical model, where demand and acquisition parameters are jointly distributed according to a multivariate Gaussian distribution with a flexible covariance matrix, and (iii) a Bayesian PCA model. As discussed in Section 4.1.5, the Bayesian PCA model is a nested specification of the proposed FIM (in which the second layer does not exist) whereas the full hierarchical model and HB-linear specifications reflect alternative (simpler) ways in which past research has modeled these types of data. To measure the accuracy of each model, we compare the predicted posterior mean vs. the actual values for the demand parameters (both the intercept and the effect of the covariates) of the 200 out-of-sample customers. Table 4 includes the results for all models across all scenarios.[19] We also include the results of estimating a hierarchical Bayesian (HB) demand-only model in which acquisition characteristics are not incorporated, to have a reference of how much error one would obtain by simply predicting the population mean.

– Insert Table 4 here –

First, under a true linear relationship (Scenario 1), the FIM predicts the individual parameters as good as the benchmark models. The RMSE of the FIM is comparable to the benchmark models, and the R-squared is equal to the benchmark models. This result verifies that the FIM does not overfit the training data or, in other words, that the additional model complexity — even when not needed — does not hurt the accuracy of predictions for customers outside the calibration sample. Second, when the relationship among the model parameters is not perfectly linear (Scenarios 2 and 3), the FIM significantly outperforms the benchmark models in all dimensions. In particular, the R-squared of the FIM is higher than that of the benchmarks, demonstrating that the model is superior at sorting customers based on their demand parameters. Moreover, the RMSE for the FIM is substantially lower than that of the benchmarks, indicating that the proposed model predicts the exact magnitude of customer parameters (e.g., purchase probability, sensitivity to marketing actions) more accurately than any of the benchmarks. These results hold when we examine the model "at scale", when we significantly increase the amount of data collected by the firm and

[19] See Appendix D.3 for more details about the specification of the benchmark models and Appendix D.4 for details on the performance metrics.

also add standard regularization techniques (e.g., LASSO) to the benchmark models. (Please see Appendix D.8 for details.)

To help understand what drives the greater accuracy of these predictions, we further explore the results for Scenario 3 (when the true relationship is positive-part). The first row of Figure 4 shows the scatter plot of the predicted ($\widehat{\beta}^{y}_{j1}$) versus actual (β^{y}_{j1}) individual demand intercepts from each model, which displays the superior performance of the FIM, as detailed in Table 4. The second row of Figure 4 shows the predicted and actual demand intercepts as function of the first acquisition variable for each model. The blue dots show the true relationship between these two variables (i.e., positive-part) whereas the red dots correspond to the relationship estimated by the model. These plots evidence that the FIM can better recover the positive-part relationship between the acquisition variables and the demand parameters.[20]

– Insert Figure 4 here –

Finally, to better understand which aspect of the model is responsible for this accuracy of predictions, we compare the BPCA and the FIM model more closely, allowing both specifications to vary the dimensionality of their latent components. Such an analysis indicates that the presence of the second layer of the DEF component is contributing significantly to the improvement in accuracy for scenarios where the relationship is not linear. The results suggest that incorporating that second layer, even if specified with low dimensionality, allows the model to flexibly capture the non-linear relationship between acquisition and demand parameters. (Please see full details in Appendix D.6.)

To sum, these analyses demonstrate the effectiveness of the FIM at overcoming the cold start problem. We have shown that the FIM can accurately infer customer parameters using only acquisition data, even when such a model is not used to simulate the true parameters. While the benchmark models fail to form accurate inferences of newly-acquired customers when the underlying relationships among variables are not perfectly linear, the FIM is flexible enough to reasonably recover those parameters. This latter point is of great importance because in reality the researcher/analyst never knows the underlying relationships among variables. Therefore, having a

[20]Note that the model performance relies empirically on the predictive power of the acquisition variables on future behavior. In our simulation analyses, we tested the model performance when adding acquisition characteristics that were unrelated to future behavior and found no evidence of model overfit. Nevertheless, we did not explore whether the model would overfit when there is no predictive power among all acquisition variables.

flexible model able to accommodate multiple forms of relationships is crucial to accurately infer customers' parameters.

5 EMPIRICAL APPLICATION

5.1 Data and model specification

Our focal firm is an international retailer that sells its own brand of beauty and cosmetic products (e.g., skincare, fragrance, haircare).[21] Customers can only purchase the company's products via owned stores, either offline (the company owns "brick and mortar" stores across many countries) or online (with one online store per country). While the company is present in many countries, most marketing functions (e.g., promotional campaigns, product introductions) are centralized and therefore operations are very consistent across markets. Like most other companies, the focal firm records the transactions of all individual customers, along with other information about the CRM activities, such as direct marketing campaigns and email marketing activities.

5.1.1 Transactional data

We obtain individual-level transactions for registered customers in the six major markets — USA, UK, Germany, France, Italy, and Spain. We observe customers from the moment they make their first purchase (starting in November of 2010). At the point of purchase, customers are asked to provide their name, email, and address so that they can receive promotions and other marketing communications from the firm. We track their behavior up to 4 years after that date (ending in November of 2014). We have $13,473$ customers, with a minimum of 3 and a maximum of 51 periods of individual observations, resulting in $287,584$ observations.[22] During this time, we observe a total of $15,985$ repeated transactions (i.e., the average number of transactions per customer is 2.19; or 1.19 repeated transactions). In addition to the behavior of the $13,473$ registered customers, we collect data on all purchases made by "anonymous" customers in all six markets — i.e., those who never shared their identity with the firm. While their behavior is not included in our main analysis (the firm can neither track their future behavior nor communicate with them via email or mail), we use these anonymous transactional data to extract product-level information which will be used to augment the cold start data and to control for shocks in distribution channels that affect the timing of the introduction of new products in specific markets.

[21]The authors thank the Wharton Customer Analytics Initiative (WCAI) for providing this data set.

[22]A period corresponds to exactly 28 days. We do not use a calendar month as our unit of analysis because we want to have the same number of days in all periods.

We specify demand as a logistic regression where $y_{it} = 1$ if customer i transacts at period t, and $y_{it} = 0$ otherwise. Specifically, $f^y(\cdot\,|)$ from (1) is defined as

$$p(y_{it} = 1) = \text{logit}^{-1}\left[\mathbf{x}_{it}^{y\prime} \cdot \boldsymbol{\beta}_i^y + \delta_{rec} \cdot \text{Recency}_{it} + \alpha_m\right], \tag{11}$$

where we control for latent attrition using recency as a covariate (Neslin et al. 2013)[23] and include market-level fixed effects to capture differences in purchase frequencies across countries (i.e., in this case $\widetilde{\mathbf{x}}_{it}^y = [\mathbf{x}_{it}^y, \text{Recency}_{it}]$ and $\boldsymbol{\sigma}^y = \{\delta_{rec}, \alpha_1, \ldots, \alpha_{M-1}\}$, with M the number of markets).

5.1.2 Marketing actions

The firm regularly sends emails and direct marketing to registered customers. The content of these promotional activities is set globally (i.e., the same promotional materials are used across countries, translated to the local language), though their intensity is set by market (e.g., the USA tend to send more emails than France).[24] In addition to promotional activity, the company uses product innovation as a marketing tool. Like other major brands in this category, the focal retailer regularly adds extensions and/or replacements to their product lines. The sense among the company managers is that such an activity not only helps in acquiring new customers but also keeps current customers more engaged with the brand. When the company introduces a new product, it does so in all markets simultaneously. There is, however, some variation across markets regarding when new products were introduced. Conversations with the company confirmed that such variation is due to differences (and random shocks) in the local distribution channels.

While direct and email marketing are observed at the individual level (we denote them by `DM` and `Email`, respectively), the availability of new products is not observed at a granular level. We create a new product introduction variable (`Introd`) by combining point-of-sale data (at the SKU level) with a firm-provided SKU list of new products. Specifically, we obtain the list of all new products introduced during the period of our study. We identify the SKUs for all products in that list and infer availability in each market from *all* purchases observed in that particular market (including all 304,497 transactions from "anonymous" customers). We assume that a new product was introduced in a market at the time the first unit of that SKU was sold. We then create

[23] As discussed in Section 4.1.1, the proposed FIM can accommodate different demand specifications such as "buy-til-you-die" models or HMMs. For our empirical application, we corroborate that adding recency is sufficient to control for latent attrition, which reduces the estimation time when compared with adding a probabilistic latent absorbing state (e.g., Chan et al. 2011).

[24] We only observe email activity sent after September 2012. Therefore, we will only consider customers acquired after that date for the estimation of the model.

a period/market-level variable representing the number of new products that were introduced in each market in each time period.

– Insert Table 1 here –

Table 1 shows the summary statistics for the marketing actions summarized across observations and across individuals. For the latter, we summarize individual average, individual standard deviation, and the individual coefficient of variation. The variation in these data is very rich both across customers and within customers.

We define the vector of demand time-variant covariates $\mathbf{x}_{it}^{y}$ as the intercept, firm-initiated marketing actions, and seasonal factors such as holiday periods,

$$\mathbf{x}_{it}^{y\prime} = \Big[1, \quad \texttt{Email}_{it}, \quad \texttt{DM}_{it}, \quad \texttt{Introd}_{m(i)t}, \quad \texttt{Season}_{m(i)t}\Big]',$$

where `Email`, `DM`, and `Introd` are the marketing actions, and `Season` is a dummy variable that equals 1 for the winter holiday, and 0 otherwise.[25]

Given the business nature of our application, the information provided by the firm about how the managers conduct their marketing actions, the rich longitudinal and cross-sectional variation in our data (Table 1), and our model specification, we argue that the potential endogenous nature of the marketing actions is not a main concern in this research (see Appendix G.1 for details). Nevertheless, in situations where these conditions do not hold (due to different strategic behavior by the firm or for data limitations), the demand model should be adjusted to account for the firm's targeting decisions. Given the flexibility of our modeling framework, those adjustments would merely involve extending the demand model to capture unobserved shocks between firm's actions and individual-level responsiveness (Manchanda et al. 2004) or adding correlations between firm decisions and unobserved demand shocks through copulas (Park and Gupta 2012), depending on how these actions are determined by the firm. Those changes would only affect the demand (sub)model and not the overall specification of the FIM.

5.1.3 (Augmented) acquisition characteristics

Transaction characteristics: We compute `Avg.Price` as the total amount in euros of the ticket divided by the number of units bought at the first transaction; `Quantity` is the total number of

[25] We compute such a variable for each market separately because the exact calendar time for the holiday period varies across countries. For example, in the USA the holiday "shopping" period covers Thanksgiving week until the last week of December (i.e., the end of Christmas), whereas in Spain the only holiday season corresponds to Christmas, which starts at the end of December and ends after the first week of January.

units bought at the first transaction; `Amount` is the total amount in euros of the ticket at the first transaction;[26] `Discount` is a dummy variable that equals 1 if the customer received discounts in the first transaction, and 0 otherwise; `Online` is a dummy variable that equals 1 if the first transaction was made online, and 0 otherwise. We also create a `Holiday` dummy variable that equals 1 if customer made their first transaction during the winter holiday period and 0 otherwise (analogously as the time-varying covariate `Season`).

Product characteristics: Directly from the observed product characteristics, we create a 10-dimensional vector that indicates whether the basket includes a product from a `Category`, including Body care, Face care, Hair care, Toiletries, etc., as defined by the focal company. Moreover, given that product innovation is very important in markets of beauty and cosmetic products, we create a `NewProduct` dummy variable that equals 1 if the customer bought a product that had been introduced in the 30 days prior to the purchase, and 0 otherwise. We also include the average `Size` of the packages in the basket, operationalized as relative size with respect to other products in the same sub-category, and a `Travel` dummy which equals 1 if the basket includes products on travel size, and 0 otherwise.

Latent representation of shopping baskets: As described in Section 3.2, we characterize each customer's first purchase by computing moments of the products included in their shopping basket. The resulting product embeddings in our empirical application is a 6-dimensional vector that represents the position of each product in a similarity space, which we call the "nature" of a product. Once those product embeddings are created, we create `BasketNature`, computed as the "average" product purchased, and `BasketDispersion`, computed as the element-wise standard deviation across products in the same basket, with missing values when the first purchase includes only one product.[27]

Formally, the vector of acquisition characteristics is specified as follows,

$$A_i = [\texttt{Avg.Price}_i, \texttt{Quantity}_i, \texttt{Amount}_i, \texttt{Discount}_i, \texttt{Online}_i, \texttt{Holiday}_i,$$
$$\texttt{Category}_i, \texttt{NewProduct}_i, \texttt{Travel}_i, \texttt{Size}_i, \texttt{BasketNature}_i, \texttt{BasketDispersion}_i].$$

[26] We transform the variables `Avg.Price` and `Amount` using a log function, and the `Quantity` using a log-log function.

[27] In addition, if a first transaction of a customer includes only SKUs of products that were not purchased in any transaction of those anonymous customers' transactions used for generating the product embeddings, then both `BasketNature` and `BasketDispersion` will have missing values as well.

The variation in the acquisition data is very rich (Table 2). For example, 22% of the sample was acquired over the holiday period, and 30% of first transactions included at least one discounted product, 35% included products in the face care category. The standard deviations of price, number of items purchased, amount, relative size, and basket dispersion are large, reflecting the heterogeneous behavior of customers across the six markets. Note that several of these acquisition characteristics are missing for some customers — for example, products for which the package size could not be retrieved from the data have missing `Package Size` observations, baskets that include single items have missing `BasketDispersion` observations, and so forth. These missing observations do not present a challenge in the estimation of the FIM — i.e., there is no need to eliminate observations or to input population averages — because of the way the acquisition characteristics enter the probabilistic model in (2).

– Insert Tables 2 and 3 here –

Consistent with the challenges mentioned in Section 3.4, some acquisition characteristics are correlated with each other (Table 3) — e.g., customers who purchased many items paid less per item (correlation= -0.330), and those who bought on discount also paid slightly lower than those who paid full price when they were first acquired (correlation= -0.200). Online first purchases tend to include more items in the basket (correlation= 0.411) and contain products in the face care category (correlation= 0.483). While it is to be expected that some of these variables will be correlated, as they capture different behaviors incurred by the *same* customer, some of these correlations might also arise from the market conditions at the moment in which a customer was acquired (e.g., if the company introduces all of its new products during the holiday, customers with `Holiday`= 1 will also have `NewProduct`= 1 and vice versa).[28] As discussed in Section 4.1.2, our modeling framework separates these two types of correlations by incorporating firm's market-level actions, $\mathbf{x}^a_{m(i)\tau(i)}$, that potentially affect these acquisition behaviors.

[28] If not accounted for, the latter case could be potentially problematic because the model would not be able to separate the predictive power of being a "holiday customer" from that of being a "new product customer." And, if the company were to change its policy in the future (e.g., introducing new products in June), our model inferences about just-acquired customers could be misleading.

Specifically, we include market-level CRM activities such as number of emails (`MarketEmail`), DMs (`MarketDM`),[29] and the number of products introduced by the firm (`Introd`) in that period.[30] That is,

$$\mathbf{x}^{a}_{m(i)\tau(i)} = \left[\texttt{MarketEmail}_{m(i)\tau(i)}, \quad \texttt{MarketDM}_{m(i)\tau(i)}, \quad \texttt{Introd}_{m(i)\tau(i)}\right]'.$$

Because the span of the acquisition data covers 4 years from 6 different markets, we have substantial variation (longitudinal and cross-sectional) to separate any firm-related systematic relationship among acquisition characteristics from correlations induced by customers' underlying preferences.

5.2 Estimation

We apply our modelling framework to this retail context to show how a firm can make meaningful inferences about newly acquired customers. The firm would do so by calibrating the FIM using historical data from its existing customers and making inferences about newly acquired customers for whom only the acquisition characteristics are observed.

We restrict our analysis to periods in which the firm was engaging in marketing activities, which span from October 2012 to November 2014 ($N = 8,985$ customers). In order to mimic the problem faced by the firm, we estimate the model with the transactional behavior of (existing) customers up to April 2014 and use those estimates to form first impressions for customers acquired after April 2014, using only their acquisition variables.[31] Specifically, we split all customers into three groups: *Training*, *Validation*, and *Test*. We randomly select customers that were acquired before April 2014 to use in our *Training* sample ($N = 5,000$) and use their behavior prior to April 2014 to train the models. Regarding the dimensionality of the FIM, and following the approach discussed in Section 4.1.4, we find that $N_1 = 13$ and $N_2 = 5$ are enough to recover the meaningful associations present in our data. The posterior distribution of $\boldsymbol{\alpha}$ is concentrated close to the origin for a set of lower level traits, indicating that $N_1 = 13$ is high enough to capture the traits that

[29] We calculate market-level number of emails and DMs as the average number of emails and DMs sent in a particular period to customers in that market. Note that the focal customer i cannot receive these marketing communications before being acquired, thus these variables are computed using the set of already existing customers at that time.

[30] Note that the number of products introduced in a particular period enters both the demand and the acquisition model ($\mathbf{x}^{y}_{it}$ and $\mathbf{x}^{a}_{m(i)\tau(i)}$, respectively). This is not problematic because the objective is different on each component. In the demand model, this variable captures the effect of introducing products at a particular period on the purchasing behavior of an existing customer for that particular period. In the acquisition model, this variable serves as a control for extracting the component of the acquisition variables that reflects individuals' traits. For example, the fact that a customer bought a new product on their first transaction could be a signal of customers traits, and/or a consequence of more products being introduced by the firm when the customer was acquired.

[31] We chose this date to reasonably balance the amount of data we need to estimate the model, with the sample size remaining for the prediction analysis.

directly affect the demand and acquisition parameters. Similarly, the posterior distribution of the computed pseudo-α shows that at least one upper level trait is not relevant for impacting the lower level traits, suggesting that $N_2 = 5$ is enough to capture the upper level traits.[32] (For further details see Appendix G.2.)

We also select another set of customers acquired during the same period for our *Validation* sample, which we will use to compare the predictive accuracy of the models at estimating demand ($N = 1,000$). Finally, we use the remaining customers acquired before April 2014, and combine them with those acquired after April 2014 to form our *Test* sample, which we will use to identify valuable customers and to inform our targeting policy ($N = 2,985$).[33]

Similarly as in Section 4.4, we estimate all models (linear HB, Bayesian PCA and FIM) using NUTS in Stan.[34] We also estimate a set of probability models (also estimated with Stan) that have been proposed in the literature to model these type of data as they explicitly account for latent attrition (e.g., Chan et al. 2011; Schweidel and Knox 2013; Schweidel et al. 2014). For completeness, we test multiple specifications varying the inclusion of time-varying covariates in the transaction process and time-invarying covariates in the attrition process, namely (1) Linear model with marketing actions + logistic attrition process (without acquisition covariates), (2) Linear model (without marketing actions) + logistic attrition with acquisition covariates, and (3) Linear model with marketing actions + logistic attrition with acquisition covariates (see details in Appendix G.3). Finally, we estimate two Machine Learning (ML) methods widely used for supervised learning (i.e., whether a customer transact) namely a feed-forward deep neural network (DNN) and a random forest (RF). Both ML models include time-varying covariates, acquisition characteristics, and market-conditions at the moment of acquisition. (See details in Appendix G.4 for details about the packages used for estimation of the ML methods and related model specifications.)

5.3 Results

5.3.1 Parameter estimates

Table 5 shows the population mean and standard deviation of each of the demand parameters. Customers in the sample have a low propensity to transact on average ($\beta^y_{intercept} = -3.110$). Email

[32]For robustness, we estimate another FIM specification with $N_2 = 2$ instead, and we find that all upper traits are relevant, suggesting that $N_2 = 2$ may not be enough to capture the non-linear relationships present in the data.

[33]Ideally, we would like to test our targeting policies using only customers acquired after the calibration period. However, given the low incidence of purchases in this empirical context, we would not observe such a group of customers for a long enough period to have reliable data to validate our predictions.

[34]We do not show the Full hierarchical model given its similar performance to the linear-HB specification.

and direct marketing communications have a positive average impact on purchase ($\beta^y_{email} = 0.111$ and $\beta^y_{dm} = 0.121$, respectively), whereas product introduction effects are not significant on average. Finally, customers return to transact more on holiday periods ($\beta^y_{season} = 0.361$). In Section 5.4 we explore the observed heterogeneity in these components (captured by the FIM) as well as the implications for the managers of the firm.

– Insert Table 5 here –

Another set of interpretable parameters of the FIM are the posterior estimates of the lower layer of the DEF component. Properly rotated, these parameters could be used to interpret the latent factors that connect acquisition characteristics and demand parameters. For the sake of brevity, in this section we focus on the model performance at solving the cold start problem and include those interpretable results in Appendix G.5.

5.3.2 Comparison with the benchmark models

Unlike the simulation exercise, in the empirical application we do not know the true value of the demand parameters (β^y_i), and therefore have to rely on the model predictions to evaluate the quality of the model. We compare the (out-of-sample) accuracy of the FIM predictions with those of the benchmark models in Table 6.[35] (For completeness, the performance of all models on the *Training* sample is presented in Appendix G.6.) The FIM outperforms all the nested and latent attrition benchmarks in out-of-sample fit (i.e., Log-Like) as well as at making predictions at the observation, customer, and period level. This results not only corroborate the results presented in Section 4.4, now on a real-world setting, but also indicate that in this application, the traditional CLV models that explicitly model attrition do not outperform the Linear HB model with recency, even when including the acquisition variables as time-invarying covariates (e.g., Chan et al. 2011). Not surprisingly, the DNN method provide the most accurate results when looking at in observation level RMSEs, with the FIM doing as well as the RF. However, when looking at customer- and period-level RMSE, the FIM outperforms all of the above models.

– Insert Table 6 here –

[35] Arguably one should test these performance metrics on a different set of customers for which we selected the FIM specification. However, most FIM specifications deliver a similar performance on this Validation sample, and thus, would perform similarly well against the benchmark models. More importantly, the main performance test of the FIM is whether it can better identify valuable customers, which we perform using the Test sample in Section 5.4.

These analyses demonstrate that the FIM outperforms the benchmark models at accurately inferring individual-level demand parameters when only acquisition characteristics are available. The benefits of the proposed model are most salient when the underlying relationship between the acquisition characteristics and the parameters governing future demand are not linear, as it is the case for many empirical applications. In the next section we illustrate the managerial value of these predictions and discuss other insights (provided by the model) that are of managerial relevance.

