



Overcoming the Cold Start Problem of Customer Relationship Management Using a Probabilistic Machine Learning Approach

Nicolas Padilla and Eva Ascarza

Abstract

The success of customer relationship management programs ultimately depends on the firm's ability to identify and leverage differences across customers—a difficult task when firms attempt to manage new customers, for whom only the first purchase has been observed. The lack of repeated observations for these customers poses a structural challenge for firms to infer unobserved differences across them. This is what the authors call the “cold start” problem of customer relationship management, whereby companies have difficulties leveraging existing data when they attempt to make inferences about customers at the beginning of their relationship. The authors 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. The authors validate their approach in a retail context and empirically demonstrate the model's ability to identify high-value customers as well as those most sensitive to marketing actions right after their first purchase.

Keywords

customer relationship management, deep exponential families, probabilistic machine learning, cold start problem

Online supplement: <https://doi.org/10.1177/00222437211032938>

Customers differ 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 estimate unobserved differences across customers using their past history. These methods allow firms to identify customers with higher versus lower expected lifetime value (e.g., Fader, Hardie, and Lee 2005; Fader, Hardie, and Shang 2010; Schmittlein, Morrison, and Colombo 1987), those who are less sensitive to a price increase (e.g., Allenby and Rossi 1998; Rossi, McCulloch, and Allenby 1996), 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, when data are limited.

Firms have traditionally relied on demographics (e.g., age, gender) and/or recency metrics (e.g., how many weeks since the last transaction) to target marketing efforts with limited data (Shaffer and Zhang 1995). These approaches, however, face practical limitations: recency metrics, for example, do

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.

not differentiate among recently acquired customers (as they all were acquired at the same time), and relevant personal information is generally difficult to collect or poses data privacy challenges. However, 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., Did the transaction occur online or offline?, Did the customer buy any new products or old bestsellers in the transaction?, Did they buy any products on discount?). While some of these characteristics may merely coincide with the moment in which the customer made their first purchase, others may carry important information, as they reflect latent customer preferences/attitudes. Thus, although firms observe a newly acquired customer on only one occasion, they now have many more cues to form a “first impression” of this customer, 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 (ML) 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 we 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 by the firm at the moment of acquisition) where the individual-level parameters of each of these submodels 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 (ARD) priors for this layer, enforcing sparsity and automatically reducing the dimensionality of the individual-level parameters, similarly as in a Bayesian principal component analysis (PCA) model and modern applications of “supervised” factor models. The model departs from the aforementioned models by allowing nonlinear 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 can 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. Because the model will recover them from the data, those (linear or nonlinear) relationships do not need to be prespecified. 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 ARD. Third, the model is scalable, being applicable to data sets 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 (i.e., “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 noncontractual, can use this framework by making minor adjustments to the demand/transactional model.

Using a set of simulation analyses, we demonstrate that the FIM inferences for newly acquired customers are 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 its 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 (e.g., price paid, number of products purchased) as well as observed product characteristics (e.g., category purchased, package size). 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—an ML 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 on the basis of 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 heavy spenders from those expected to yield less value, immediately after they make their first purchase. 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 are likely to include fewer units. We also find nonlinear relationships between acquisition characteristics and customer responsiveness to marketing actions. For example, the differences in email sensitivities across customers who received discounts on their first purchase exist only 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 newly acquired customers. Methodologically, our article 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, which is 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 ML and natural language processing to create a matrix of product “embeddings” that enable firms to characterize (even recently acquired) customers on the basis of 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 article 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 relative to 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.

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, Park, and Jamal (2014), who, building on the foundations of the Beta-Geometric/Beta-Binomial model (Fader, Hardie, and Shang 2010), incorporate the effect of direct marketing activity and past customer activity on the latent attrition process and the customer's purchase propensity. In addition, Knox and Van Oest (2014) and Braun, Schweidel, and Stein (2015) incorporate the effect of the customer service experience and customer complaints on the latent attrition process of the Beta-Geometric/Negative Binomial Distribution model (Fader, Hardie, and Lee 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. Although the referenced models could be used to incorporate a handful of prespecified acquisition variables, they are not well-suited to extract relevant information from noisy and redundant variables, as is 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 previous research.

On a substantive level, our work relates to Gopalakrishnan, Bradlow, and Fader (2016), who propose a framework for multicohort data able to predict the behavior of new cohorts of customers for whom little transactional data are available. These authors 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. (2021), 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, thus, provide less value to the firm. Our work also relates to Loupos, Nathan, and Cerf (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 it is challenging to secure a second purchase.