5.4 Overcoming the cold start problem

First, we investigate how accurately the firm can identify "heavy spenders" using only the data from their first transaction. We do so by leveraging the information from customers in the *Test* sample. Specifically, we combine the estimates of the models (calibrated with the *Training* sample) and the acquisition characteristics observed for customers in the *Test* sample, and infer their individual-level demand parameters (see Appendix G.7) to predict each individual's expected number of transactions. We then compare these inferences with their actual behavior using two sets of prediction metrics (Table 7). First, we compute the RMSE on the individual-level average number of transactions per period.[36] Second, based on each individual's expected number of transactions, we flag whether a customer belongs to the top 10% and top 20% of highest average number of transactions and report the proportion customers correctly identified/classified in each group.[37] For reference, we compare those figures with what a random classifier would predict (shown in the last row).

– Insert Table 7 here –

As Table 7 shows, the FIM can predict reasonably well the value of customers: the FIM has a lower RMSE than the Linear HB and the Bayesian PCA models, only outperformed by the RF and the DNN. Moreover, Linear HB and BPCA are significantly better than the baseline at identifying valuable customers, which proves that acquisition characteristics carry valuable information to predict the value of customers. Nevertheless, the FIM significantly improves the identification of valuable customers over the benchmark models, including the DNN, being able to correctly identify 40.5% of customers in the Top 10% and 47.7% of customers in the Top 20%. These results are consistent with the notion that, because the FIM captures the non-linearities in the relationship between acquisition characteristics and future demand parameters, it does an excellent

[36] Using our notation, the individual level average number of transactions per period is $\bar{Y}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} y_{it}$.

[37] We make predictions and compute recovery rates for each draw of the posterior distribution and report posterior means and 95% CPI.

job — significantly better than the benchmarks — at sorting customers based on their expected value inferred from their acquisition characteristics.

Similarly, a firm would use the FIM to identify which customers are the most sensitive (or least sensitive) to marketing interventions; information that will be instrumental in increasing the effectiveness of its marketing actions (e.g., Ascarza 2018). Unfortunately, our data does not enable us to quantify the exact value that the focal firm could extract from a FIM-based targeting approach — ideally, one would run a field experiment to test the effectiveness of targeting policies based on the predictions of the FIM. Nevertheless, combining the results from Section 4.4, where we demonstrate the model's ability to predict the (individual-level) demand intercept as well as the sensitivity to the covariates, with the results in Table 7, where we corroborate some of those findings in our empirical application, we are confident that implementing targeting policies based on predictions of the FIM would generate incremental revenues to the firm. We trust that future research will be able to quantify these benefits empirically.

Second, we use the FIM results to explore the acquisition variables that better characterize "heavy spenders" (separately from light users), customers with "high sensitivity to email" (from those who are better left out in the email campaigns), and those who are "most sensitive to direct marketing" campaigns. Based on the model predictions, we split customers from the *Test* sample in three groups: Top 10%, Middle 80% and Bottom 10% for each of the three categories and summarize the average value of each of the (standardized) variables observed at the moment of acquisition. Figure 5 shows the results when sorting customers on the basis of expected future value. Several interesting findings emerge: Consistent with the patterns observed when exploring the predictive power of the acquisition variables (Figure 1) we find that the Top 10% heavy spenders are less likely to be acquired during the holiday period, more likely to being acquired offline, and tend to buy expensive and discounted products in their first purchase, compared to those at the Bottom 10%. They are also characterized to buy certain types of products, as indicated by the high chance to include Perfume and Hair products in their first transaction (less likely to contain products in the Body Care, Home and Services categories), as well as by a high score in dimension 4 of the product embeddings.[38]

– Insert Figure 5 here –

[38]This dimension is related to products such as "Grape Line Showers" and "Olive Harvest Conditioner," see Table A.1 in Appendix A.

We repeat the analysis now sorting customers based on their predicted sensitivity to email (Figure 6) and predicted sensitivity to DM (Figure 7). Consistent with the previous findings, several acquisition characteristics exhibit a non-linear relationship with the sensitivities to marketing actions. Both the Top 10% and Bottom 10% email sensitivity groups are less likely to buy in the Body Care category during their first transaction, compared with the remaining 80% of customers in between. Customers who are the most sensitive to email marketing are more likely to be acquired online, buy less expensive products, and fewer units at their first purchase. With respect to DM, low sensitive customers buy fewer units and more expensive products in their first transaction, while high sensitive customers are more likely to buy relatively small sized products, recently introduced products, and products in the Perfume Category at their first purchase.

– Insert Figures 6 and 7 here –

Finally, we use the inferred demand parameters from these test customers to explore the relationships between the magnitude of the demand parameters and the acquisition characteristics. Figure 9 shows the individual level posterior mean of the demand parameter vs. the acquisition characteristics for a set of demand parameters and acquisition characteristics. In particular, we find that these plots corroborate that there are non-linear relationships that the model allows to uncover.[39] Figure 8 explore possible interactions by presenting box plots of individual level posterior mean demand parameters and pairs of discrete acquisition characteristics. The model replicates the model-free insights shown in Figure 2: (1) the relationship between the intercept and whether the customer was acquired during the winter holiday season (`Holiday`) depends on whether the customer purchased a travel-sized product (`Travel Size`), and (2) the relationship between the intercept and whether the customer purchased discounted products at acquisition (`Discount`) depends on whether the customer purchased a recently introduced product (`New Product`). Moreover, the model not only captures these relationships for the intercept but also for other demand parameters. For instance, the holiday season lift is higher for customers that were acquired during a past holiday season compared to those that were not, but this difference is considerably larger for those that did not purchased a travel-sized product when acquired. Also, the differences in email sensitivities across customers that received discounts on their first purchase only exist for those who purchased a recently introduced product at acquisition.

[39]Note, that these plots show marginal relationships of demand parameters and acquisition characteristics (i.e., one at a time) where indeed the model cover relationships accounting for all acquisition characteristics.

– Insert Figures 8 and 9 here –

6 CONCLUSION

We have developed a modeling framework (FIM) that, leveraging information collected when customers are acquired, enables firms to overcome the cold start problem of CRM. Using a probabilistic machine learning approach, the model connects underlying acquisition and demand parameters using a set of hidden factors modeled via deep exponential families. The multi-layer structure with flexible relationships among layers enables the researcher or analyst to be agnostic about the (assumed) underlying relationship among variables. The hidden factors automatically extract relevant information from existing data — i.e., identify the traits that relate acquisition characteristics with future outcomes — overcoming the challenge (commonly faced by firms) of maintaining significant amounts of redundant and irrelevant data in their customer databases.

We have illustrated the benefits of using the FIM in a retail setting. First we have shown how the focal firm can further leverage its existing database to augment the cold start data using readily-available techniques. We have further demonstrated how subtle signals extracted from the augmented data by the FIM enables the focal firm to make individual-level inferences about just-acquired customers, for example, distinguish high-value customers from those unlikely to purchase again and those most and least sensitive to marketing interventions, such as email campaigns or direct marketing. We leverage the model predictions to identify characteristics of first transactions that are predictive of customer behavior in future periods. For example, compared to the rest, Top 10% heavy spenders are more likely to be acquired online and their first purchases to be expensive and discounted products, and customers identified as most sensitive to email marketing to also be more likely to be acquired online but buy less expensive products, and their first purchases to be of fewer units.

These findings suggest that firms can meaningfully categorize customers based on characteristics of their first transactions. We believe this approach to customer segmentation to be promising in relying neither on sometimes difficult to obtain customer-provided data (Dubé and Misra 2017) and nor on external sources of data that could pose privacy concerns. The resulting insights can be used both to prune acquisition data and inform decisions about the types of variables worth collecting from customers that make a first transaction or first visit a company's website. Our research shows that firms leave value on the table by not fully leveraging the multiple behaviors ob-

served when a customer makes a first transaction, and provides a general framework for extracting meaningful but hard-to-pinpoint relationships imprinted in subtle ways in "cold start" data.

While this research highlights the value of using the FIM to tackle the cold start problem of CRM, it is also important to acknowledge some limitations of the present research. The simulation analyses enabled us to validate the accuracy of the model at inferring individual-level parameters, but doing so in an empirical setting, in which only realized purchases are observed, is more difficult. We leave it to future research to examine and quantify the effectiveness of targeting policies based on the predictions of the FIM. Regarding the model specification, we investigated model performance using linear and logistic specifications for the demand and acquisition models. Although the proposed FIM is extremely flexible so as to be adaptable to other modeling frameworks, we have not empirically tested the model's performance in more complex structures. The current model estimation is computationally feasible for datasets with thousands of customers, dozens of time periods, and a handful of variables (as in our empirical application). Although the model scales readily to situations with more acquisition variables (and the model does not need to be fully trained when making inferences on new customers), increasing the sample size to, for example, millions of customers will increase estimation time substantially, constraining the ability to gauge customers' first impressions in a timely manner. For such cases, variational inference implemented in recent deep probabilistic programming languages that allow for black-box variational inference methods (e.g., Pyro) might be a better way to estimate and use the model. We look forward to reading and exploring such approaches in future research.

A natural extension to this research would be to investigate a wider range of acquisition characteristics and the relevance thereof to customers' first impressions in different contexts. The results of our empirical application could be built on to further augment the data from first purchases and incorporate other acquisition characteristics that, although not currently collected (e.g., whether the customer visited the store alone or with family), could be valuable in identifying which marketing actions are most likely to increase future sales. We encourage further research to investigate these research settings and identify additional drivers and methods that might help companies overcome the cold start problem.

The main goal of this work being to provide a flexible model that overcomes the cold start problem, we have not formally investigated the latent traits that drive all the observed behaviors. It would be relevant for researchers and marketers to identify individual traits that characterize

shopper behavior, to which end customer behavior in a variety of contexts could be measured and estimated in a unifying FIM framework. We hope that this research opens up new avenues for understanding "universal" shopping traits and identifies the behaviors that best relate to those generalizable findings.

References

Allenby, G. M. and Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89(1-2):57–78.

Anderson, E., Chaoqun, C., Israeli, A., and Simester, D. (2020). Do harbinger products signal which new customers will stop purchasing?

Ansari, A. and Mela, C. F. (2003). E-customization. *Journal of Marketing Research*, 40(2):131–145.

Artun, O. (2014). What are those new holiday customers worth? [Online; accessed 5-February-2017] https://www.internetretailer.com/2014/12/19/what-are-those-new-holiday-customers-worth.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98.

Bishop, C. M. (1999). Bayesian PCA. In *Advances in neural information processing systems*, pages 382–388.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Braun, M., Schweidel, D. A., and Stein, E. (2015). Transaction attributes and customer valuation. *Journal of Marketing Research*, 52(6):848–864.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–29.

Chan, T. Y., Wu, C., and Xie, Y. (2011). Measuring the lifetime value of customers acquired from Google Search Advertising. *Marketing Science*, 30(5):837–850.

Chen, F., Liu, X., Proserpio, D., and Troncoso, I. (2020). Product2vec: Understanding product-level competition using representation learning. *Available at SSRN*.

Datta, H., Foubert, B., and Van Heerde, H. J. (2015). The challenge of retaining customers acquired with free trials. *Journal of Marketing Research*, 52(2):217–234.

Dew, R. and Ansari, A. (2018). Bayesian nonparametric customer base analysis with model-based visualizations. *Marketing Science*, 37(2):216–235.

Dew, R., Ansari, A., and Li, Y. (2020). Modeling dynamic heterogeneity using gaussian processes. *Journal of Marketing Research*, 57(1):55–77.

Dubé, J.-P. and Misra, S. (2017). Scalable price targeting. Technical report, National Bureau of Economic Research.

Fader, P. S., Hardie, B. G., and Jerath, K. (2007). Estimating clv using aggregated data: the tuscan lifestyles case revisited. *Journal of Interactive Marketing*, 21(3):55–71.

Fader, P. S., Hardie, B. G. S., and Lee, K. L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, 24(2):275–284.

Fader, P. S., Hardie, B. G. S., and Shang, J. (2010). Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*, 29(6):1086–1108.

Forbes (2015). Big data: A game changer in the retail sector. [Online; accessed 23-September-2017] https://www.forbes.com/sites/bernardmarr/2015/11/10/big-data-a-game-changer-in-the-retail-sector/.

Gopalakrishnan, A., Bradlow, E. T., and Fader, P. S. (2016). A cross-cohort changepoint model for customer-base analysis. *Marketing Science*, (December).

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

Jacobs, B. J., Donkers, B., and Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, 35(3):389–404.

Karaletsos, T. and Rätsch, G. (2015). Automatic relevance determination for deep generative models. *arXiv preprint arXiv:1505.07765*.

Knox, G. and van Oest, R. (2014). Customer complaints and recovery effectiveness: A customer base approach. *Journal of Marketing*, 78(5):42–57.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Kumar, M., Eckles, D., and Aral, S. (2020). Scalable bundling via dense product embeddings. *arXiv preprint arXiv:2002.00100*.

Lewis, M. (2006). Customer acquisition promotions and customer asset value. *Journal of Marketing Research*, 43(2):195–203.

Loupos, P., Nathan, A., and Cerf, M. (2019). Starting cold: The power of social networks in predicting non-contractual customer behavior. *Available at SSRN 3001978*.

MacKay, D. J. (1995). Probable networks and plausible predictions—a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469–505.

Manchanda, P., Rossi, P. E., and Chintagunta, P. K. (2004). Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research*, 41(4):467–478.

Mcauliffe, J. D. and Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mnih, A. and Hinton, G. E. (2009). A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088.

Mohamed, S., Ghahramani, Z., and Heller, K. A. (2008). Bayesian exponential family pca. *Advances in neural information processing systems*, 21:1089–1096.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Neslin, S. A., Taylor, G. A., Grantham, K. D., and McNeil, K. R. (2013). Overcoming the "recency trap" in customer relationship management. *Journal of the Academy of Marketing Science*, 41(3):320–337.

Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3):521–543.

Park, S. and Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4):567–586.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. (2016). Deep survival analysis. *arXiv preprint arXiv:1608.02158*.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015). Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Research and Markets (2016). Overview & evolution of the global retail industry. [Online; accessed 23-September-2017] https://www.researchandmarkets.com/research/tqh2xb/.

RJMetrics (2016). The ecommerce holiday customer benchmark. [Online; accessed 5-February-2017] https://rjmetrics.com/resources/reports/the-ecommerce-holiday-customer-benchmark/.

Rossi, P. E., McCulloch, R. E., and Allenby, G. M. (1996). The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340.

Ruiz, F. J., Athey, S., and Blei, D. M. (2017). Shopper: A probabilistic model of consumer choice with substitutes and complements. *arXiv preprint arXiv:1711.03560*.

Schmitt, P., Skiera, B., and Van den Bulte, C. (2011). Referral programs and customer value. *Journal of Marketing*, 75(1):46–59.

Schmittlein, D. C., Morrison, D. G., and Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science*, 33(1):1–24.

Schweidel, D. A. and Knox, G. (2013). Incorporating direct marketing activity into latent attrition models. *Marketing Science*, 32(3):471–487.

Schweidel, D. A., Park, Y.-h., and Jamal, Z. (2014). A multiactivity latent attrition model for customer base analysis. *Marketing Science*, 33(2):273–286.

Shaffer, G. and Zhang, Z. J. (1995). Competitive coupon targeting. *Marketing Science*, 14(4):395–416.

Steffes, E. M., Murthi, B. P. S., and Rao, R. C. (2011). Why are some modes of acquisition more profitable? A study of the credit card industry. *Journal of Financial Services Marketing*, 16(2):90–100.

Uncles, M. D., East, R., and Lomax, W. (2013). Good customers: The value of customers by mode of acquisition. *Australasian Marketing Journal*, 21(2):119–125.

Verhoef, P. C. and Donkers, B. (2005). The effect of acquisition channels on customer loyalty and cross-buying. *Journal of Interactive Marketing*, 19(2):31–43.

Villanueva, J., Yoo, S., and Hanssens, D. M. (2008). The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth. *Journal of Marketing*, 45(1):48–59.

Voigt, S. and Hinz, O. (2016). Making digital freemium business models a success: Predicting customers' lifetime value via initial purchase information. *Business & Information Systems Engineering*, 58(2):107–118.

Wang, Y. and Blei, D. M. (2019). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

7 TABLES AND FIGURES

Table 1: Summary of time-varying marketing actions.

Marketing action	Statistic	Mean	SD	N
Email	Across observations	3.267	4.686	287,584
	Indiv. average	4.272	3.612	13,473
	Indiv. st. dev.	3.404	1.790	13,473
	Indiv. coeff. of variation	1.425	1.082	13,336
Direct Marketing	Across observations	1.006	1.889	287,584
	Indiv. average	1.329	1.018	13,473
	Indiv. st. dev.	1.731	0.769	13,473
	Indiv. coeff. of variation	2.031	1.205	13,455
Products introduced	Across observations	0.923	1.264	287,584
	Indiv. average	0.657	0.532	13,473
	Indiv. st. dev.	0.755	0.534	13,473
	Indiv. coeff. of variation	1.354	0.478	11,927

Table 2: Summary statistics of selected acquisition characteristics.

Variable	Description	Mean	SD	N
Avg. price (€)	Average price per unit, in euros	11.642	10.237	13,473
Quantity	Total number of units purchased	4.934	5.298	13,473
Amount (€)	Total ticket amount, in euros	39.567	38.433	13,473
Holiday	Whether customer was acquired during the Holiday	0.220	−−	13,473
Discount	Whether discounts were applied in transaction	0.302	−−	13,473
Online	Whether the transaction was online	0.176	−−	13,473
New product	Whether a new product was purchased	0.431	−−	13,473
Travel	Whether a travel-size product was purchased	0.397	−−	13,473
Package Size	Average size of products (relative to its subcategory)	1.080	0.701	13,352
Avg. BasketDispersion	Average basket dispersion across all dimensions	1.338	0.660	9,928
Face Care	Whether a product in the Face Care category was purchased	0.352	−−	13,473
Hair Care	Whether a product in the Hair Care category was purchased	0.120	−−	13,473

Note: For the sake of simplicity, we omit the descriptive statistics for the 6 BasketNature variables and 8 remaining product categories. We also aggregate the BasketDispersion variables, by averaging across all dimensions of the *word2vec* representations. Missing values correspond to first purchases including products with missing information and for the case of BasketDispersion, those with only one item in the basket.

Table 3: Correlations among selected acquisition characteristics.

	Avg. price	Quantity	Amount	Size	Holiday	Discount	Online	New product	Travel	Face care
Avg. price	1.000									
Quantity	-0.330									
Amount	0.251	0.594								
Size	0.396	-0.238	0.038							
Holiday	-0.082	0.179	0.090	-0.027						
Discount	-0.200	0.285	0.184	-0.160	0.055					
Online	-0.241	0.411	0.168	-0.097	0.056	-0.049				
New product	-0.036	0.250	0.248	-0.055	0.068	0.066	0.106			
Travel	-0.350	0.347	0.122	-0.348	0.088	0.289	0.009	0.149		
Face care	-0.066	0.366	0.298	-0.113	0.051	0.096	0.483	0.177	0.083	
Hair care	-0.124	0.261	0.121	-0.091	-0.016	0.084	0.266	0.139	0.063	0.155

Note: We dropped missing values in pairwise computations only.

Table 4: Accuracy of predictions of demand parameters for (out-of-sample) customers

	Scenario 1 Linear		Scenario 2 Quadratic/interactions		Scenario 3 Positive part	
	R-squared	RMSE	R-squared	RMSE	R-squared	RMSE
Intercept						
HB demand-only	0.001	6.703	0.020	7.624	0.007	8.514
Linear HB	0.988	**0.734**	0.711	4.113	0.783	4.056
Full hierarchical	**0.988**	0.735	0.704	4.164	0.781	4.091
Bayesian PCA	0.988	0.736	0.706	4.484	0.780	4.329
FIM	0.988	0.738	**0.888**	**2.661**	**0.928**	**2.987**
Effect of covariates						
HB demand-only	0.005	2.562	0.004	4.589	0.001	4.604
Linear HB	0.986	0.303	0.258	3.969	0.736	2.363
Full hierarchical	0.986	0.303	0.258	3.970	0.733	2.378
Bayesian PCA	**0.986**	**0.301**	0.245	4.364	0.738	2.752
FIM	0.986	0.302	**0.515**	**3.229**	**0.745**	**2.325**

Table 5: Parameter estimates of FIM.

Demand parameter		**Posterior statistics** Post. mean	Post. sd	PCI 2.5%	PCI 97.5%
Intercept	Pop. mean	−3.110	0.051	−3.205	-3.024
	Pop. std. dev.	0.364	0.086	0.245	0.549
Email	Pop. mean	0.111	0.026	0.061	0.163
	Pop. std. dev.	0.167	0.031	0.110	0.235
DM	Pop. mean	0.121	0.028	0.067	0.174
	Pop. std. dev.	0.137	0.023	0.094	0.182
Product introductions	Pop. mean	−0.058	0.048	−0.164	0.024
	Pop. std. dev.	0.213	0.046	0.128	0.310
Season	Pop. mean	0.361	0.072	0.235	0.502
	Pop. std. dev.	0.362	0.065	0.245	0.505

Table 6: Comparison with benchmark models (*Validation* sample).

Model	Log-Like	RMSE Observation	Customer	Period
Linear HB	−2134.6	0.247	1.307	4.570
Latent Attrition w/ Acq.	−2367.4	0.249	1.403	4.951
Latent Attrition w/ Mktg. Actions	−2194.1	0.250	1.361	4.499
Latent Attrition w/ Acq.+Mktg. Actions	−2384.5	0.253	1.421	4.722
Bayesian PCA	−2010.0	0.240	1.184	4.240
Feed-Forward DNN	−−	**0.235**	1.095	7.468
Random Forest	−−	0.236	1.118	6.783
FIM	**−1927.0**	0.236	**1.046**	**4.058**

Note: Log-Like corresponds to the log expected posterior predictive density.

Table 7: Identifying valuable customers using *Test* customers.

		% customers correctly classified	
Model	**RMSE**	**Top 10%**	**Top 20%**
Linear HB	0.157	0.151	0.253
Latent Attrition w/ Acq.	0.520	0.113	0.207
Latent Attrition w/ Mktg. Actions	0.303	0.213	0.248
Latent Attrition w/ Acq.+Mktg. Actions	0.242	0.090	0.191
Bayesian PCA	0.138	0.208	0.313
Feed-Forward DNN	**0.098**	0.349	0.450
Random Forest	0.106	0.193	0.310
FIM	0.131	**0.401**	**0.477**
Baseline (random)	–	0.100	0.200
	–	(0.067,0.127)	(0.170,0.230)

Note: The proportion of top spenders is computed by predicting over the observed periods, computing the average number of transactions per period, and selecting customers with highest predicted values.

Figure 1: Observed (mean) repeated transactions as a function of a sample of augmented acquisition characteristics. All acquisition variables are constructed from the first transaction of each customer. Repeated transactions do not include the first transaction. Error bars represent standard errors.

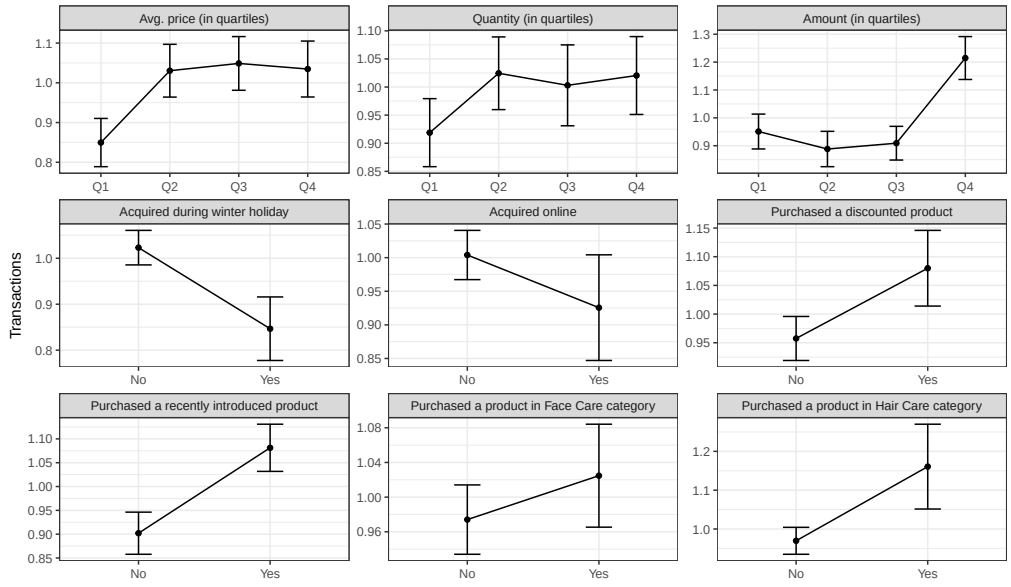


Figure 2: Observed (mean) repeated transactions as a function of interactions among acquisition characteristics. All acquisition variables are constructed from the first transaction of each customer. Repeated transactions do not include the first transaction. Error bars represent standard errors.

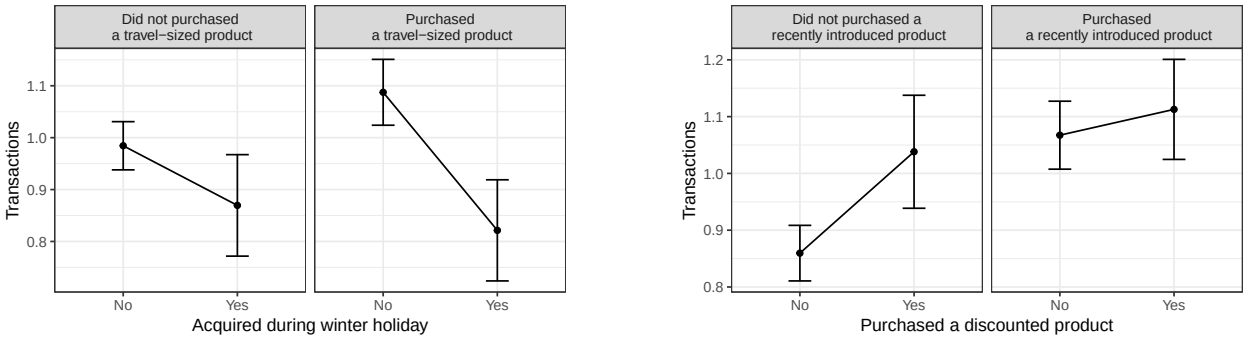


Figure 3: Graphical model of first impressions

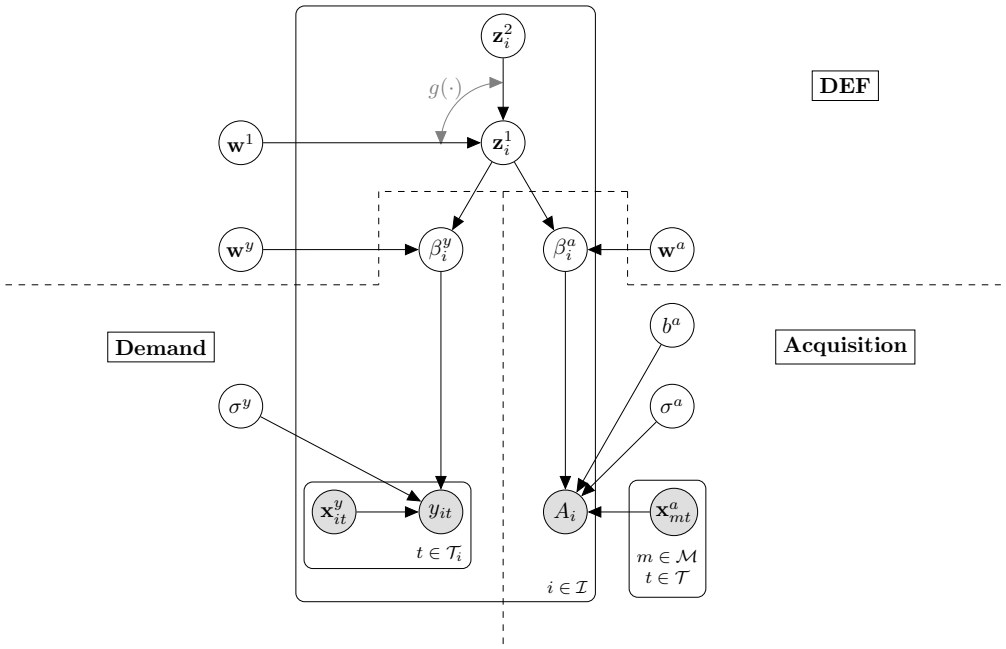


Figure 4: Visualization of model performance for Scenario 3: positive-part individual results of intercept. The first row shows the scatter plot of the individual true vs. posterior mean for each model. The second row shows the individual posterior mean (red) and true (blue) as a function of acquisition variable 1 (A_1).

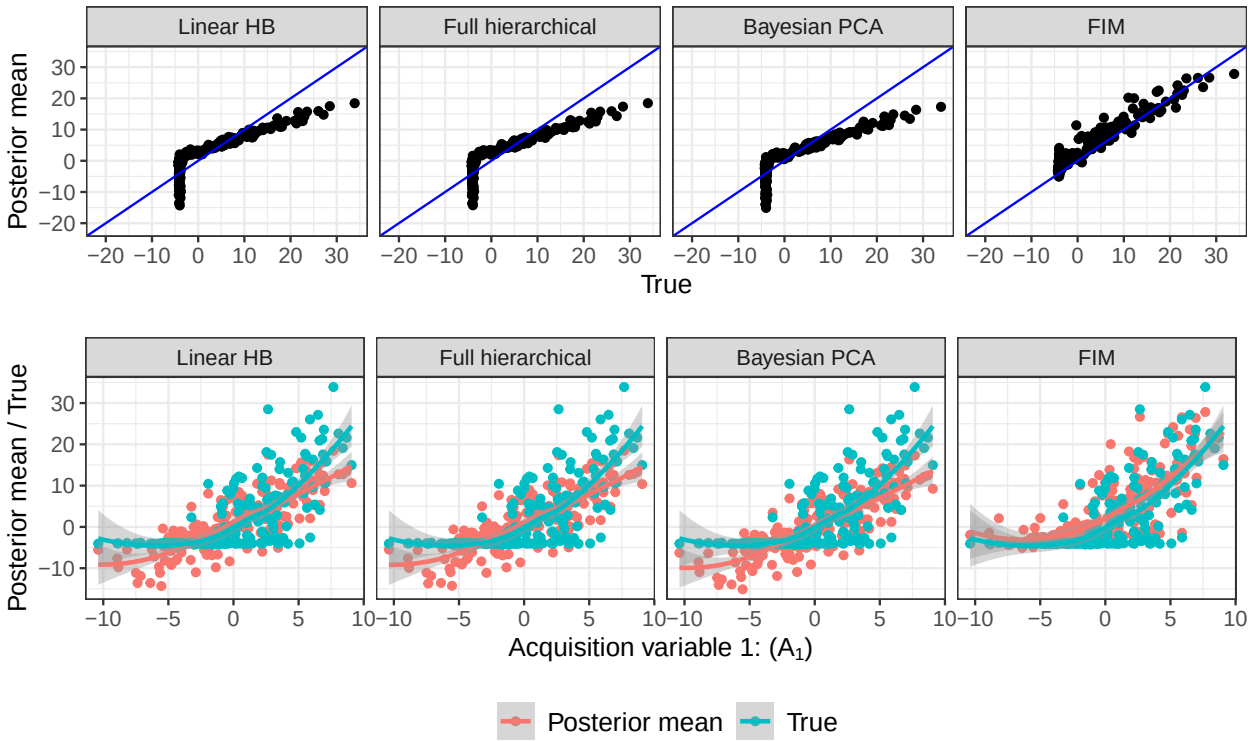


Figure 5: Acquisition characteristics for customers with top/middle/low CLV.

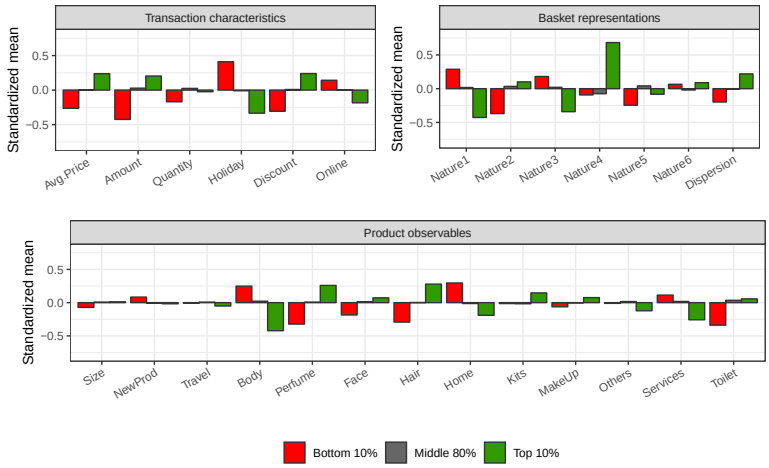


Figure 6: Acquisition characteristics for customers with top/middle/low sensitivity to Email.

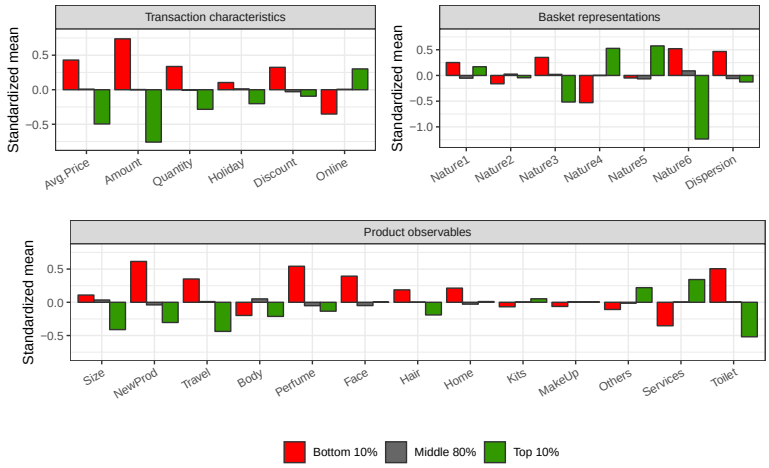


Figure 7: Acquisition characteristics for customers with top/middle/low sensitivity to DM.

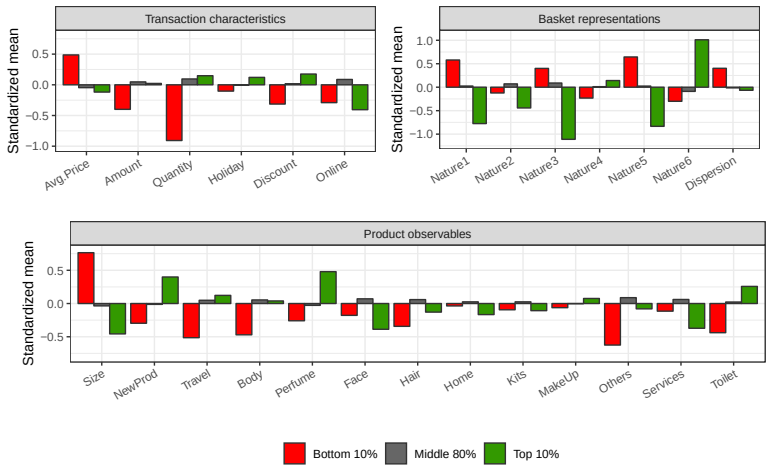


Figure 8: Empirical relationship between the posterior mean and some of the (binary) acquisition characteristics

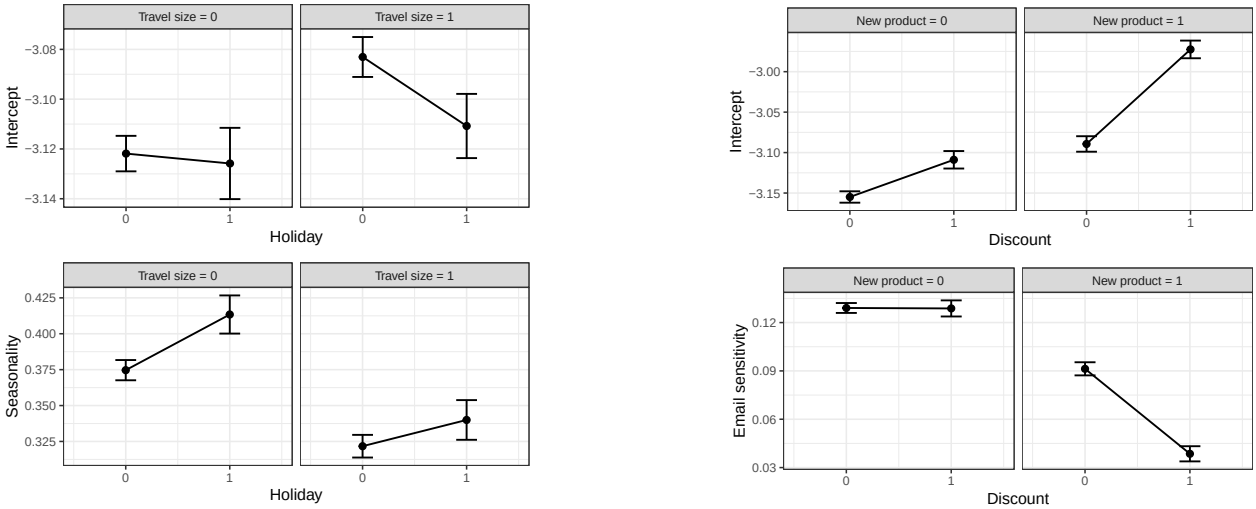
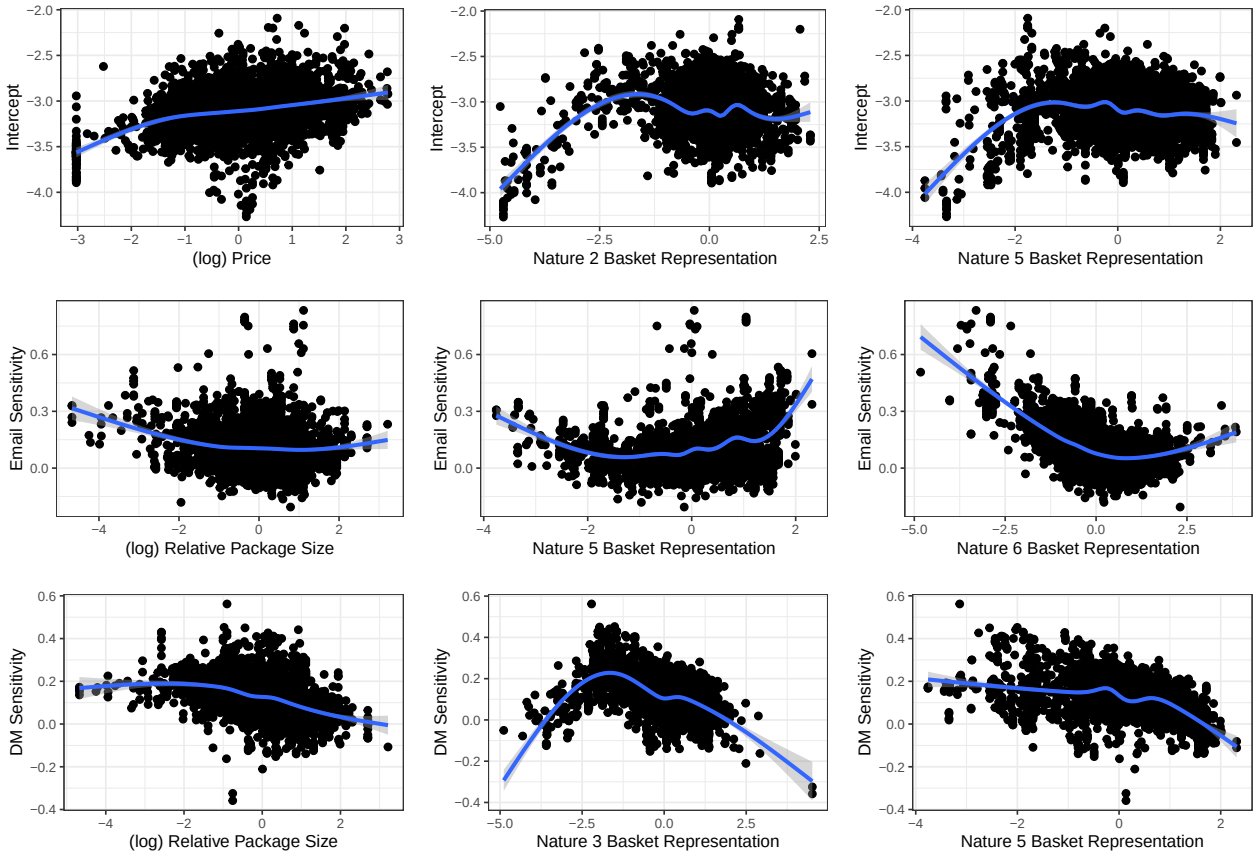


Figure 9: Empirical relationship between the posterior mean and some (continuous) acquisition characteristics



WEB APPENDICES

Overcoming the Cold Start Problem of CRM
using a Probabilistic Machine Learning Approach

Nicolas Padilla
Eva Ascarza

Nicolas Padilla is an Assistant Professor of Marketing, London Business School (email: npadilla@london.edu). Eva Ascarza is the Jakurski Family Associate Professor of Business Administration, Harvard Business School (email: eascarza@hbs.edu).

Contents

A Augmenting the acquisition characteristics via product embeddings

While one could attempt to directly include the product-level purchase incidence as acquisition characteristics, such an approach would suffer from high levels of sparsity (i.e., unique SKUs are purchased rather infrequently over the first transaction of the customers in the calibration data). Instead, we rely on embedding models that have been developed to overcome the challenge that large "vocabularies" have on computing probabilities of multinomial outcomes. (Specifically, how to efficiently compute/approximate the large denominator of the softmax). As described in Section 3.2 (Augmenting cold start data with acquisition characteristics), we use the transactional data from anonymous customers to create product embedding vectors, i.e., vectors representations of all products available, that captures the nature of products, as perceived by the customers. In essence, we leverage the co-occurrences of products in customers' baskets to infer similarities across products.

A.1 Data processing

The anonymous transactions include 304,497 transactions and 4,730 unique product codes (corresponding to unique SKUs specified by the firm). Many of those product codes are very similar in nature, as they only reflect slight modifications of the exact same product, different sizes, or travel-size packaging. Because those pieces of information are already captured by the acquisition characteristics (`NewProduct`, `Travel`, and `Size`), we aggregate the product code to unique combinations of product sub-category (e.g., liquid soap, bath, beauty oils) and product line (e.g., shea butter, chamomile, fresh-summer). This characterization of product codes results in 515 unique products in the data.

A.2 Word2vec algorithm

To capture latent semantic patterns among products in the same transaction, we use Word2vec, a word embedding method in Natural Language Processing (NLP), to map words into numerical vectors. Word2vec is proposed by Mikolov et al. (2013) who develop two architectures to take advantages of word context: continuous bag-of-words (CBOW) and continuous skip-gram (SG). The

first model predicts a word based on its neighbor words, and the second model predicts surrounding words based on a given word. We use the SG model to generate a "product vector."

More specifically, let $T = \{T_1, T_2, ..., T_H\}$ be the set of transactions, $Q = \{q_1, q_2, ..., q_M\}$ be the set of unique products, $V = \{V_{q_1}, V_{q_2}, ..., V_{q_M} | V_{q_i} \in \mathbb{R}^N\}$ be the set of product vectors. Then, the SG model optimizes V by maximizing the loss function:

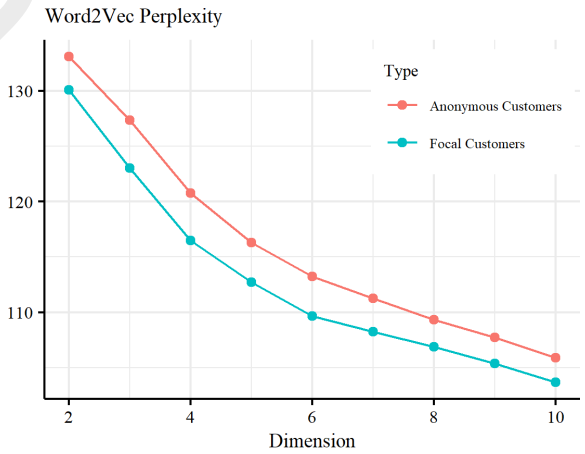
$$L = \sum_{T_i \in T} \sum_{q_i \in Q} \sum_{1<=j<=M, j \neq i} \log P(q_j | q_i), \tag{A.12}$$

where P is the probability of observing product q_j given the occurrence of product q_i in the same transaction. The probability function is defined by the softmax:

$$P(q_j | q_i) = \frac{e^{V_{q_i}^T V_{q_j}}}{\sum_{k=1}^{M} e^{V_{q_i}^T V_{q_k}}}. \tag{A.13}$$

A straightforward softmax calculation requires an evaluation of all M products in the denominator, so we speed up the computation by using hierarchical softmax (Mnih and Hinton 2009) to approximate the conditional probability. We implement the model via the Python package Gensim (Řehůřek and Sojka 2010) and train the model on anonymous customers till the loss L is stable. The hyper parameters in Gensim are: sg=1, negative=0, hs=1, window=10000, min_count=1, random_seed=4. We set a large sliding window size so that all product combinations are selected.