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, Hardie, and Jerath 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 CLV (e.g., Chan, Wu, and Xie 2011; Datta, Foubert, and Van Heerde 2015; Lewis 2006; Schmitt, Skiera, and Van den Bulte 2011; Steffes, Murthi, and Rao 2011; Uncles, East, and Lomax 2013; Verhoef and Donkers 2005; Villanueva, Yoo, and Hanssens 2008). Although our work investigates relationships between acquisition-related variables and subsequent customer behavior, it 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 causal 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 examining not only 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?), thus extracting more information from the initial transaction. The latter is especially relevant for managers and analysts in large retail and hospitality businesses, among others, given that such information is not only easily observed but also typically already in their databases.

From a methodological perspective, we contribute to the literature on applying probabilistic ML methods to marketing (Dew and Ansari 2018; Dew, Ansari, and Li 2020; Jacobs, Donkers, and Fok 2016). More specifically, our work relates to the literature on applying DEFs (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 PCA (Bishop 1999; Mohamed, Ghahramani, and Heller 2008).

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 (Marr 2015). Second, retail represents a large proportion of the total economy, with revenues accounting for 31% for the global gross domestic product (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 (e.g., hospitality, entertainment business, nonprofit organizations) that face similar data challenges when implementing CRM programs.

The Cold Start Problem

Consider a retailer that sells cosmetic/beauty products via both 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 in (e.g., 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, Hardie, and Shang 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, Park, and Jamal 2014). However, when making these types of inferences for recently acquired customers, for whom the firm has no transactional history or 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 to differentiate them, therefore diminishing the effectiveness of future CRM activities.

The premise of this research is that, while it is the lack of (historical) data that causes the cold start problem, firms now have access to other data sources that, properly leveraged, can help them overcome the cold start problem. Granted, if firms observed only that the customer made a transaction, it would be very difficult to overcome this problem. However, most firms not only know when a customer made their first transaction but also record details such as the channel/store used, the exact product the customer purchased, the price paid, whether they used a discount, the time of the day, and so forth.³ We propose leveraging those existing data and extract what we call “acquisition characteristics” from each customer's first transaction.⁴ We contend that these acquisition characteristics/

² This is the specific context of our empirical application. The full set of details about the focal firm and the data are presented in the “Empirical Application” section; in this section, we present only the relevant information to motivate the business problem and the modeling challenges.

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 visited the website previously (without 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 use only the first transaction because that is the data that all customers (i.e., just acquired and existing) have in common, which is the key to making inferences about recently acquired customers. Adding information about each subsequent transaction might add precision to the individual-level inferences of repeat users but

choices can be informative about underlying customer differences that can predict customers' future behavior. 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 customer behavior down the road.

Augmenting Cold Start Data with Acquisition Characteristics

Consider our focal retailer, which 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, Wu, and Xie 2011). The caveat of doing so is that so 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 capture only a coarse representation of customers' heterogeneity. We aim to fully augment the acquisition data to broaden the amount of information that would (potentially) be linked to future behavior, therefore increasing marketers' ability to solve the cold start problem.

Specifically, using the (existing) data from each first transaction, we propose augmenting cold start data with three types of acquisition variables: (1) transaction characteristics (e.g., channel, price paid, holiday season) and (2) product characteristics⁵ (e.g., product category, package size), which are easily extracted from the transactional database, and (3) shopping basket (latent) representation. The latter type of data aims to capture the "nature" of products that the customer purchased, 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 cosmetics market, certain ingredients or aromas characterize lines of products. It is possible that customers who discover the brand by buying products of a certain nature are similar in the 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 and then aggregate it at the basket level.⁶

not necessarily to the inferences of recently acquired customers, which is our main focus.

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 in the first transaction.

⁶ One alternative to this solution would be to include a dummy variable per (available) stockkeeping unit (SKU). This approach would be straightforward in business contexts where the product space is small. However, when the firm offers a large selection of items or SKUs—as 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 from learning across customers. For those cases, we recommend using a lower-dimensional vector representing the product space, as we do in this research.