Figure A.1: Model selection for Word2vec: Perplexity when varying the number of dimensions from 2 to 10.



We calibrate the Word2vec algorithm using $N = 2, 3, ..., 10$ dimensions to represent the set of 515 products available in the data and compare the model performance over the number of dimensions (Figure A.1). We select the model with 6 dimensions based on the (lower) rate of decline.[1] As a result, we have a matrix of product embeddings that maps each product to a 6-dimensional vector that represents the position of the product within a multi-dimensional space that captures product similarities.

A.3 Interpreting the product dimensions

One could interpret those dimensions by identifying the products that score high in each of the dimensions (Table A.1). While not all dimensions are easy to interpret, some clearly capture characteristics defining the nature of the product. For example, looking at the products that score high in the first dimension, we infer that it represents aromas and items for the household. The fifth dimension seems to capture kits and other uncategorized items whereas the sixth dimension represents a specific line of beauty called Fleur Cherie.

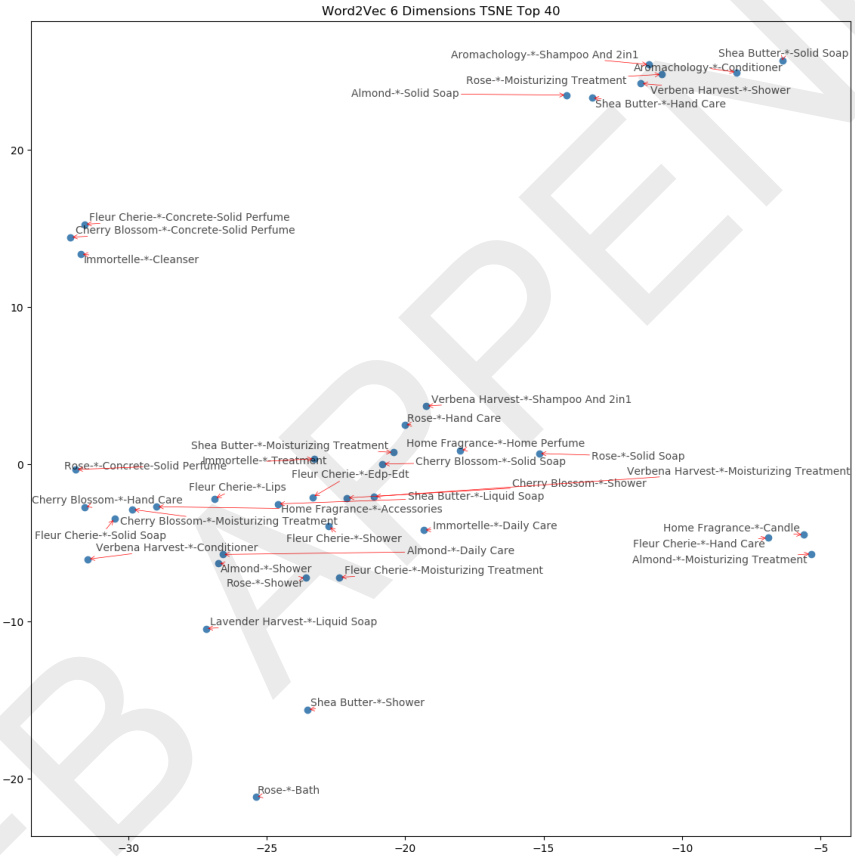
Table A.1: Top 5 products per dimension of the product embeddings.

Dimension 1	**Dimension 2**
Furniture-*-Others	Immortelle-*-Accessories
Aromachology-*-Accessories	Collection De Grasse-*-Accessories
Aromachology-*-Beauty Oils	027-*-Others
Home Fragrance-*-Accessories	Collection De Grasse-*-Shampoo And 2in1
Relaxing Recipe-*-Home Perfume	Verbena Harvest-*-Conditioner
Dimension 3	**Dimension 4**
Furniture-*-Others	Grape-*-Shower
Orange Harvest-*-Lips	Fleur Cherie-*-Concrete-Solid Perfume
Bonne Mere-*-Others	Olive Harvest-*-Conditioner
Homme-*-Edp-Edt	Shea Butter-*-Body Sun Care
Relaxing Recipe-*-Kits	Grape-*-Body Scrub
Dimension 5	**Dimension 6**
027-*-Others	Fleur Cherie-*-Solid Soap
Almond-*-Kits	Fleur Cherie-*-Shower
Bonne Mere-*-Kits	Bonne Mere-*-Others
Others-*-Lips	Fleur Cherie-*-Edp-Edt
Immortelle-*-Moisturizing Treatment	Fleur Cherie-*-Moisturizing Treatment

[1] A company with a larger product space would calibrate the model with a greater number of dimensions and pick the dimensionality that is best suited for their application.

In addition to creating the product embeddings that will be used to augment the data, this methodology can also be used to visualize similarities across products. For example, Figure A.2 visualizes the 40 most popular products in the anonymous data. Because showing the 6 dimensions would be cumbersome, we apply TSNE (t-distributed stochastic neighbor embedding; algorithm for dimensionality reduction that is well-suited to visualizing high-dimensional data) and visualize the data in a two-dimensional space. It appears to be four clusters representing similarities across these products.

Figure A.2: Visual representation of the product embeddings



A.4 Product mapping for first purchase data

Finally, once the product embeddings are created, we characterize the first purchase from our focal customers by taking the average of the embeddings of each product in the basket (`BasketNature`) and by computing the standard deviation of all products in the basket (`BasketDispersion`), which has missing value if the first purchase only included one product. Note that four products from the first purchase data were not present in the data from anonymous customers and therefore have missing values in the `ProductNature` variable as well.

B Brief description of DEFs

DEFs are deep generative probabilistic models that describe a set of observations $\mathbf{D}_i$ with latent variables layered following a structure similar to deep neural networks. The lowest layer describes the distribution of the observations, $p(\mathbf{D}_i|\mathbf{z}_i^1, \mathbf{W}^0) = f\left(\mathbf{D}_i|\mathbf{W}^{0\prime}\mathbf{z}_i^1\right)$ and the top layers describe the distribution of the layer just below them. As in deep neural networks, DEFs have two sets of variables: layer variables ($\mathbf{z}_i^\ell$) and weights matrices ($\mathbf{W}^\ell$) for the ℓ'th layer. Each layer variable $\mathbf{z}_i^\ell$ is distributed according to a distribution in the exponential family with parameters equal to the inner product of the previous layer parameters $\mathbf{z}_i^{\ell+1}$ and the weights $\mathbf{W}^\ell$, by

$$p(z_{i,k}^\ell|\mathbf{z}_i^{\ell+1}, \mathbf{w}^\ell) = EXPFAM_\ell\ \left(z_{i,k}^\ell|\ g_\ell\left(\mathbf{w}_k^{\ell\prime}\cdot\mathbf{z}_i^{\ell+1}\right)\right) \qquad \ell \in \{1,\ldots,L-1\},$$

where $z_{i,k}^\ell$ is the k'th component of vector $\mathbf{z}_i^\ell$, $\mathbf{w}_k^\ell$ is the k'th column of weight matrix $\mathbf{W}^\ell$, $EXPFAM_\ell(\cdot)$ is a distribution that belongs to the exponential family and governs the ℓ'th layer, and $g_\ell(\cdot)$ is a link function that maps the inner product to the natural parameter of the distribution, allowing for non-linear relationships between layers. The top layer is purely governed by a hyperparameter η, that is, $p(z_{i,k}^L) = EXPFAM_L\ \left(z_{i,k}^L|\eta\right)$.

Similar to deep unsupervised generative models, DEF models are suitable to find interesting exploratory structure in large data sets. For example, DEFs have been applied to textual data (newspaper articles), binary outcomes (clicks) and counts (movie ratings), being found to give better predictive performance than state-of-the-art models (Ranganath et al. 2015).

C Model priors and automatic relevance determination component

We detail the specification of the automatic relevance determination component that creates sparcity in the weights $\mathbf{W}^y$, $\mathbf{W}^a$, and $\mathbf{W}^1$ and the prior distribution.

C.1 Automatic relevance determination

Following Bishop (1999) we define $\boldsymbol{\alpha}$ as a positive vector of length N_1 (number of traits in the lower layer z_i^1), to control the activation of each trait. Note that $\mathbf{W}^y$ is matrix of size $D_y \times N_1$, where D_y is the length of the demand parameters $\boldsymbol{\beta}_i^y$; and $\mathbf{W}^a$ is matrix of size $P \times N_1$, where P is the length of the acquisition parameters $\boldsymbol{\beta}_i^a$.

We assume that the component associated with the n'th row (demand parameter) and k'th column (trait) of $\mathbf{W}^y$ is modeled by:

$$p(\mathbf{w}_{nk}^y) = \mathcal{N}\left(\mathbf{w}_{nk}^y | 0, \sigma^y \cdot \alpha_k\right) \tag{C.14}$$

where σ^y is the parameter that captures the variance of the demand model outcome (e.g., the variance of the error term in a linear regression). For identification purposes, we assume $\sigma^y = 1$ for logistic regressions. For other demand models, σ^y controls the scale of $\mathbf{W}^y$, and therefore should be defined accordingly. Note that if the vector of covariates $\mathbf{x}_{it}^y$ is not standardized, then this distribution should also consider the scale of the covariates.

Similarly, we model $\mathbf{w}_{pk}^a$, the component associated with the p'th row (acquisition behavior) and k'th column (trait) $\mathbf{W}^a$, by:

$$p(\mathbf{w}_{pk}^a) = \begin{cases} \mathcal{N}\left(\mathbf{w}_{pk}^a | 0, \alpha_k\right) & \text{if } p \text{ is discrete} \\ \mathcal{N}\left(\mathbf{w}_{pk}^a | 0, \sigma_p^a \cdot \alpha_k\right) & \text{if } p \text{ is continuous} \end{cases}, \tag{C.15}$$

where σ_p^a is the variance of the error term in the acquisition model for variable p. This variable again corrects for the scale of $\mathbf{w}_{pk}^a$ so it matches the scale of acquisition behavior p.

Finally, note that matrix $\mathbf{W}^1$ is of size $N_1 \times N_2$. We model $\mathbf{w}_{km}^1$, the component associated with the k'th row (lower layer) and m'th column (higher layer) of $\mathbf{W}^1$, using a sparse gamma

distribution:

$$p(\mathbf{w}_{km}^1) = \text{Gamma}\left(\mathbf{w}_{km}^1 | 0.1, 0.3\right) \tag{C.16}$$

C.2 Model priors

We model the prior distribution of the set of parameters using

$$\begin{aligned} p(\mathbf{W}^y, \mathbf{W}^a, \boldsymbol{\alpha}, \mathbf{W}^1, \boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}^a) =& p(\mathbf{W}^y, \mathbf{W}^a, \boldsymbol{\alpha}, \mathbf{W}^1, \boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}_a) \\ =& p(\mathbf{W}^y | \boldsymbol{\alpha}, \boldsymbol{\sigma}^y) \cdot p(\mathbf{W}^a | \boldsymbol{\alpha}, \boldsymbol{\sigma}^a) \cdot p(\mathbf{W}^1) \cdot p(\boldsymbol{\alpha}) \\ & \cdot p(\boldsymbol{\mu}^y) \cdot p(\boldsymbol{\mu}^a) \cdot p(\boldsymbol{\sigma}^y) \cdot p(\boldsymbol{\sigma}^a) \cdot p(\mathbf{b}^a) \end{aligned}$$

In our estimated models, $\boldsymbol{\sigma}^y$ is a positive scalar σ^y when the demand model is a regression and it does not exist when the demand model is a logistic regression; and $\boldsymbol{\sigma}_p^a$ is a positive scalar σ_p^a if the p'th acquisition behavior is continuous, and it does not exists if it is discrete. We use the automatic relevance determination component, described in Appendix C.1, for the terms $p(\mathbf{W}^y|\boldsymbol{\alpha}, \boldsymbol{\sigma}^y)$, $p(\mathbf{W}^a|\boldsymbol{\alpha}, \boldsymbol{\sigma}^a)$, and $p(\mathbf{W}^1)$. Denoting N_{ac} the number of firm-level controls for the acquisition model (i.e., dimension of $\mathbf{x}_{m\tau}^a$), and P_c the number of discrete acquisition variables, we model the remaining terms by:

$$\begin{aligned} p(\boldsymbol{\alpha}) &= \prod_{k=1}^{N_1} \text{InverseGamma}(\alpha_k | 1, 1), \\ p(\boldsymbol{\mu}^y) &= \prod_{k=1}^{D_y} \mathcal{N}(\mu_k^y | 0, 5), \\ p(\boldsymbol{\mu}^a) &= \prod_{p=1}^{P} \mathcal{N}(\mu_p^a | 0, 5), \\ p(\mathbf{b}^a) &= \prod_{n=1}^{N_{ac}} \prod_{p=1}^{P} \mathcal{N}(b_{np}^a | 0, 5), \\ p(\sigma^y) &= \log \mathcal{N}(\sigma^y | 0, 1), \qquad \text{(if demand model is a regression)}, \\ p(\boldsymbol{\sigma}^a) &= \prod_{p=1}^{P_c} \log \mathcal{N}(\sigma_p^a | 0, 1) \end{aligned} \tag{C.17}$$

D Further details about the simulation analyses

In this appendix we provide further details about the simulation exercise described in Section 4.4 (Model performance).

D.1 Simulation design

We simulate demand and acquisition behavior for $2,200$ customers. We first simulate acquisition and demand parameters ($\boldsymbol{\beta}_i^a$ and $\boldsymbol{\beta}_i^y$ respectively), and then use those to simulate the observed behaviors (A_i and $y_{i1:T}$ respectively). The data from 2,000 customers will be used to calibrate the models while the remaining 200 individuals will be used to evaluate the performance of each of the estimated models. For those (hold out) customers, we will assume that only the acquisition characteristics are observed, we will use each estimated model to infer customers' demand parameters and then will compare those inferences with the true parameters.

For our simulation study, we assume that acquisition and demand parameters are correlated, that is, observing acquisition behavior can partially inform demand parameters. For this purpose, we generate the individual demand parameters as a function of the acquisition parameters. To cover a variety of relationships among variables we use a linear, quadratic/interactions, and a positive-part (i.e., max) function, therefore exploring linear as well as non-linear relationships. Furthermore, to test whether the model can account for redundancy and irrelevance of variables in the acquisition characteristics collected by the firm, we assume that some acquisition variables are correlated among them and that other acquisition variables are totally independent of future demand. For clarity of exposition and brevity's sake, we first assume a small number of acquisition variables. Because many empirical contexts will likely have a large number of acquisition variables, we then extend the analysis to incorporate dozens of variables and show how the model performs at a larger scale.

D.2 Data generation process

Generate individual-level parameters

First, we generate seven acquisition parameters for seven corresponding acquisition characteristics. In order to resemble what real data would look like, and to test whether our model can account for redundancy in the acquisition data (e.g., the number of items purchased and total

amount spent at acquisition being highly correlated), we make some of these acquisition parameters highly correlated among themselves. We operationalize such a relationship by assuming that six of the seven parameters are driven by two main factors $\mathbf{f}_i = \begin{pmatrix} f_{i1} \\ f_{i2} \end{pmatrix}$, where $\mathbf{f}_i \sim N(0, I_2)$. Furthermore, we set the seventh acquisition parameter to be independent of other acquisition parameters as well as independent to future demand parameters. The rationale behind this structure is to resemble the situation in which the acquisition data includes irrelevant data and therefore test whether the model is robust to random noise. More specifically,

$$\begin{aligned} \beta^a_{ip} &\sim N\left(\mu^a_p + B_{1p} \cdot f_{i1}\ , \sigma^{ba}_p\right), && p = 1, \ldots, 3 \\ \beta^a_{ip} &\sim N\left(\mu^a_p + B_{2p} \cdot f_{i2}\ , \sigma^{ba}_p\right), && p = 4, \ldots, 6 \\ \beta^a_{i7} &\sim N\left(\mu^a_7\ , \sigma^{ba}_p\right), \end{aligned} \tag{D.18}$$

where β^a_{ip} is the p^{th} component of acquisition vector $\boldsymbol{\beta}^a_i$, μ^a_p is the mean of the p^{th} acquisition parameter; B_{1p} and B_{2p} represent the impact of factors 1 and 2 respectively on the p^{th} acquisition parameter; and $\sigma_{ba} = 0.1$ the standard deviation of the uncorrelated variation of the p^{th} acquisition parameter. The values used to generate factors f_{i1} and f_{i2} are presented in Table D.2.

Table D.2: True values for factors f_{i1} and f_{i2} impact on acquisition parameters (B_{1p} and B_{2p}).

Acquisition parameter	**Weight factors**	
	B_{1p}	B_{2p}
Factor 1, f_{i1}		
Acq. variable 1	3.0	0.0
Acq. variable 2	2.0	0.0
Acq. variable 3	-2.5	0.0
Factor 2, f_{i2}		
Acq. variable 4	0.0	3.5
Acq. variable 5	0.0	-2.0
Acq. variable 6	0.0	-3.0
Independent		
Acq. variable 7	0.0	0.0

Second, we generate the individual customer parameters for demand; these are the values that the firm is interested in inferring ($\boldsymbol{\beta}^y_i$. We generate three parameters governing the demand model: an intercept and two covariate effects. We generate these individual demand parameters

β_{ik}^{y} as a function of the acquisition parameters $\boldsymbol{\beta}_i^a$, following a general form

$$\beta_{ik}^{y} \sim N\left(\mu_k^y + g_k(\boldsymbol{\beta}_i^a|\Omega_k), \sigma_k^{by}\right), \qquad k = 1, \ldots, 3, \tag{D.19}$$

where $g_k(\boldsymbol{\beta}_i^a|\Omega_k)$ is the function that represents the relationship between acquisition and demand parameters. Because our goal is to investigate the accuracy of the model (compared to several benchmarks) in contexts in which the relationship between acquisition and demand parameters could take different forms, we vary g_k to capture a variety of scenarios:

- Scenario 1: Linear

$$g_k(\boldsymbol{\beta}_i^a|\Omega_k) = \boldsymbol{\omega}_k^{1\prime} \cdot \boldsymbol{\beta}_i^a \tag{D.20}$$

This relationship would exist when, for example, customers with a strong preference for discounted products at the moment of acquisition are also more likely to be price sensitive in future purchases.

Table D.3: Simulated values for $\boldsymbol{\omega}_k^1$ in the Linear scenario

Variable	**Demand variables**		
	Intercept	Covariate 1	Covariate 2
ω_{k1}^1	0.30	-0.69	-0.03
ω_{k2}^1	0.86	-0.61	-1.37
ω_{k3}^1	-1.44	-0.35	-0.03
ω_{k4}^1	-0.05	-0.10	0.12
ω_{k5}^1	1.16	-0.06	0.71
ω_{k6}^1	-0.12	0.10	0.93

- Scenario 2: Quadratic/interactions

$$g_k(\boldsymbol{\beta}_i^a|\Omega_k) = \boldsymbol{\omega}_k^{1\prime} \cdot \boldsymbol{\beta}_i^a + \boldsymbol{\beta}_i^{a\prime} \cdot \Omega_k^2 \cdot \boldsymbol{\beta}_i^a \tag{D.21}$$

This pattern captures situations in which the relationship between an acquisition variable and future demand depends on other acquisition-related parameters, or when such a relationship is quadratic. For example, it is possible that a strong preference for discounted products

at the acquisition moment relates to price sensitivity in future demand *only* if the customer was purchasing for herself/himself, or outside the holiday period. In that case, the relationship between demand parameters and acquisition variables will be best represented by an interaction term.

Table D.4: Simulated values for $\boldsymbol{\omega}_k^1$ and Ω_k^2 in the Quadratic/Interaction scenario

Variable	**Demand variables**		
	Intercept	Covariate 1	Covariate 2
ω_{k1}^1	0.30	-0.69	-0.05
ω_{k2}^1	0.86	-0.61	-1.04
ω_{k5}^1	1.16	-0.06	0.36
ω_{k3}^1	-1.44	-0.35	-0.27
ω_{k4}^1	-0.05	-0.10	0.10
ω_{k6}^1	-0.12	0.10	-1.11
Ω_{k11}^2	-0.01	0.06	0.00
Ω_{k22}^2	0.41	0.34	0.00
Ω_{k33}^2	-0.01	0.05	0.00
Ω_{k44}^2	0.01	-0.04	0.00
Ω_{k55}^2	0.17	-0.24	0.00
Ω_{k66}^2	-0.21	-0.11	0.00
Ω_{k12}^2	-0.36	-0.27	0.00
Ω_{k13}^2	-0.01	0.12	0.00
Ω_{k14}^2	-0.05	-0.01	0.00
Ω_{k15}^2	0.11	-0.08	0.00
Ω_{k16}^2	0.08	-0.16	0.00
Ω_{k23}^2	-0.01	-0.18	0.00
Ω_{k24}^2	0.24	0.10	0.00
Ω_{k25}^2	-0.24	-0.29	0.00
Ω_{k26}^2	-0.06	0.04	0.00
Ω_{k34}^2	0.17	0.07	0.00
Ω_{k35}^2	0.14	-0.14	0.00
Ω_{k36}^2	0.36	-0.10	0.00
Ω_{k45}^2	0.08	0.04	0.00
Ω_{k46}^2	-0.17	-0.15	0.00
Ω_{k56}^2	0.29	-0.17	0.00

- Scenario 3: Positive part

$$g_k(\boldsymbol{\beta}_i^a|\Omega_k) = \boldsymbol{\omega}_k^{1\prime} \cdot \begin{pmatrix} \max\{\boldsymbol{\beta}_{i1}^a, 0\} \\ \vdots \\ \max\{\boldsymbol{\beta}_{iP}^a, 0\} \end{pmatrix} \tag{D.22}$$

This pattern captures situations in which an acquisition variable relates to future demand parameters, but only if the former passes a certain threshold. For example, the number of items purchased at the moment of acquisition might relate to the likelihood of purchasing again in the category, but only above a certain threshold that reflects strong parameters for such a category.

Table D.5: Simulated values for $\boldsymbol{\omega}_k^1$ in the Positive part scenario

Variable	**Demand variables**		
	Intercept	Covariate 1	Covariate 2
ω_{k1}^1	0.34	0.00	0.30
ω_{k2}^1	0.00	0.00	0.86
ω_{k3}^1	0.00	0.00	-1.44
ω_{k4}^1	0.00	0.28	-0.05
ω_{k5}^1	0.00	0.00	1.16
ω_{k6}^1	0.00	0.00	-0.12

For each scenario, we generate the intercept (β_{i1}^y) and the effect of the first covariate (β_{i2}^y) according to the functions $g_1(\cdot)$ and $g_2(\cdot)$ as described in equations (D.20)–(D.22), while maintaining the effect of the second covariate (β_{i3}^y) to be a linear function of the acquisition variables. Furthermore, to compare parameters in the same scale across scenarios, we scale demand parameters such that the standard deviation across individuals is equal across all scenarios.

Simulate individual-level behaviors

Once the individual-level parameters are generated, we simulate behaviors using the generated acquisition and demand parameters for each scenario, a set of market-level covariates $\mathbf{x}_{m(i)}^a$ for the acquisition model, and individual and time-variant covariates $\mathbf{x}_{it}^y$ for the demand model. We assume

a Gaussian distribution for all behaviors,

$$A_{ip} \sim N(\beta^a_{ip} + {\mathbf{x}^a_{m(i)}}' \cdot \mathbf{b}^a_p, \sigma^a_p), \qquad p = 1, \ldots, 7 \tag{D.23}$$

$$y_{it} \sim N({\mathbf{x}^y_{it}}' \cdot \boldsymbol{\beta}^y_i, \sigma^y), \qquad t = 1, \ldots, 20. \tag{D.24}$$

with $\sigma^a = 0.5$, $\mathbf{x}^a_{m(i)} \sim \mathcal{N}(0, 1)$, $\mathbf{b}^a \sim \mathcal{N}(0, 2)$, $\sigma^y = 0.5$, and $\mathbf{x}^y_{it} \sim$ Bernoulli(0.5).

D.3 Estimated models

Given the observed behaviors (A_{ip} and y_{it}) and the covariates ($\mathbf{x}^a_{m(i)}$ and $\mathbf{x}^y_{it}$), we estimate the model parameters. In addition to our proposed FIM, we use four benchmark models to infer β^y_j: (1) a hierarchical Bayesian demand-only model in which acquisition variables are not incorporated, (2) a linear model, where individual demand parameters are a linear function of the acquisition characteristics, (3) a full hierarchical model, where individual demand and acquisition parameters are jointly distributed according to a multivariate Gaussian distribution with a flexible covariance matrix, and (4) a Bayesian PCA model, identical to our proposed model, without the higher layer. For all models we assume the same linear demand model as in the data generation process, equation (D.24). We describe these models in more detail.

D.3.1 Hierarchical Bayesian (HB) demand-only model This first benchmark is a *HB demand-only* model that does not incorporate acquisition variables. That is,

$$\boldsymbol{\beta}^y_i | \boldsymbol{\mu}^y, \Sigma^y \sim \mathcal{N}(\boldsymbol{\mu}^y, \Sigma^y),$$

where $\boldsymbol{\mu}^y$, and Σ^y are the population mean vector and covariance matrix respectively.

We acknowledge that such a model would fail to provide individual-level demand parameter estimates for customers that are not in the calibration sample. In other words, the best this model can provide is to draw the estimates from the population distribution. We include this benchmark to illustrate the problem of estimating parameters when only one observation per customer is observed and most importantly, to have a reference of how much error we should obtain if the model only captured random noise.

D.3.2 Linear HB model The second benchmark is the *linear HB model*, which is an extension of the previous model with the mean demand parameters being a linear function of the acquisition characteristics and market level covariates. That is,

$$\boldsymbol{\beta}_i^y = \boldsymbol{\mu}^y + \Gamma \cdot A_i + \Delta \cdot \mathbf{x}_{m(i)}^a + \mathbf{u}_i^y, \qquad \mathbf{u}_i^y \sim \mathcal{N}(0, \Sigma^y),$$

where Γ capture the linear explanatory power of acquisition characteristics A_i, and Δ allows to control for market-level covariates $\mathbf{x}_{m(i)}^a$.

In this model, we incorporate both acquisition characteristics as well as market-level covariates to control for firm's actions that may be correlated with acquisition characteristics (e.g. average price paid and promotional activity). Note that this model resembles the first simulated scenario in which the relationship between acquisition and demand parameters was assumed to be linear. As such, this model should be able to predict demand parameters in the first scenario most accurately.

D.3.3 Full hierarchical model For the third benchmark, we endogenize the acquisition characteristics by modeling them as an outcome. Similar to our proposed FIM (described in Section 4.1) (Model development), the full hierarchical model estimates acquisition and demand parameters jointly, with the difference that these two sets of parameters are modeled using a standard hierarchical model, rather than connected via DEF models. That is, the full hierarchical model assumes that

$$\boldsymbol{\beta}_i = \begin{pmatrix} \boldsymbol{\beta}_i^y \\ \boldsymbol{\beta}_i^a \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma),$$

where $\boldsymbol{\mu}$ is the population mean vector of all individual parameters (demand and acquisition), and Σ is the population covariance matrix of these parameters, capturing correlations within demand and acquisition parameters as well as across those types of parameters.

Because of the Gaussian specification for $\boldsymbol{\beta}_i$, this model imposes a linear relationship between $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$; this is, the conditional expectation of $\boldsymbol{\beta}_i^y$ given $\boldsymbol{\beta}_i^a$, is linear in $\boldsymbol{\beta}_i^a$. As such, this model is mathematically equivalent to the linear HB model. However, the full hierarchical model differs from the linear model if acquisition behavior A_i is not linear in $\boldsymbol{\beta}_i^a$ (e.g. logit or log-normal. Moreover, if the number of acquisition characteristics increases, the full hierarchical model becomes more

difficult to estimate due to the dimensionality of the covariance matrix. In this simulation exercise we assume a linear (Gaussian) acquisition model and therefore the linear and full hierarchical models should provide equivalent results. Nevertheless, this is not the case in the empirical application as we incorporate binary acquisition characteristics modeled using a logit specification.

D.3.4 Bayesian PCA The fourth benchmark is the closest to our proposed model, with the omission of the higher layer of traits ($\mathbf{z}_i^2$). Analogously as in our model, we model individual demand and acquisition parameters as a linear function of a set of traits,

$$\boldsymbol{\beta}_i^y = \boldsymbol{\mu}^y + \mathbf{W}^y \cdot \mathbf{z}_i^1 \tag{D.25}$$

$$\boldsymbol{\beta}_i^a = \boldsymbol{\mu}^a + \mathbf{W}^a \cdot \mathbf{z}_i^1. \tag{D.26}$$

In this Bayesian PCA model, we model the first layer $\mathbf{z}_i^1$ as a vector of independent standard Gaussian variables,

$$\mathbf{z}_{ik}^1 \sim \mathcal{N}(0, 1).$$

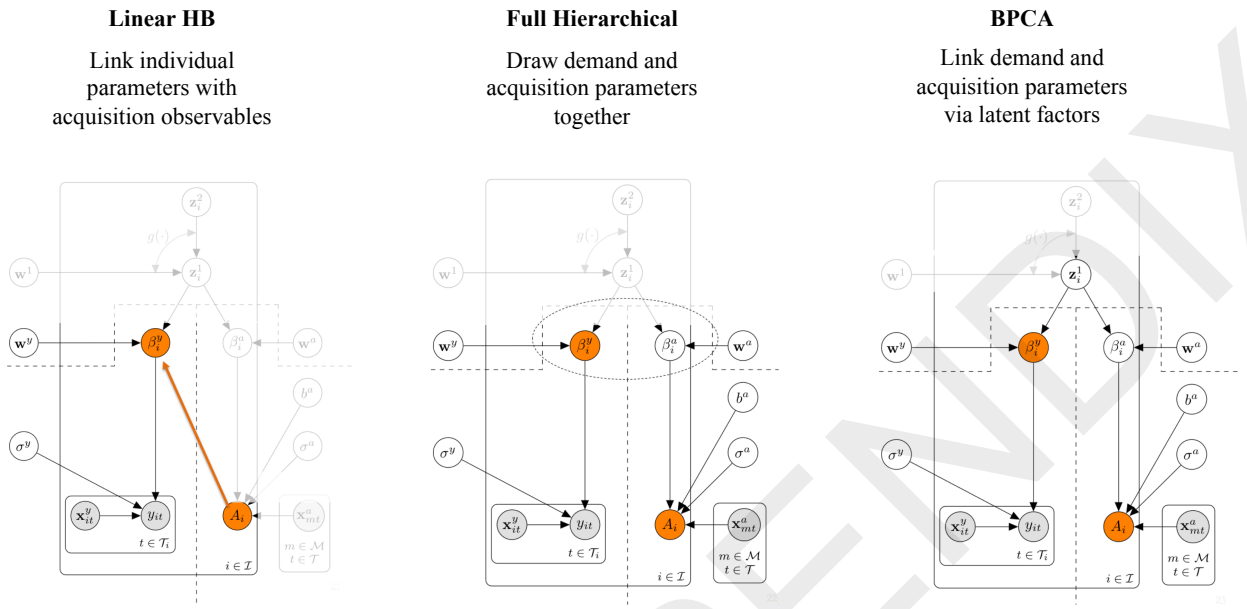
Note that like the linear HB and full hierarchical specifications, the PCA also imposes a linear relationship between $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$. However this approach is different from those because it allows for data dimensionality reduction via the latent factors. Similarly, as in our proposed model, we use sparse Gaussian priors on $\mathbf{W}^y$ and $\mathbf{W}^a$, using an automatic relevance determination model to automatically select the number of traits.

As discussed in Section 4.1.5 (Bringing it all together), the Bayesian PCA model is a nested specification of the proposed FIM (in which the second layer does not exist) whereas the full hierarchical model and HB-linear specifications reflect alternative (simpler) ways in which past research has modeled these types of data. Figure D.3 visually shows how each of these approaches compares with our proposed modeling framework.

D.4 Assessing model performance

We calibrate each model using acquisition and demand data for $2,000$ customers. This step resembles the firm calibrating each of the models (our proposed model as well as the benchmark models) with the historical data. First, we corroborate that all models are equally capable of recovering the

Figure D.3: Visualization of the benchmark models



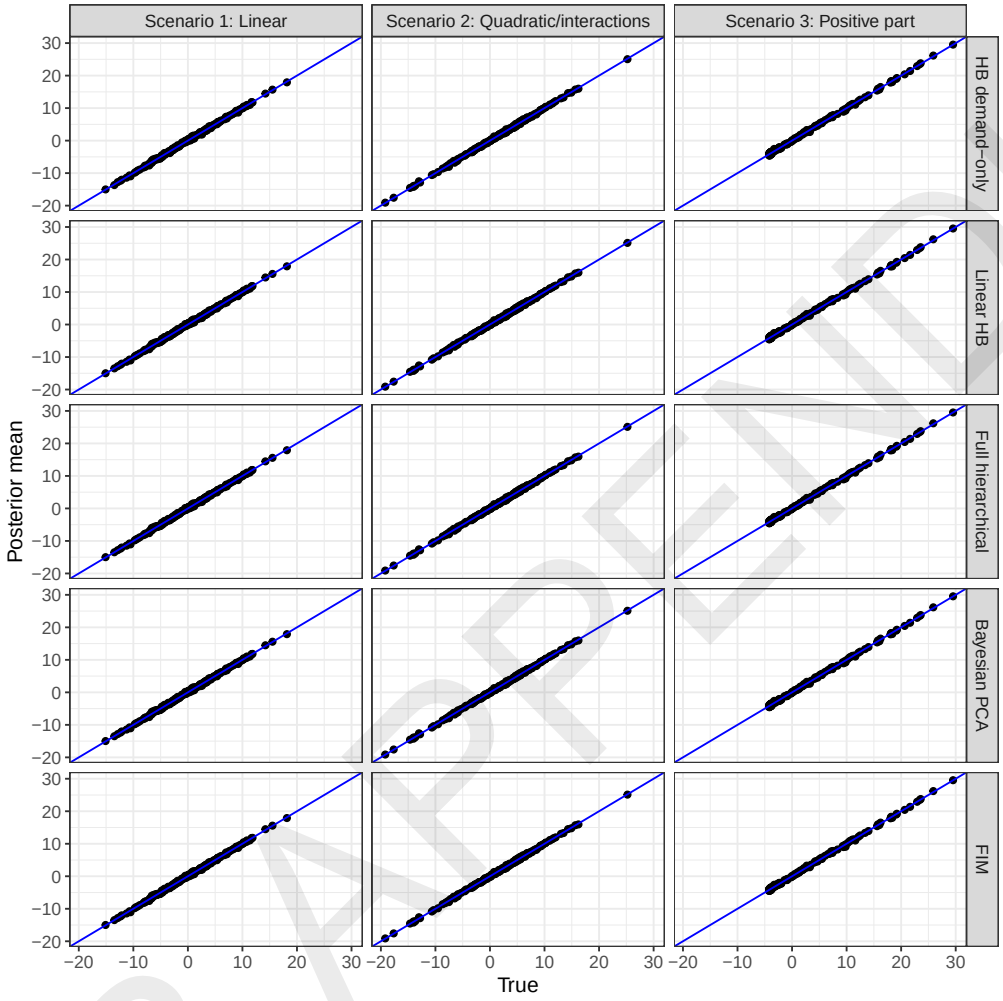
individual-level parameters for customers in the calibration sample. In particular, we confirm that the in-sample predictions for $\boldsymbol{\beta}_i^y$ are almost perfect for all model specifications and for all scenarios (see Figure D.4 for the in-sample predictions). In other words, all models are equally capable of accurately estimating individual-level demand parameters for in-sample customers.

Then, we evaluate the ability of each model to form first impressions of newly-acquired customers. Under each scenario, we use the estimates of each model to predict the individual-level demand parameters for the remaining 200 customers, using only their acquisition data, and compare those predictions with the true values. This task requires the computation of the individual posterior mean for each individual ($\widehat{\boldsymbol{\beta}}_j^y = E(\boldsymbol{\beta}_j^y | A_j, \mathcal{D})$) by integrating over the estimated density $p(\boldsymbol{\beta}_j^y | A_j, \mathcal{D})$,

$$\widehat{\boldsymbol{\beta}}_j^y = \int \boldsymbol{\beta}_j^y \cdot p(\boldsymbol{\beta}_j^y | A_j, \mathcal{D}) d\boldsymbol{\beta}_j^y.$$

While the procedure described in Section 4.3 is valid for all models, the expectation $E(\boldsymbol{\beta}_j^y | A_j, \mathcal{D})$ can be computed directly for some of the benchmark models, which we do for simplicity. For example, for the HB demand-only model, this procedure reduces to compute the expectation of individual draws of $\boldsymbol{\beta}_j^y$ from the population mean, which converges to the posterior mean of the population

Figure D.4: Individual posterior mean vs. true intercepts of the demand model. Each dot represents a customer from the calibration set. In blue, the 45 degree line represents perfect predictive power.



mean $\boldsymbol{\mu}^y$. For the linear HB model, it reduces to use the linear formulation and the posterior mean estimates of $\boldsymbol{\mu}^y$, Γ, and Δ. For the full hierarchical model, the Bayesian PCA model, and our proposed FIM, where acquisition is modeled as an outcome, we compute the posterior of $\boldsymbol{\beta}_j^y$ given A_j using HMC as described in Section 4.3 (Model inferences for newly acquired customers).

Figure D.5 shows the scatter plot of the predicted ($\widehat{\beta}_{j1}^y$) versus actual (β_{j1}^y) individual demand intercepts from each model, for each scenario.[2] Not surprisingly, the HB demand-only model that does not incorporate acquisition behavior in the model (top row of Figure D.5) cannot distinguish

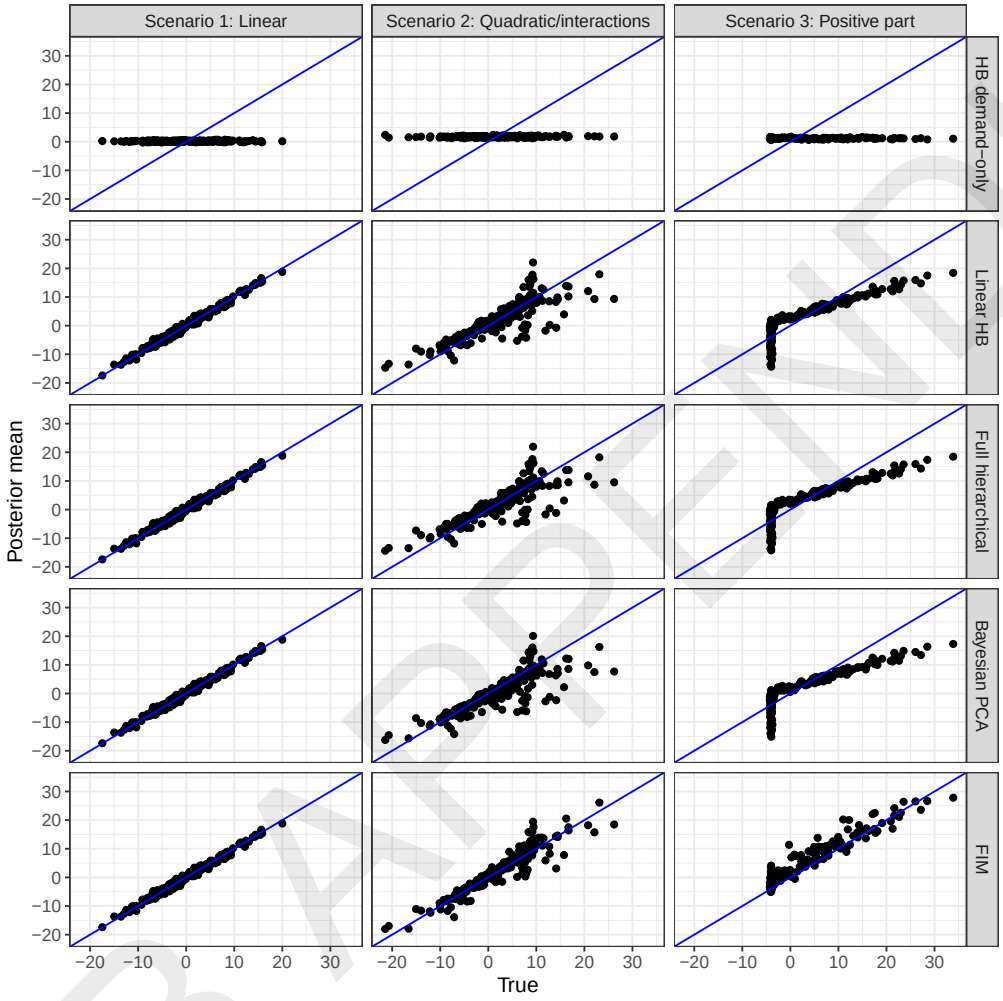
[2]For brevity's sake, we present the results for one parameter of the demand model (the intercept), but the results hold for all other parameters as well.

(hold out) individuals from their population mean. Turning our attention to the other model specifications, we start analyzing the scenario in which the relationship between acquisition and demand parameters is linear (left-most column of Figure D.5). Under this scenario, all models are equally capable of predicting demand estimates for (hold out) customers using only their acquisition data. This result is not surprising for the benchmark models as their mathematical specification resembles that of the simulated data. However, when the relationship between the acquisition and demand parameters is not perfectly linear (as it is the case in scenarios 2 and 3), all benchmark models struggle to predict these individual-level estimates accurately. On the contrary, the proposed FIM is flexible enough to recover these parameters rather accurately. Note that the flexibility of the FIM comes at no overfitting cost; that is, even when the relationship is a simple linear relationship, our model recovers the parameters as well as the benchmark models, which assume a linear relationship by construction.

To explore the differences in accuracy more systematically, we compute two different measures of fit: (1) the (squared) correlation between true $\boldsymbol{\beta}_j^y$ and predicted $\widehat{\boldsymbol{\beta}}_j^y$ (i.e., R-squared) — measuring the model's accuracy in sorting customers (e.g., differentiating customers with high vs. low value, more vs. less sensitivity to marketing actions) — and the root mean square error (RMSE) — measuring the accuracy on predicting the value/magnitude of the parameter itself.

The results are presented in Table 4 of the main manuscript, confirming the results from Figure D.5. Under a true linear relationship (Scenario 1), the FIM predicts the individual parameters as good as the benchmark models. The RMSE of the FIM is comparable to the benchmark models, and the R-squared is equal to the benchmark models. However, when the relationship among the model parameters is not perfectly linear (Scenarios 2 and 3), the FIM significantly outperforms the benchmark models in all dimensions. In particular, the R-squared of the FIM is higher than that of the benchmarks, demonstrating that the model is superior at sorting customers based on their demand parameters. Moreover, the RMSE for the FIM is substantially lower than that of the benchmarks, indicating that the proposed model predicts the exact magnitude of customer parameters (e.g., purchase probability, sensitivity to marketing actions) more accurately than any of the benchmarks.

Figure D.5: Individual posterior mean vs. true intercepts of the demand model. Each dot represents a customer from the hold out set; i.e., only their acquisition characteristics are used to form first impressions about their individual-level parameters. In blue, the 45 degree line represents perfect predictive power.



D.5 Interpreting the model parameters and results

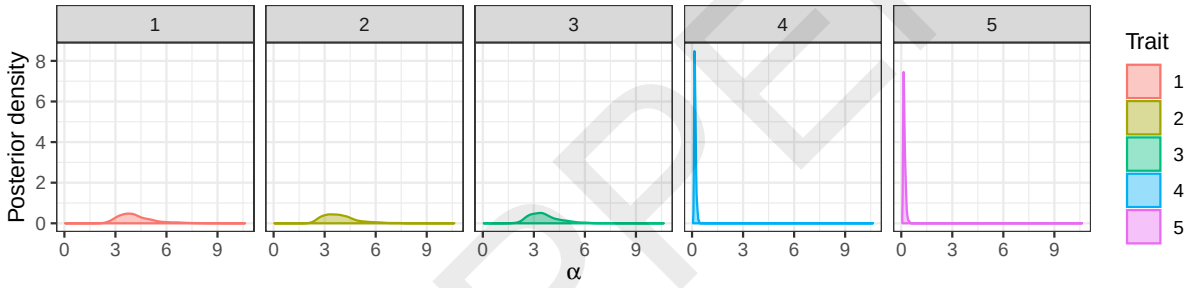
To get a better sense of what the model is doing and what its parameters capture, we explore in detail the model estimates and compare those with the parameters used to simulate the data. We do so for the linear case, as it is the easiest to interpret the relationships among variables. For this particular exercise, we select the FIM with 5 dimensions in the lower layer and 3 in the top layer.[3] We start by evaluating the number of traits captured by the FIM; this is an insight that can be obtained in two ways. First, looking at the posterior estimates for $\boldsymbol{\alpha}$, parameters that determined

[3]Results are equivalent for other specifications of the model.

the weights of the lower layer to check how many dimensions of the lower layer are activated in the model. Second, by looking at the specific weights, $\mathbf{W}^y$ and $\mathbf{W}^a$, between the lower layer and the model parameters and interpret their meaning based on their magnitude.