Previous literature has used different methods to encode such information, from human coding based on full description of the product to ML 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., Chen et al. 2020; Jacobs, Donkers, and Fok 2016; Kumar, Eckles, and Aral 2020; Ruiz, Athey, and Blei 2017). 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), an ML approach widely used for natural language processing, to map each item to a multidimensional 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. (For details about how we process the transaction data and create the product embeddings using the word2vec algorithm, see Web Appendix A.) Once we represent each product by a continuous vector, we can easily characterize the first purchase of any customer by computing moments of the product vectors in that basket.

In summary, 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 and the type of products purchased.⁷

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).

Indeed, acquisition characteristics are predictive of customers' future transactions. Consistent with common belief in the industry (e.g., Artun 2014; RJMetrics 2016), customers who were acquired during the holiday season are less valuable to the firm, as we find that they are less likely to purchase in the

⁷ In our empirical application, this vector has 31 dimensions. We present further details in the "Empirical Application" section.

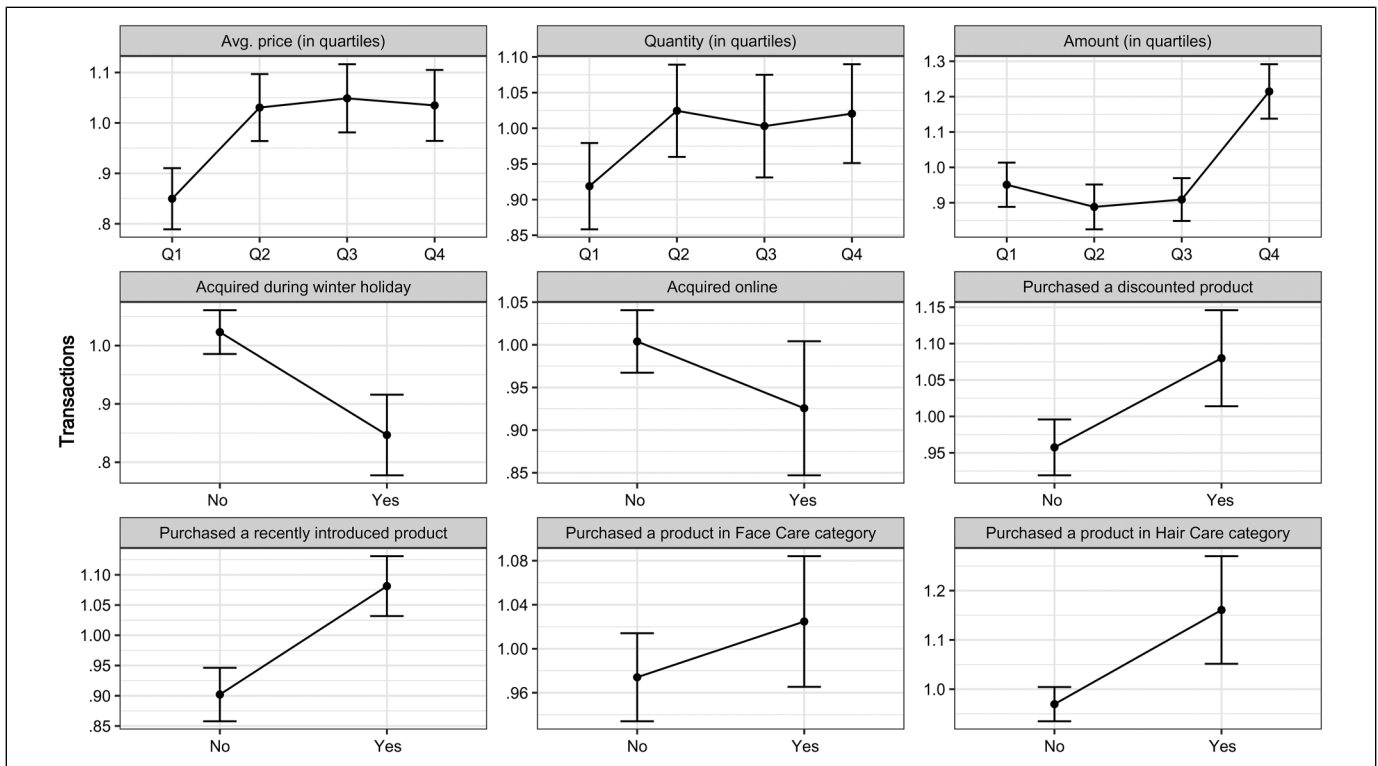


Figure 1. Observed (mean) repeated transactions as a function of a sample of augmented acquisition characteristics.