We know from the simulations (Appendix D.1) that the data were generated from three factors: two factors generating 6 acquisition characteristics that relate to demand parameters, and another independent factor that generated one acquisition variable that was irrelevant for the demand model. Figure D.6 shows the posterior distribution for $\boldsymbol{\alpha}$. While the model was specified to have 5 dimensions in the lower layer, it is obvious that the model only "needs" three, one of which is irrelevant in the demand specification.

Figure D.6: Posterior distribution of $\boldsymbol{\alpha}$



We show in Table D.6 the posterior mean of the rotated weight traits on demand parameters and acquisition parameters. The first two traits capture most of the variance across individuals for demand and acquisition parameters, while the other traits capture residual variance. First, trait 1 captures the associations among acquisition variables 1 through 3, whereas trait 2 captures the associations of acquisition variables 4 through 6. Second, both traits capture relationships with demand: trait 1 is negatively correlated with intercept and positively correlated with both covariates, whereas trait 2 is negatively correlated with intercept and covariate 2 (effect on covariate 1 is not significantly different from zero).

Table D.6: Posterior mean of lower layer weights ($\mathbf{W}^y$ and $\mathbf{W}^a$) for FIM.

Variable	Trait 1	Trait 2	Trait 3	Trait 4	Trait 5
Intercept	**-5.55**	**-2.14**	**0.04**	0.00	-0.00
Covariate 1	**2.28**	-0.53	**0.10**	-0.00	0.00
Covariate 2	**2.91**	**-3.63**	-0.04	-0.00	0.00
Acq. variable 1	**-2.78**	0.07	-0.04	-0.01	0.00
Acq. variable 2	**-1.84**	-0.03	0.02	0.00	0.00
Acq. variable 3	**2.30**	0.05	-0.02	0.00	-0.00
Acq. variable 4	-0.31	**3.40**	0.02	-0.01	0.01
Acq. variable 5	0.18	**-1.95**	0.00	0.01	**0.02**
Acq. variable 6	0.26	**-2.91**	-0.05	0.01	0.01
Acq. variable 7	-0.01	0.02	-0.03	-0.02	0.01

Note: In bold parameters such that corresponding CPI do not contain zero

Now, we are interested in comparing these insights with the true values used for the simulation, specifically how these estimated traits relate to the true factors in the data generation process. In the data generation process, demand parameters are generated from acquisition parameters. Instead, the FIM gives us the overall associations of the traits with demand parameters, and not the one-to-one relationships between acquisition variables and demand parameters. Therefore, in order to assess whether our model can capture the essence of the insights the "true" effect of factors f_{i1} and f_{i2} on acquisition parameters and demand parameters in Table D.7. For the acquisition parameters, these true effects are B_{1p} and B_{2p} from (D.18) (whose values are shown in Table D.2). For the demand parameters, these effects can be obtained by replacing (D.18) in (D.19), which reduces to $\boldsymbol{\omega}_k^{1\prime} B_1$ and $\boldsymbol{\omega}_k^{1\prime} B_2$ for the effects of factors 1 and 2, respectively.

Table D.7: True associated effects of factors on demand and acquisition variables.

Demand/acquisition parameter	Variable	Factors	
		1	**2**
Intercept	$\boldsymbol{\omega}_1^{1\prime} B_f$	6.20	-2.10
Covariate 1	$\boldsymbol{\omega}_2^{1\prime} B_f$	-2.40	-0.57
Covariate 2	$\boldsymbol{\omega}_3^{1\prime} B_f$	-2.77	-3.76
Acq. variable 1	B_{f1}	3.00	0.10
Acq. variable 2	B_{f2}	2.00	0.00
Acq. variable 3	B_{f3}	-2.50	0.00
Acq. variable 4	B_{f4}	0.00	3.50
Acq. variable 5	B_{f5}	0.00	-2.00
Acq. variable 6	B_{f6}	0.00	-3.00
Acq. variable 7	B_{f7}	0.00	0.00

By comparing Tables D.6 and D.7 we observe that: (1) trait 1 captures the reverse of factor 1 ($\widehat{z}_{i1}^1 \approx -f_{i1}$); and (2) trait 2 captures factor 2 ($\widehat{z}_{i2}^1 \approx f_{i2}$). This result implies that our model is able to capture and deliver meaningful insights that relate to the true data generation process.

D.6 Why is the model giving superior performance?

A natural question to ask is, why is the proposed model outperforming the benchmark models? As described in Section 4.1 (Model development), the DEF component of the proposed model is very flexible at capturing underlying relationships between the model parameters. Such a property enables the model to capture non-linear relationships between acquisition characteristics and the parameters that drive customer demand. This is unlike the benchmarks whose specification imposes a linear relationship among the variables. As such, even though the in-sample predictions of all the models are very accurate (Figure D.4), when any of the benchmark models are used to make (out-of-sample) predictions for newly-acquired customers, the predicted values differ dramatically from the actual values (Figure D.5).

Table D.8: Squared correlation (true vs predicted) for Covariate 1; Quadratic/Interaction Scenario.

	Dim. Upper layer		
	Bayesian PCA	FIM	
Dim. Lower layer	0	1	2
1	0.209	0.207	0.209
2	0.237	0.304	0.306
3	0.257	0.402	0.404
4	0.250	0.539	0.425
5	0.252	0.538	0.641
6	0.250	0.509	0.612
7	0.250	0.451	0.627
8	0.243	0.525	0.571

To better corroborate that it is the DEF component that brings the non-linearities, we compare in greater detail the predictions of the BPCA model with those of the FIM. We pick the BPCA (among the other benchmarks) because that is the only model that is mathematically nested to our proposed model. In turn, the BPCA is the closest to the FIM, with the difference that it does not have an upper layer (and its corresponding non-linear link function). Table D.8 shows the squared correlation (true vs. predicted) for Covariate 1 of the second scenario (Quadratic/Interaction), for the BPCA and the FIM models, as we vary the number of dimensions. The first column corresponds to the fit of the BPCA model, as we increase the number of dimensions. We see an improvement in fit as we increase the number of dimensions from 1 to 2, and to 3; and no improvement after that, with the best fit obtained being around 0.25. However, the jump in fit is tremendous when we allow the model to have an upper layer (even if it only includes 1 dimension).[4] Such an upper layer is the model component that allows for flexible relationships relationships. The same results hold when looking at the third scenario (Positive-part).

To conclude, the upper layer of the DEF — the component that allows the model to capture non-linear relationships among variables — is responsible for the great improvement in the model's ability to predict (out-of-sample) individual-level parameters when the underlying relationship between acquisition characteristic and the demand parameter is not linear.

[4]We discuss the importance of the dimensionality of the upper layer in Appendix D.7.

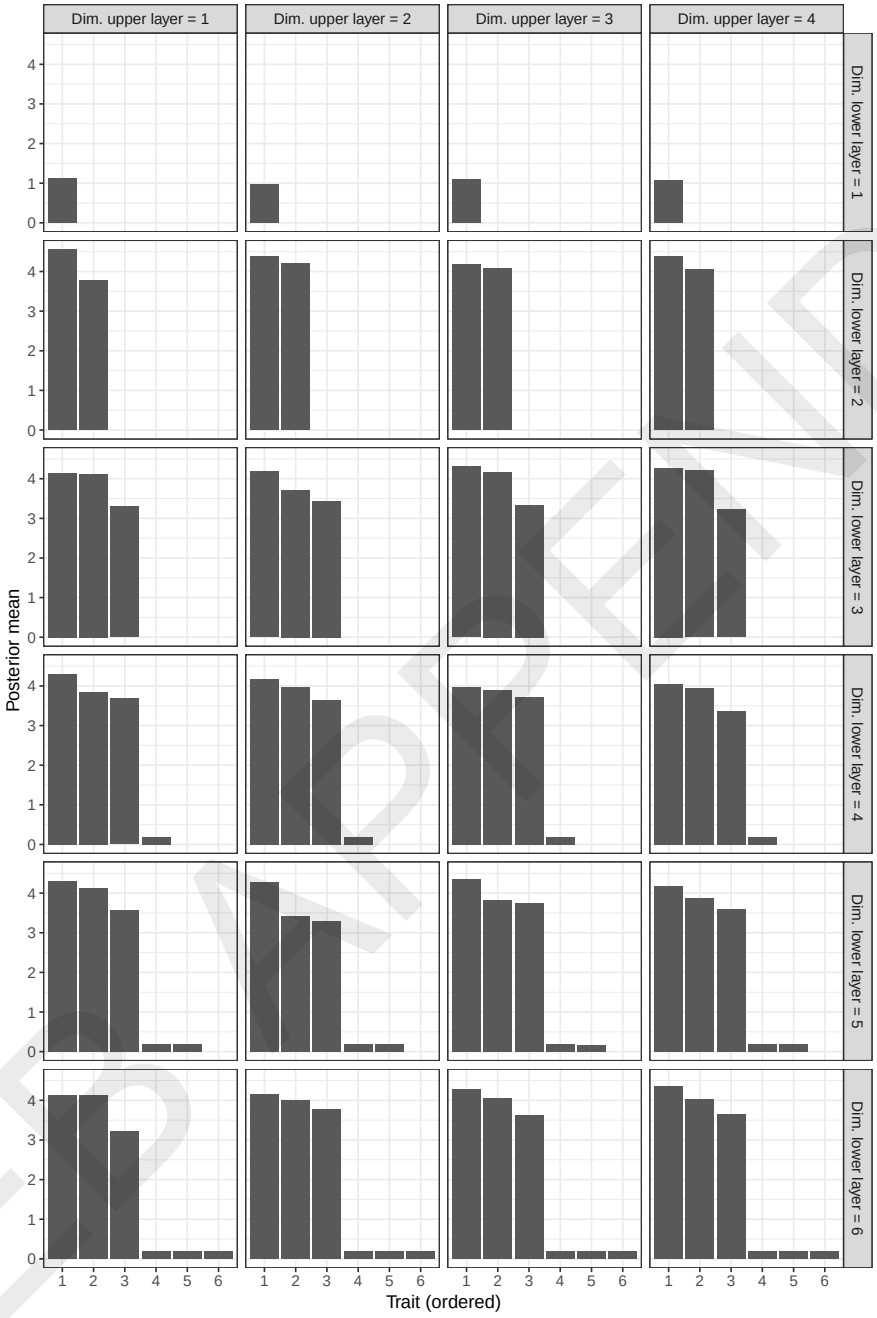
D.7 Exploring the number of dimensions per layer

As described in Section 4.1.3 (Linking acquisition and future demand: Deep probabilistic model), we take a hybrid approach to model selection in which we make sure that the number of pre-specified dimensions is large enough — a phenomenon that can be validated from the model parameters — while we rely on the priors of the model to ensure regularization. In this appendix we leverage the simulation results to provide further details about the model selection procedure and to corroborate the two premises that drive our model selection approach. Specifically, we present empirical evidence that (a) one can ensure that the model has a "large enough" number of dimensions by examining the posteriors of the Gaussian ARD priors, and (b) as long as the layers have enough dimensions to capture meaningful interrelations and priors induce sparsity on the weight traits, increasing the number of dimensions on each layer would only lead to higher computational cost, without the corresponding loss in out of sample performance.

To illustrate how one can use the posterior of the Gaussian ARD priors to ensure that the number of dimensions is "large enough," we revisit the model examined in Appendix D.5 in which the simulated behavior was generated by three factors, one of which had no impact on the demand parameters, and the FIM specification included 5 traits in the lower layer (e.g., $N_1 = 5$). As seen in that section, the FIM results not only recover that data generation process (Tables D.6 and D.6), but also informs of the number of dimensions in the lower layer (Figure D.6). In this appendix we expand the results presented in Figure D.6 by showing the posterior estimates for $\boldsymbol{\alpha}$ for FIM specifications with different values for N_1 and N_2 (Figure D.7).

As it is evident from the figure, the model detects that the data were generated from three latent traits (as long as the FIM is specified with $N_1 \geqslant 3$) and in cases where the FIM allows for larger dimensionality, the model "shuts down" the rest of the traits. In other words, regardless of the dimensionality of the top layer (N_2), when the number of traits in the lower layer is not enough, the model does not "shut down" any component. However, once N_1 is large enough (in this case $N_1 = 3$, as it was used to generate the data), the posterior mean of α_4, α_5 and so on, are all close to zero. These results corroborate that the posterior distribution of the Gaussian ARD variances can be used to show when the model has a "large enough" number of dimensions.

Figure D.7: Posterior mean of $\boldsymbol{\alpha}$ as a function of number of dimensions in lower layer (N_1) and upper layer (N_2). Components are sorted in decreasing order per model.



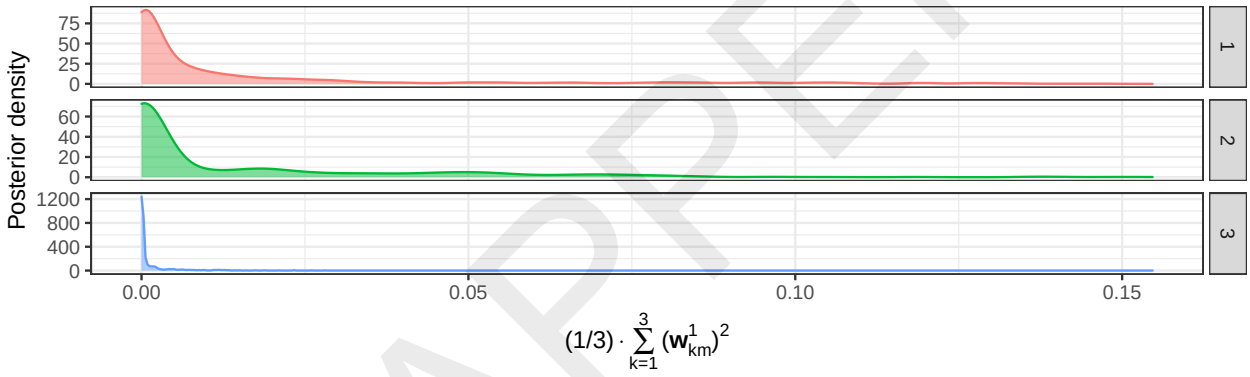
In contrast to the usefulness of α to detect relevant lower traits, our model does not have an analogous parameter to explore how many upper level traits are enough to capture the relevant non-linear interrelations. Instead, each component of the upper weight $\mathbf{W}^1$, $\mathbf{w}^1_{km}$ (for lower trait k and upper trait m), has i.i.d. sparse gamma priors, which by themselves induce regularization. In order to summarize each upper level trait in a way that can help us determine whether they make

an impact on those 6 relevant lower layer traits, we compute a pseudo-α_m^1 for each upper trait m using the weight matrix $\mathbf{W}^1$. Similarly to how the lower level weights $\mathbf{W}^y$ and $\mathbf{W}^a$ are related to $\boldsymbol{\alpha}$ (i.e., variance of zero-centered Gaussian distributions), we compute these pseudo-α^1's by averaging the square of all weights associated with a fixed upper level trait and those 6 relevant lower level traits, as described by

$$\text{pseudo-}\alpha_m^1 = \frac{1}{6}\sum_{k=1}^{6}\left(\mathbf{w}_{km}^1\right)^2.$$

We show the posterior this quantity in Figure D.8. Not surprisingly, these posterior distributions are concentrated close to the origin, which suggests that no upper trait is relevant for this scenario (as the data were generated linearly).

Figure D.8: Posterior distribution of pseudo-$\boldsymbol{\alpha}^1$ (Linear scenario).

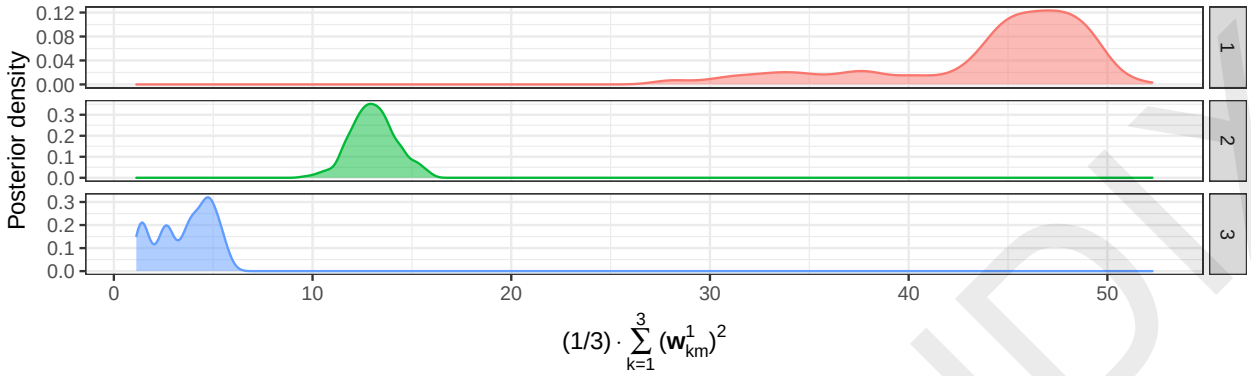


More interestingly, we further explore this quantity using a scenario in which the model requires to capture non-linear relationships, such as the one with Interactions. Figure D.9 shows the posterior of pseudo-α^1 for two FIM specifications with different values of N_2. First, Figure D.9a clearly shows that the FIM with $N_1 = 5$ and N_2 estimated for the Interactions scenario, unlike the FIM estimated using the linearly simulated data, has all three upper traits being relevant in the model. Second, if we estimate a FIM with more upper traits ($N_1 = 5$, $N_2 = 5$) the model starts to "shut down" the less relevant traits, indicating that such a model is enough to recover the non-linear relationships present in those data.

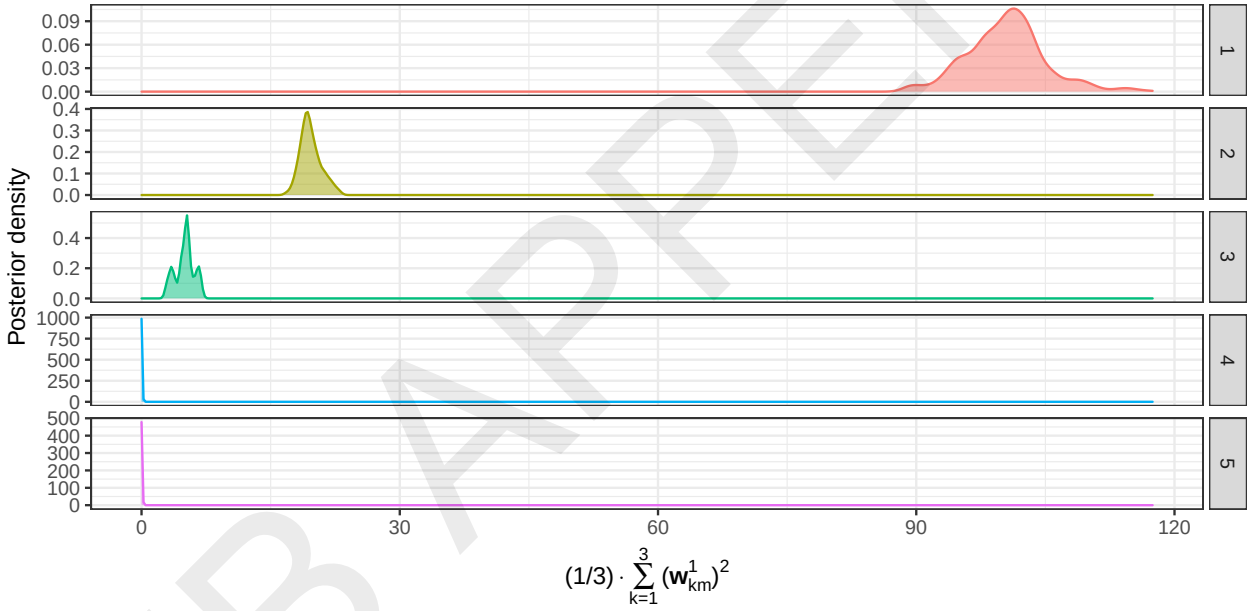
Finally, we leverage the results of multiple estimated FIM specifications over the Interactions scenario and show that once the FIM specification contains the dimensions "needed" by the data, the performance of the model remains the same even if we add dimensions to the DEF component.

Figure D.9: Posterior distribution of pseudo-$\boldsymbol{\alpha}^1$ (Interactions scenario).

(a) FIM ($N_1 = 5$, $N_2 = 3$).



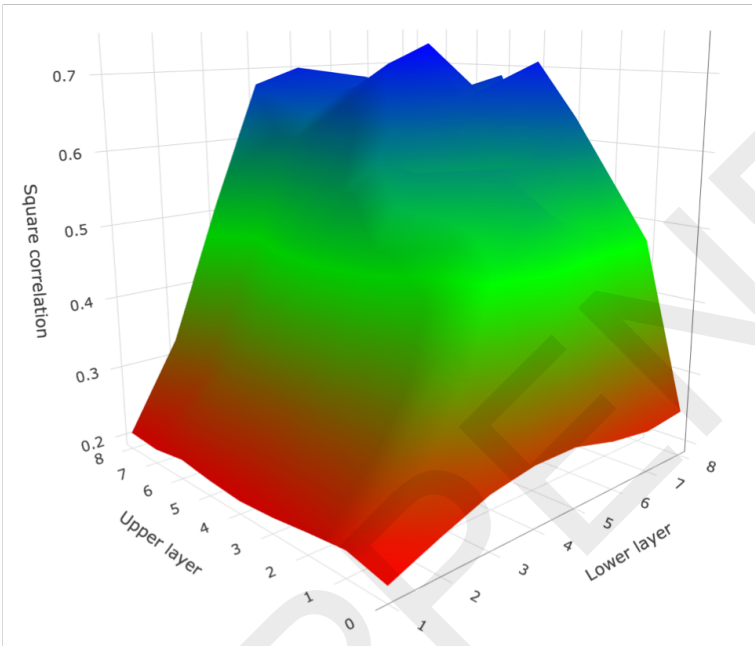
(b) FIM ($N_1 = 5$, $N_2 = 5$).



To illustrate this phenomenon, we focus on the performance of the FIM at predicting the parameter for the sensitivity to the first covariate (bottom half of middle columns in Table 4). Figure D.10 shows the squared correlation between simulated and predicted values of the parameter of interest (higher numbers imply better model performance). The figure shows a notable improvement in performance as we increase the dimensionality of the lower layer from 1 to 2, 3, and 4. However, once $N_1 > 3$, the model performs very similarly as more layers are added to DEF. Similarly, we observe a radical increase in performance as one increases the dimensionality of the upper layer (from 0 to 1, 2 and 3); reaching a point in which more dimensions do not alter the performance

of the model. In other words, the performance in out-of-sample recovery of demand parameters flattens, once the model has a "large enough" number of dimensions.

Figure D.10: Square correlation between simulated and predicted β for Covariate 1 in Scenario 2: Interaction



D.8 Model performance "at scale"

While the analysis thus far assumed a handful number of acquisition variables, many firms collect a larger quantity of behaviors when a customer makes their first transaction. These firms do not necessarily know a priori which variables can be most predictive of demand parameters, and if so, what the underlying relationship between these variables would be. In this section we show that models that incorporate all interactions fail to recover demand parameters when the number of acquisition variables is large, whereas the FIM can accurately infer these non-linear relationships. We maintain a similar simulation structure, where acquisition parameters are driven by factors, but instead we now have 5 factors and 60 acqusition behaviors, where acquisition behavior is driven by one and only one factor, and each factor generates 12 acquisition parameters. We start by describing the simulation details and their differences to the main analysis in Appendix D.1. Then, we describe the additional estimated models, specifically those that include interactions. Finally,

similarly as in Appendix D.4, we show the models' ability to infer demand parameters for out of sample customers.

D.8.1 Simulation details We assume there are 3 demand parameters (intercept and two covariates) and 60 acquisition parameters, for 60 acquisition characteristics. We generate these acquisition parameters as being highly correlated among each other by assuming these parameters are driven by one of five factors $f_{i1}, \ldots, f_{i5}$. Similarly as in Equation (D.18), we generate acquisition parameters by:

$$\begin{aligned}
\beta^a_{ip} &\sim N\left(\mu^a_p + B_{1p} \cdot f_{i1}\ , \sigma^b_p\right), && p = 1, \ldots, 12 \\
\beta^a_{ip} &\sim N\left(\mu^a_p + B_{2p} \cdot f_{i2}\ , \sigma^b_p\right), && p = 13, \ldots, 24 \\
\beta^a_{ip} &\sim N\left(\mu^a_p + B_{3p} \cdot f_{i3}\ , \sigma^b_p\right), && p = 25, \ldots, 36 \\
\beta^a_{ip} &\sim N\left(\mu^a_p + B_{4p} \cdot f_{i4}\ , \sigma^b_p\right), && p = 37, \ldots, 48 \\
\beta^a_{ip} &\sim N\left(\mu^a_p + B_{5p} \cdot f_{i5}\ , \sigma^b_p\right), && p = 49, \ldots, 60, \qquad \text{(D.27)}
\end{aligned}$$

where μ^a_p is the mean of the p^{th} acquisition parameter; $B_{\ell p}$ represent the impact of factor ℓ respectively on the p^{th} acquisition parameter; and σ_p the standard deviation of the uncorrelated variation of the p^{th} acquisition parameter.

The rest of the simulation design is identical as the simulation in Section 4.3 (Model inferences for newly acquired customers), with a different set of parameters Ω. In order to incorporate noise and to allow for different acquisition parameters to inform demand parameters, we relate demand parameters only to a subset of acquisition parameters. Specifically, we choose Ω such that demand parameters are only affected by acquisition parameters from three out of the five factors. We achieve this by setting to zero Ω values for the remaining acquisition parameters. The intercept is a function of the acquisition parameters from factors 1, 2 and 3 (i.e., $\Omega_{1p} = 0,\ \forall p = 37, \ldots, 60$). Covariate 1 is a function of the acquisition parameters from factors 1, 2 and 4 (i.e., $\Omega_{2p} = 0,\ \forall p = 25, \ldots, 36, 49, \ldots, 60$). Covariate 2 is a function of the acquisition parameters from factors 2, 3 and 4 (i.e., $\Omega_{3p} = 0,\ \forall p = 1, \ldots, 12, 49, \ldots, 60$). Similarly as in the main simulation analysis, Covariate

2 is always a linear function of acquisition parameters for all scenarios. The values we use for Ω are specific to each scenario:

- **Linear**: Following (D.20), we define $\omega^1_{kp} \sim \mathcal{N}(0, 2)$ for all non-zero ω^1_{kp}.
- **Quadratic/Interaction**: Following (D.21), we define $\omega^1_{kp} \sim \mathcal{N}(0, 2)$ for all non-zero ω^1_{kp}; and $\Omega^2_{kpp'} \sim \mathcal{N}(0, 1)$ for all non-zero $\Omega^2_{kpp'}$.
- **Positive part**: To avoid attenuating the effect of the non-linear function by combining a large number of non-linear functions of correlated acquisition parameters, we fix the effect to the intercept and the first covariate as a function of only one acquisition parameter from each of the three factors that determine that demand parameter. Following (D.22), we define $\omega^1_{3p} \sim \mathcal{N}(0, 2)$ for all non-zero ω^1_{3p}, and:

$$\omega^1_{1,1} = 12.5 \qquad \omega^1_{1,13} = -8 \qquad \omega^1_{1,25} = 4$$
$$\omega^1_{2,1} = -7.5 \qquad \omega^1_{2,13} = -4 \qquad \omega^1_{2,37} = 8.$$

Finally, to compare parameters in the same scale across scenarios, we standardize demand parameters such that the population standard deviation is 2.

D.8.2 Estimated models In addition to all models described in Appendix D.3, we estimate a Linear HB model where we include all interactions of acquisition parameters,

$$\boldsymbol{\beta}^y_i = \boldsymbol{\mu}^y + \Gamma \cdot \widetilde{A}_i + \Delta \cdot \mathbf{x}^a_{m(i)} + \mathbf{u}^y_i, \qquad \mathbf{u}^y_i \sim \mathcal{N}(0, \Sigma^y),$$

where $\widetilde{A}_i$ includes all acquisition characteristics, their squares, and all two-way interactions among them.

We also estimate a Lasso model with all interactions, which is identical to the Linear HB model with interactions, but we exchange the Gaussian prior for a Laplace prior to enforce regularization using a different functional form.

D.8.3 Results We estimate all models except the full hierarchical model, which is computationally unstable given that now there are 60 acquisition variables, and therefore we need a 63×63 covariance matrix. Note that in theory, and in practice as we showed in Appendix D.4, the full hierarchical model is equivalent to a Linear HB model. Therefore, removing this model from the analysis does not bias our benchmark.

We show in Table D.9 the out of sample prediction of intercept, and the two covariates under all three scenarios for all models. We replicate the main results from Appendix D.4. Both the Linear HB and Bayesian PCA models perform well in the Linear scenario. The FIM performs as good as these models in the Linear scenario, and outperforms these linear models in the Quadratic/Interaction and the Positive part scenarios. More importantly, both models that include all interactions, Linear and Lasso, do not perform well in any scenario.

Table D.9: Model at scale results

Model	Intercept		Covariate 1		Covariate 2	
	R-squared	RMSE	R-squared	RMSE	R-squared	RMSE
Linear						
HB demand-only	0.000	2.018	0.000	2.038	0.000	2.003
Linear HB	0.990	0.198	0.987	0.231	0.983	0.264
Linear with interactions	0.202	4.267	0.166	4.825	0.121	5.265
Lasso with interactions	0.161	5.916	0.115	6.129	0.108	5.561
Bayesian PCA	0.990	0.197	0.988	0.229	0.983	0.265
FIM	0.990	0.206	0.987	0.230	0.983	0.262
Quadratic/Interaction						
HB demand-only	0.004	2.060	0.000	2.133	0.007	2.084
Linear HB	0.231	1.808	0.398	1.663	0.994	0.167
Linear with interactions	0.147	4.064	0.201	4.331	0.246	4.125
Lasso with interactions	0.147	4.212	0.211	4.871	0.236	4.181
Bayesian PCA	0.243	1.790	0.408	1.646	0.994	0.167
FIM	0.598	1.456	0.681	1.432	0.994	0.165
Positive part						
HB demand-only	0.003	2.010	0.005	2.030	0.017	1.965
Linear HB	0.723	1.059	0.746	1.019	0.990	0.201
Linear with interactions	0.232	3.990	0.165	4.916	0.122	4.414
Lasso with interactions	0.161	4.493	0.088	5.336	0.186	5.032
Bayesian PCA	0.728	1.052	0.747	1.017	0.991	0.196
FIM	0.884	0.699	0.853	0.825	0.991	0.192

E Rotation of traits

In order to obtain insights about the traits, we post process the posterior sample by carefully rotating the lower weights parameters across draws to define a consistent sign and label of those traits.

First, we define the vectors $\boldsymbol{\beta}_i^{ya} = \begin{pmatrix} \boldsymbol{\beta}_i^y \\ \boldsymbol{\beta}_i^a \end{pmatrix}$, and $\boldsymbol{\mu}^{ya} = \begin{pmatrix} \boldsymbol{\mu}^y \\ \boldsymbol{\mu}^a \end{pmatrix}$ of length $(D_y + P)$, and the matrix $\mathbf{W}^{ya} = \begin{bmatrix} \mathbf{W}^y \\ \mathbf{W}^a \end{bmatrix}$ of size $(D_y + P) \times N_1$. Second, we rewrite (5) and (6) as:

$$\boldsymbol{\beta}_i^{ya} = \boldsymbol{\mu}^{ya} + \mathbf{W}^{ya} \cdot z_i^1. \tag{E.28}$$

Let D the number of posterior draws obtained using HMC, and $d = 1, \ldots, D$ one draw from the posterior distribution. For a sample $\left\{\mathbf{W}_d^{ya}, \{z^1\}_i\right\}_{d=1}^D$, where traits may switch signs and labels, we are interested in constructing $\left\{\widetilde{\mathbf{W}}_d^{ya}, \{\tilde{z}_{id}^1\}_i\right\}_{d=1}^D$ with "consistent labels and signs", such that:

$$\mathbf{W}_d^{ya} \cdot z_{id}^1 = \widetilde{\mathbf{W}}_d^{ya} \cdot \tilde{z}_{id}^1 \qquad \forall i, d$$

Intuitively, we are interested in finding the major traits that explain heterogeneity.

In order to build this sample, we use two steps:

1. **Fix labels**:

   We obtain the singular value decomposition (SVD) of $\mathbf{W}_d^{ya} = \mathbf{U}_d \cdot \mathbf{D}_d \cdot \mathbf{V}_d'$, where $\mathbf{U}_d$ is an orthogonal matrix of size, $(D_y + P) \times N_1$, $\mathbf{D}_d$ is a diagonal matrix of size $N_1 \times N_1$ with non-negative diagonal values sorted in decreasing order, and $\mathbf{U}_d$ is a orthogonal matrix of size $N_1 \times N_1$. We define $\widehat{\mathbf{W}}_d^{ya} = \mathbf{U}_d \cdot \mathbf{D}_d$, and $\hat{z}_{id}^1 = \mathbf{V}_d' \cdot z_{id}^1$. Note that we have $\mathbf{W}_d^{ya} \cdot z_{id}^1 = \mathbf{U}_d \cdot \mathbf{D}_d \cdot \mathbf{V}_d' \cdot z_{id}^1 = \widehat{\mathbf{W}}_d^{ya} \cdot \hat{z}_{id}^1$.

   This construction allow us to choose the labels of the traits that explain the most variance in decreasing order, similarly as in Bayesian PCA (Bishop 2006), which are unlikely to switch across posterior samples for well behaved samples of the product $\mathbf{W}_d^{ya} \cdot z_{id}^1$, which is identified in our model. However, the sign of the traits are not uniquely determined by the SVD. Note

that if we multiply by -1 a column of $\mathbf{U}_d$, and we also multiply by -1 the same corresponding row of $\mathbf{V}'_d$, then we would also obtain a valid SVD.[5]

2. **Fix signs**:

   We are interested in fixing a sign for each traits across draws of the posterior distribution, however some trait weights may change sign across the posterior. In order words, the posterior distribution may have its mode close to the origin, and therefore the weights may take values both positive and negative. Therefore, we choose the sign of each trait by observing the behavior it impacts the most (demand or acquisition), and we choose the sign such that the weight of this trait on that behavior does not change sign across draws of the posterior sample.

   More formally, let $k = 1, \ldots, N_1$ a trait (a column of $\mathbf{W}_d^{ya}$), and $n(k)$ the behavior (a row of $\mathbf{W}_d^{ya}$) that is most impacted by trait k, which we operationalize by computing the posterior mean of the absolute value of $\widehat{w}_{nk}^{ya}$, the weight of trait k on behavior n (i.e., the nk'th component of matrix $\widehat{\mathbf{W}}^{ya}$), and choosing the maximum:

$$n(k) = \underset{n=1,\ldots,(D_y+P)}{\arg\max} \left\{ \frac{1}{D} \sum_{d=1}^{D} \text{abs}\left(\widehat{w}_{nk,d}^{ya}\right) \right\} \tag{E.29}$$

   Then, we change the sign of the trait so $\mathbf{w}_{n(k)k,d}^{ya}$ is always positive, by defining $\widetilde{I}_d$ a diagonal matrix of size $N_1 \times N_1$, where its k diagonal value is:

$$(\widetilde{I}_d)_{kk} = \text{sign}\left(\widehat{w}_{n(k)k,d}^{ya}\right)$$

   Finally, we construct our sample by:

$$\begin{aligned} \widetilde{\mathbf{W}}_d^{ya} &= \widehat{\mathbf{W}}_d^{ya} \cdot \widetilde{I}_d && \forall d \\ \tilde{z}_{id}^1 &= \widetilde{I}_d \cdot \hat{z}_{id}^1 && \forall i, d \end{aligned}$$

[5] Let $\widetilde{I}$ a diagonal matrix of size $N_1 \times N_1$ where each of its diagonal values are either 1 or -1, then we have that $\left(\mathbf{U}_d \cdot \widetilde{I}\right) \cdot \mathbf{D}_d \cdot \left(\mathbf{V}_d \cdot \widetilde{I}\right)' = \mathbf{U}_d \cdot \widetilde{I} \cdot \mathbf{D}_d \cdot \widetilde{I} \cdot \mathbf{V}'_d = \mathbf{U}_d \cdot \mathbf{D}_d \cdot \mathbf{V}'_d$.

F Algorithm for newly-acquired customers

With reference to (10), once we have estimated the full model using the calibration data, we can form first impressions of newly acquired customers using the following procedure:

Algorithm 1 Forming first impressions

Input A sample of the population parameters drawn from the posterior $\{\Theta_m\}_{m=1}^M$
Acquisition characteristics A_j of focal customer j.
Output A sample of $\boldsymbol{\beta}_j^y$ drawn from $p(\boldsymbol{\beta}_j^y|A_j, \mathcal{D})$
for all $d \leftarrow 1 : S$ **do**
Draw $\Theta_d \sim p(\Theta|\mathcal{D})$ from sample $\{\Theta_m\}_{m=1}^M$
Draw $\mathbf{Z}_{jd} \sim p(\mathbf{Z}_j|\Theta_d, A_j)$ ▷ Using MCMC, HMC or VI
Compute $\boldsymbol{\beta}_{jd}^y \leftarrow \boldsymbol{\mu}_d^y + \mathbf{W}_d^y \cdot \mathbf{z}_{jd}^1$
end for
Return $\left\{\boldsymbol{\beta}_{jd}^y\right\}_{d=1}^S$

Note that the step "Draw $\mathbf{Z}_{jd} \sim p(\mathbf{Z}_j|\Theta_d, A_j)$" involves sampling from a posterior distribution for which we do not have access to a closed form distribution. Instead, using the approximation described in (10), we use HMC to approximately sample from this posterior for each draw $\Theta_d \sim p(\Theta|\mathcal{D})$. Note that as in this sub-model, only $\mathbf{Z}_j$ of the focal customer j is unknown, an HMC algorithm that samples from this posterior is computationally fast even if this algorithm has to be run inside the loop for each value of d.

G Empirical application: Additional results

G.1 Possible sources of endogeneity in the model components

Like most demand models including firm's marketing actions, we face the risk of introducing endogenous variables in our model, potentially preventing us from obtaining unbiased estimates of the customers' parameters. If that were the case, the relationships between acquisition characteristics and demand parameters captured by the model would likely reflect the firms strategies, and not the true underlying interrelations that the FIM intends to capture.

Given the intended applications for this modeling framework, there are three mechanisms by which endogeneity concerns would arise: (unobserved) *temporal shifts* that systematically affect both the time-varying covariate and the overall demand, *static targeting rules*, whereby some customer characteristics (unobserved to the researcher) makes a customer more/less prone to receive marketing actions, while such a characteristic is also correlated to other components of the model, and *dynamic targeting rules*, whereby the presence/absence of the marketing action is driven by an unobserved customer state, which is also correlated with the individual propensity to transact with the firm. The former case is likely to be present if, for example, the firm introduced products or ran specific campaigns only when periods of lower/higher level of demands were expected. The second case corresponds to situations in which marketing actions such as e-mails are prioritized to customers of certain characteristics, for example, those who usually transact online, which is likely to be correlated with one of the acquisition characteristics. The third case is that in which the firm targets only customers who exhibit a behavior that is correlated with demand, for example, send an email to customers who have visited the online store in the last week, or those who abandoned a basket before purchase, etc.

First, we explore the extent to which these phenomena might present in our application. According to the managers of the focal firm, marketing actions are decided in two steps. First, the firm chooses periods in which it will engage in promotional activity (i.e., run a marketing campaign). This decision is made from the headquarters, runs several times through the year (with special campaigns run during the holidays) and affects all markets simultaneously. Second, managers in each focal market choose the set of customers who will receive each campaign, with the

proportion of customers not being determined consistently. The only variable that some markets include in their targeting rules is recency (i.e., time since last purchase). The introduction of new products follows a similar process — i.e., the decision being made globally, the implementation affected also by local factors such as distribution shocks in each of the markets — with the main difference being that the second step does not vary across customers of the same market.

Therefore, regarding potential (omitted) temporal shifts, the only variable that could systematically affect the presence/absence of promotional activity in all markets is the holiday season, which is not omitted as it is included in the model. Regarding (static) targeting rules, we confirm with the firm and verify with the data that these were not present in our application. Nonetheless, it is worth noting that when such targeting rules are present (e.g., the firm contacts customers based on demographic information), because the model includes unobserved heterogeneity on purchase frequency (first element of $\boldsymbol{\beta}_i^y$), the identification of the individual-level sensitivity to promotional activity comes mainly from individual differences across periods, for which we have rich variation during the four years of available data. Finally, regarding dynamic targeting rules, it is indeed the case that some customers (in the most sophisticated markets) are more/less prone to receive emails and DM based on their purchase activity. However, our model not only includes unobserved heterogeneity on purchase frequency — capturing the customers' base level of activity — but also includes the recency of purchase, alleviating the endogeneity concerns arising from potential correlation between the firm's targeting policies and customers' propensity to transact in a particular period.

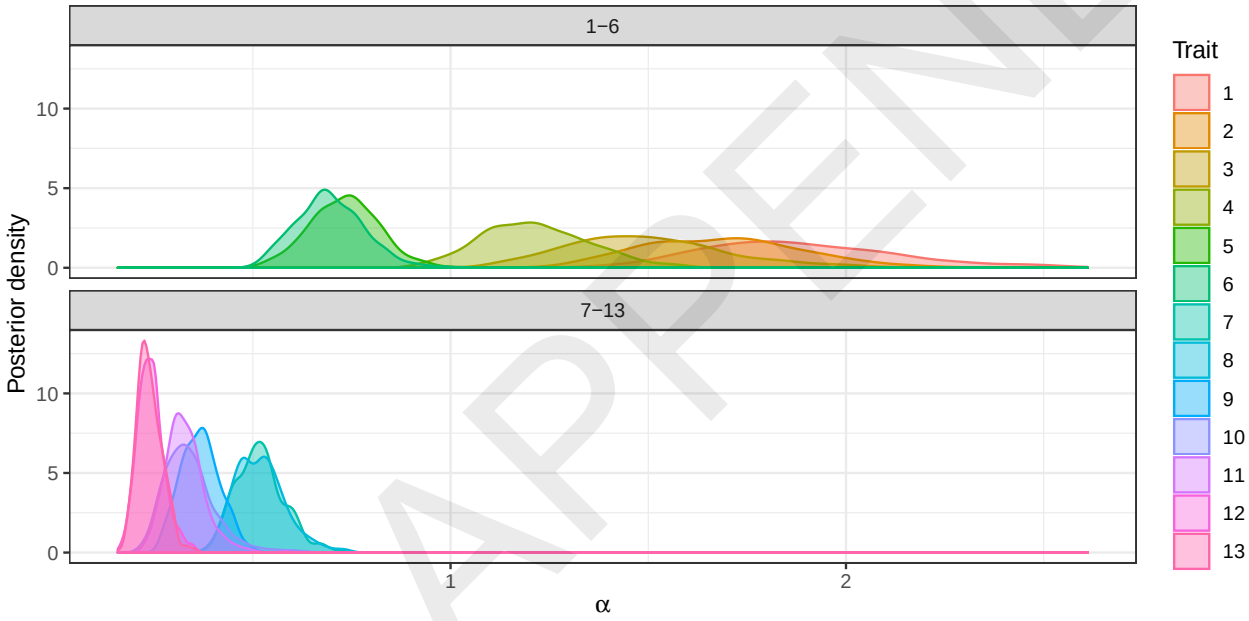
To conclude, given the business nature of our application, the rich variation in our data (Section 5.1.2, Marketing actions), and our model specification, we argue that the potential endogenous nature of the marketing actions is not a main concern in this research. Nevertheless, in situations where these conditions do not hold (due to different strategic behavior by the firm or for data limitations), the demand model should be adjusted to account for the firm's targeting decisions. Given the flexibility of our modeling framework, those adjustments would merely involve extending the demand model to capture unobserved shocks between firm's actions and individual-level responsiveness (Manchanda et al. 2004) or adding correlations between firm decisions and unobserved demand shocks through copulas (Park and Gupta 2012), depending on how these actions

are determined by the firm. Those changes would only affect the demand (sub)model and not the overall specification of the FIM.

G.2 Exploring the latent factors

Figure G.11 shows the posterior distribution of weight variances $\boldsymbol{\alpha}$ for each one of the 13 traits. As described in Appendices C.1 and D.7, each trait parameter α_k controls whether traits are activated by regularizing the weights ($\mathbf{W}^y$ and $\mathbf{W}^a$) related to the k'th trait.

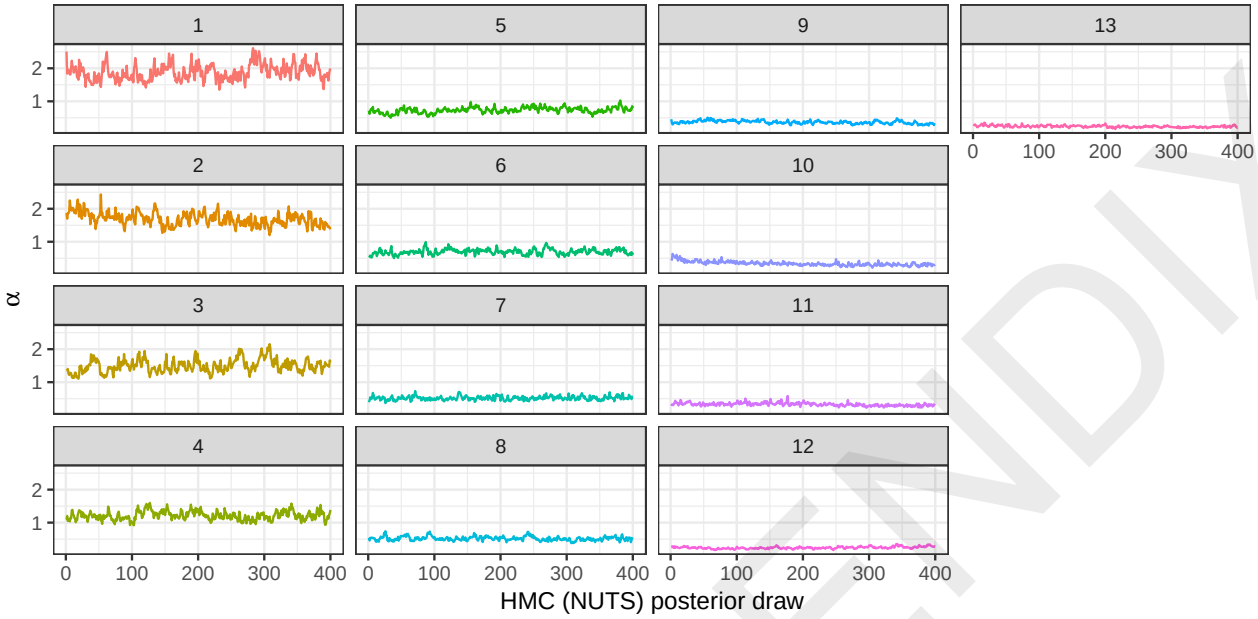
Figure G.11: Posterior distribution of $\boldsymbol{\alpha}$.



We conclude that the first 6 traits carry most of the weight at "connecting" acquisition and demand variables. (Note that the convergence of these parameters, in Figure G.12, shows no evidence of label switching or rotation of these traits.) This is not to say that the other traits irrelevant. In turn, those other traits add to the prediction accuracy of the model. However, for deriving insights from the model parameters, we choose to explore the handful of traits that carry most of the information.

Following the discussion in Appendix D.7, we plot in Figure G.13a the posterior density of the computed pseudo-α for each upper trait for the FIM model used in our empirical application ($N_1 = 13$, $N_2 = 5$). We find that the relevance of the fifth upper traits is significantly lower than the relevance of the first three traits. This result suggests that $N_2 = 5$ is enough to capture the

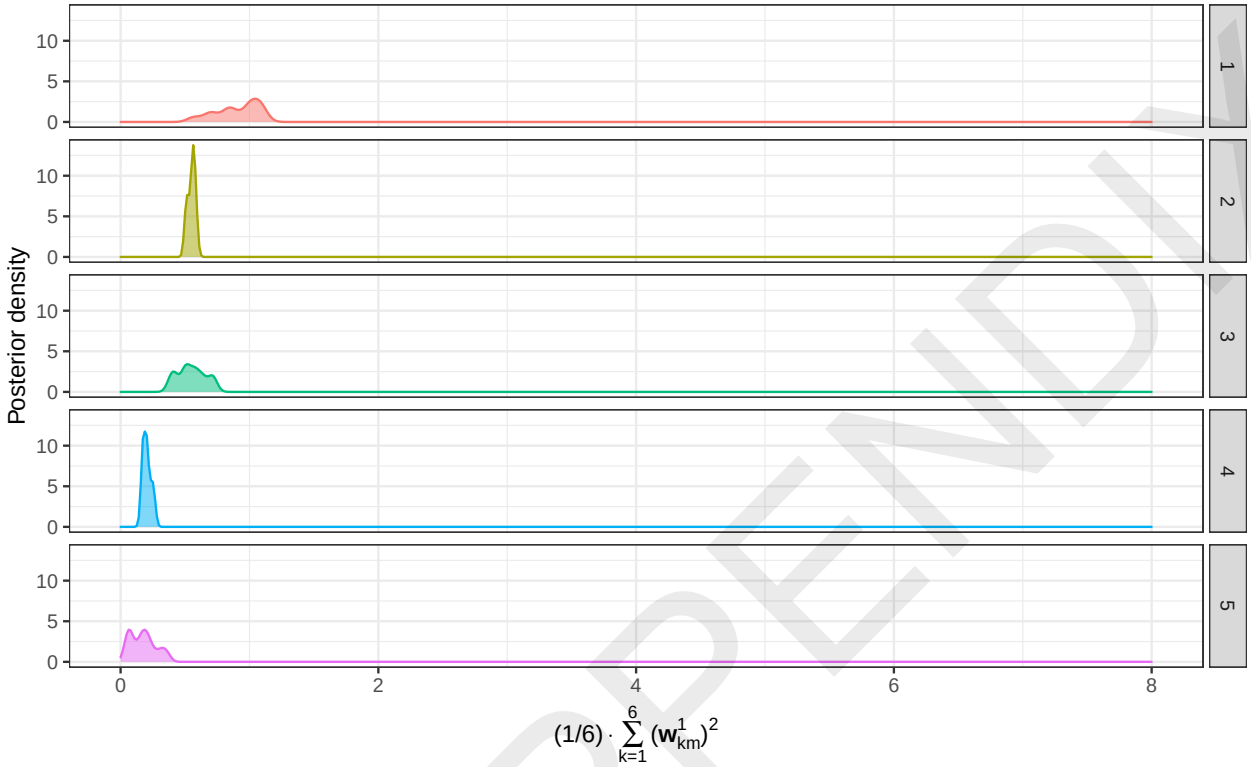
Figure G.12: Convergence of $\boldsymbol{\alpha}$.



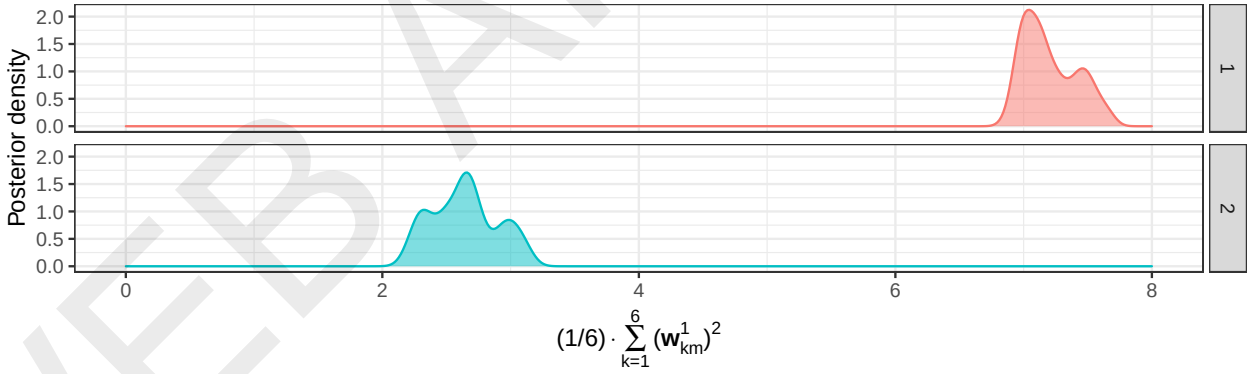
non-linear interrelations present in the data. For robustness, we estimate another FIM specification with $N_2 = 2$ instead, and we find that all upper traits are relevant, suggesting that $N_2 = 2$ may not be enough to capture the non-linear relationships present in the data.

Figure G.13: Posterior distribution of pseudo-$\boldsymbol{\alpha}^1$.

(a) FIM ($N_1 = 13$, $N_2 = 5$).



(b) FIM ($N_1 = 13$, $N_2 = 2$).



G.3 Latent attrition benchmarks models

We estimate three additional non-nested benchmark models (borrowed from the CRM literature) that do account for latent attrition: (1) Linear model with marketing actions + logistic attrition process (without acquisition covariates), (2) Linear model (without marketing actions) + logis-

tic attrition with acquisition covariates, and (3) Linear model with marketing actions + logistic attrition with acquisition covariates.

For all the aforementioned models we define purchase incidence (y_{it}) given attrition, which we denote as h_{it}, and we have that $p(y_{it} = 1|h_{it} = 1) = 0$, $p(h_{it} = 0|h_{it-1} = 1) = 0$, and

$$p(h_{it} = 1|h_{it-1} = 0) = \text{logit}^{-1}\left[\beta_i^h\right],$$

where β_i^h is a (scalar) parameter that captures the individual log-odds of attrition. In all specifications, we model the purchase incidence parameters $\boldsymbol{\beta}_i^y$ as a linear function of acquisition characteristics as described in Appendix D.3.2.

The models differ in the inclusion of marketing actions into the demand given attrition component and modeling of the attrition parameter β_i^h as displayed in Table G.10.

Table G.10: Latent attrition benchmarks models.

	Demand $p(y_{it} = 1\|h_{it} = 0)$	Attrition parameter β_i^h
Latent Attrition		
w/ Acq.	$\text{logit}^{-1}\left[\beta_{i1}^y + \alpha_m\right]$	$\beta_i^h = \mu^h + \Gamma^h \cdot A_i + \Delta^h \cdot \mathbf{x}_{m(i)}^a + u_i^h$
w/ Mktg. Actions	$\text{logit}^{-1}\left[\mathbf{x}_{it}^{y\prime} \cdot \boldsymbol{\beta}_i^y + \alpha_m\right]$	$\beta_i^h = \mu^h + u_i^h$
w/ Acq.+Mktg. Actions	$\text{logit}^{-1}\left[\mathbf{x}_{it}^{y\prime} \cdot \boldsymbol{\beta}_i^y + \alpha_m\right]$	$\beta_i^h = \mu^h + \Gamma^h \cdot A_i + \Delta^h \cdot \mathbf{x}_{m(i)}^a + u_i^h$

Note that in all specifications we model jointly the unobserved individual components of purchase incidence and attrition parameters by $[\mathbf{u}_i^y, u_i^h] \sim \mathcal{N}(0, \Sigma^{yh})$.

G.4 Details on the (Machine Learning) benchmark models

We estimate the Feed-Forward DNN model (hidden layer with ReLu as activation function, sigmoid output and cross-entropy loss) using package `torch` in R. We select the value of the weight decay based on the loss calculated using hold-out data in the training sample. After evaluating the values$=0.01, 0.005, 0.001, 0.0005, 0.0001$, the value that provides better performance is 0.0001, which we use to estimate the model on the full training sample using 10 epochs. We set the number of hidden dimensions to 128 after corroborating that larger dimensionality does not lead to better fit of the model.

We estimate the Random Forest (RF) using the package `ranger` in R. We finetune the number of trees (num.trees), number of variables to possibly split at in each node (mtry), and fraction to

sample (sample.fraction) via cross-validation using the training sample. The resulting values, which we use to estimate the model in the full training data are, num.trees= 1000, sample.fraction = 0.3, mtry = 6.

G.5 Interpreting the latent traits

Finally, we further explore the posterior estimates of the (lower layer) hidden traits and their relationship with the demand and acquisition parameters to provide additional insights into customer traits and behaviors. We begin by analyzing which latent traits capture the most salient relationships in the data. We do so by exploring the posterior estimates of the parameters governing the ARD component of the model and find that six traits carry most of the "weight" at connecting acquisition and demand parameters. (Please see Appendix G.2 for details.) Then, we investigate the correlations among these traits (Table G.11), exploring whether customers that score high in a particular trait also score high (or low) in another trait. Note that these traits *do not capture segments* in the population (e.g., groups of customers of similar characteristics) but rather traits that capture the multiple dimensions of customer behavior. In other words, every customer has a score for each of the traits, being not only possible but very likely that customers score high in more than one trait. In our data, customers who score high in Trait 4 also tend to score high in Trait 6 (correlation= 0.553). On the contrary, those same customers have the tendency to score low in Trait 5 (correlation= -0.268).

Table G.11: Posterior mean of correlations across customers of individual lower level traits $\mathbf{z}_i^1$.

	Trait 1	**Trait 2**	**Trait 3**	**Trait 4**	**Trait 5**
Trait 1	1.000				
Trait 2	-0.144				
Trait 3	0.101	-0.113			
Trait 4	0.130	0.185	0.170		
Trait 5	-0.026	-0.141	-0.057	-0.268	
Trait 6	0.129	0.242	0.258	0.553	-0.361

An obvious question to ask is: What do these traits represent? To answer that question we compute the posterior mean of the weights of each of the rotated trait on each of the acquisition and demand parameters (Table G.12). Looking at the weights to the demand parameters, we learn that the first trait is the most relevant in explaining heterogeneity in the base propensity to buy. Scoring high on this "high-frequency" trait also relates to a positive response to product

introductions in future demand. This first trait is negatively correlated with whether the first purchase was made online and whether that purchase contained a product in the Home category; but positively correlated with whether the customer purchased a product in the Hair Care category. Interestingly this trait is also positively correlated with first transaction baskets containing products that score high on dimension 4 of the Basket Nature product embeddings. Moreover, customers that score high on this trait are more likely to buy at their first purchase smaller sized products and travel sized products.

Table G.12: Rotated traits weights' on acquisition and demand variables

Parameter	Trait					
	1	**2**	**3**	**4**	**5**	**6**
Demand ($\mathbf{W}^y$)						
Intercept	0.133	0.129	-0.106	-0.072	-0.002	0.024
Email	-0.018	-0.016	0.046	0.027	-0.015	-0.004
DM	0.010	0.038	-0.003	-0.001	0.013	-0.004
Product introductions	0.044	0.085	0.001	-0.029	-0.026	0.009
Season	-0.025	0.058	0.027	0.085	0.004	0.005
Acquisition ($\mathbf{W}^a$)						
Avg. price (log)	-0.109	0.022	-0.644	-0.370	0.039	0.313
Amount (log)	-0.021	0.076	-0.541	0.305	0.209	0.425
Quantity (log-log)	0.074	0.066	0.050	0.647	0.174	0.130
Package size (log)	-0.143	0.052	-0.087	-0.205	0.016	0.217
Holiday	0.029	-0.110	0.053	0.159	0.085	0.170
Discount	0.298	-0.073	0.280	0.414	0.133	0.029
Online	-0.382	1.368	0.581	6.830	0.019	0.146
New product	0.007	0.216	-0.283	0.544	0.354	0.234
Travel	0.470	-0.928	0.440	0.724	0.413	0.037
Category: Body Care	0.248	-4.922	-0.112	2.916	-0.072	-0.016
Category: Body Perfume	-0.025	0.436	-1.152	0.554	0.462	0.079
Category: Face Care	0.352	0.610	0.051	0.745	0.234	0.718
Category: Hair Care	1.267	1.178	-0.514	1.930	-0.631	-0.595
Category: Home	-1.097	-0.051	-0.336	1.836	1.073	-0.417
Category: Kits	0.285	0.227	-0.469	0.803	-0.100	0.225
Category: Make Up	0.377	0.528	0.334	1.149	-0.137	0.001
Category: Others	-0.134	0.230	0.623	1.845	0.387	0.029
Category: Services	-0.006	0.110	-0.501	5.762	-0.545	0.102
Category: Toiletries	0.239	0.733	0.200	1.190	0.607	-0.268
BasketNature dimension 1	-0.104	-0.022	-0.071	0.083	0.078	-0.112
BasketNature dimension 2	0.042	0.012	-0.011	-0.003	0.110	-0.035
BasketNature dimension 3	0.193	0.082	0.034	-0.040	-0.180	0.153
BasketNature dimension 4	0.200	0.105	-0.021	0.136	-0.167	0.005
BasketNature dimension 5	-0.035	0.003	0.001	0.025	0.009	0.154
BasketNature dimension 6	0.120	-0.017	0.141	-0.102	0.012	0.010
BasketDispersion dimension 1	-0.150	0.012	-0.166	0.256	0.237	-0.238
BasketDispersion dimension 2	-0.033	0.026	-0.105	0.196	0.114	-0.151
BasketDispersion dimension 3	-0.045	-0.094	-0.155	0.379	0.039	-0.120
BasketDispersion dimension 4	0.113	0.086	-0.216	0.406	-0.087	-0.082
BasketDispersion dimension 5	-0.137	0.123	-0.154	0.360	0.155	-0.195
BasketDispersion dimension 6	-0.033	-0.020	-0.159	0.462	0.078	-0.160

Another interesting trait is number four, which is associated with lower propensities to buy (intercept) and higher activity during the holiday season (Season variable). This "holiday-customer" trait is positively correlated with whether customers have been acquired online and during the Holiday season. This trait is positively associated with less expensive products and more units on the first transaction. With respect to the type of products associated with the first purchase, customers that score high on this trait are more likely to buy in the Body Care, Hair Care and Home categories. (Note that this trait is capturing some of the associations among acquisition variables reported in Table 3 — e.g., [Online-FaceCare]= 0.48 — allowing the model to clean redundancies in the acquisition characteristics and tie the main trait to demand variables.) Finally, this "holiday-customer" trait is related with very diverse baskets (with respect to the type of products purchased in the first transaction), as indicated by its positive weights on Basket dispersion in all six dimensions.

G.6 FIM predictive accuracy using in-sample customers

Table G.13 shows the performance of all models on the *Training* sample. The first two columns show the in-sample fit for each of the models, for which we compute log-likelihood and Watanabe-Akaike Information Criterion (WAIC) (Watanabe 2010). Columns 3 through 6 show different measures of out-of-sample prediction accuracy, computed for customers in the training sample, but using the time periods that were not included in the estimation (i.e., periods after April 2014). We compute log-likelihood as well as the root mean square error (RMSE) for behavioral predictions. In particular, we compare the predicted and actual number of transactions at the observation level (i.e., at the customer/period level), at the customer level, calculating the total number of transactions per customer (in "future" periods), and at the period level, computing the total number of transactions per period. While the HB benchmark model fit the in-sample data better than our proposed model, the FIM outperforms all benchmarks in the out-of-sample predictions. In other words, whereas the hierarchical models are very flexible at capturing heterogeneity in the training data, such a model is likely overfitting the data, as reflected in the out-of-sample predictions. On the other hand, the FIM forecasts the out-of-sample behavior of existing customers with greater accuracy.

Table G.13: Model fit and prediction accuracy for the *Training* sample

	In-sample		Out-of-sample (future periods)			
	Log-Like	WAIC	Log-Like	RMSE		
Model				Observation	Customer	Period
HB - Linear	**−7843.0**	**17807.8**	−5511.1	0.202	0.723	62.841
Latent Attrition w/ Acq	−7880.1	17507.7	−6126.5	0.201	0.750	78.810
Latent Attrition w/ Mktg. Actions	−7781.1	17715.5	−5786.0	0.206	0.767	74.525
Latent Attrition w/ Acq+Mktg. Actions	−7612.8	17438.2	−6476.8	0.209	0.812	81.143
Bayesian PPCA	−8482.4	18361.4	−5137.2	0.191	0.573	35.696
Feed-Forward DNN	−−	−−	−−	**0.189**	0.556	53.410
Random Forest	−−	−−	−−	0.193	0.616	133.598
FIM ($N_1 = 13, N_2 = 5$)	−9135.4	18885.7	**−5096.4**	0.190	**0.533**	**32.313**
Other FIM specifications						
FIM ($N_1 = 12, N_2 = 2$)	−8654.0	18555.7	−5097.2	0.191	0.558	32.612
FIM ($N_1 = 12, N_2 = 5$)	−8952.1	18927.6	−5116.7	0.190	0.541	32.762
FIM ($N_1 = 13, N_2 = 2$)	−8587.6	18399.0	−5140.1	0.192	0.578	35.454
FIM ($N_1 = 14, N_2 = 2$)	−8683.6	18531.9	−5131.8	0.191	0.561	33.824
FIM ($N_1 = 14, N_2 = 5$)	−8613.9	18465.3	−5147.6	0.191	0.571	34.423

G.7 Population distribution and individual-level posterior distributions

Figure G.14 summarizes the inferred individual posterior distributions of the demand parameters of *Test* customers using their acquisition characteristics. The top row of Figure G.14 shows the degree of heterogeneity that the FIM infers. How uncertain are those inferences at the individual level? In order to answer that question, for each demand parameter, we sort customers based on their posterior means, and compute their 95% CPI. The second row of Figure G.14 shows the uncertainty at the individual level that the model can infer these parameters: each customer is represented horizontally, where the shaded area shows their 95% CPI and the white line, their posterior mean. Using this figure we can show that for the case of the intercept of the demand model, can clearly separate some customers based on their acquisition characteristics: the bottom customers in the figure (i.e., those with individual posterior means between -2.5 and -2) have clearly higher intercept than the top customers (i.e., those with individual posterior means around -4) as the 95% CPI of the latter group does not overlap with the posterior means of the former.

Figure G.14: Population distribution and individual-level posterior distribution for customers in the *Test* sample. The top row shows an histogram of individual-level posterior means for each demand parameter. The bottom row shows customers sorted by posterior means, where the shaded area and the white line represent the individual-level 95% CPI and posterior mean, respectively.