Notes: 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.

Notes: 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.

future. In contrast, 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 for those who purchased products from the hair care category. Interestingly, this model-free analysis also suggests that some of these relationships are likely to be nonlinear. 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 nonlinear 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 depending on whether they were acquired during the winter holiday season, coupled with whether they purchased travel-size products. We find that purchasing travel-size products moderates the relationship between being acquired during the holidays and the future number of

transactions. In the panel on the right, we observe that purchasing a discounted product in the first transaction signals lower value only if the 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 be performed for only a subsample of customers—those we observe for a relatively long period of time (e.g., 15 months)—to have a fair comparison across customers over the same number of periods. Second, this type of analysis examines each variable independently (Figure 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 light on 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 is 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.

Modeling Challenges

Our solution to overcome the cold start problem ultimately depends on the model's ability 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 to 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 the data would be redundant. For example, a price-conscious customer may purchase a set of travel-sized, cheap products that are discounted. Although the price and discount variables 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 the customer's preference 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 characteristics 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 nonlinear 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 nonlinear 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 because 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. We combine a flexible demand specification (that can be applied to a wide range of marketing contexts) with state-of-the-art ML methods (addressing nonlinearities and data redundancy) within a Bayesian framework (that extracts signals from the acquisition characteristics while handling missing data). The resulting modeling framework is a flexible probabilistic ML model that links the individual-level parameters governing a 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 nonlinear) without the need to prespecify those relationships a priori. Moreover, the model explicitly accounts for interrelations among acquisition data, which helps regularize the flexible model and avoid 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 address some (but not all) modeling challenges. For example, we compare it with a standard hierarchical Bayesian (HB) model with acquisition characteristics included as covariates, a fully hierarchical model where acquisition characteristics and demand are jointly correlated using a multivariate Gaussian distribution, and a (supervised) Bayesian PCA that aims to reduce dimensionality of acquisition characteristics as well as demand parameters.

Finally, 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 ML approaches such as a random forest (RF) and a deep neural network (DNN; proven to capture non-linear relationships very well). Our model stands out from these ML benchmarks in two main ways. Methodologically, it can be easily combined with multiple demand specifications and 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.

Modeling Framework

Model Development

Our modeling framework, the FIM, comprises three main components: (1) the demand model, the main outcome of interest to the firm, which could include customers' transactions, purchase volume, and so on; (2) the acquisition model, which captures all customer outcomes that are observed to the firm at the moment of acquisition; and (3) the probabilistic model, which links the underlying customer parameters influencing these two types of behaviors through hidden traits.

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

$$p(y_{it}|\tilde{\mathbf{x}}_{it}^y, \boldsymbol{\beta}_i^y, \boldsymbol{\sigma}^y) = f^y(y_{it}|\tilde{\mathbf{x}}_{it}^y, \boldsymbol{\beta}_i^y, \boldsymbol{\sigma}^y) \quad (1)$$

$$i \in \{1, \dots, I\}, t \in \{1, \dots, T_i\},$$

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 $\tilde{\mathbf{x}}_{it}^y$ includes the observed covariates for customer i at period t . Finally, $f^y(\cdot)$ is the probability density function/probability mass function (pdf/pmf) for outcome y_{it} ; for example, if the outcome of interest is purchase incidence, we would specify $p(y_{it} = 1) = \text{logit}^{-1}[\mathbf{x}_{it}^{y'} \times \boldsymbol{\beta}_i^y]$.⁸

Acquisition model. We denote \mathbf{A}_i as the vector of characteristics that are collected at the moment of acquisition, and \mathbf{a}_{ik} as 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?) and 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). 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 drives the acquisition process to change the pool of acquired customers.

Modeling the acquisition characteristics as an output not only enables 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 the "Linking Acquisition and Future Demand: Deep Probabilistic Model" subsection.) Specifically, we denote

$$p(\mathbf{a}_{ip}|\boldsymbol{\beta}_{ip}^a, \boldsymbol{\sigma}_p^a, \mathbf{x}_{m(i)\tau(i)}^a) = f_p^a(\mathbf{a}_{ip}|\boldsymbol{\beta}_{ip}^a, \boldsymbol{\sigma}_p^a, \mathbf{x}_{m(i)\tau(i)}^a) \quad (2)$$

$$i \in \{1, \dots, I\}, p \in \{1, \dots, P\},$$

where P is the number of different types of behaviors collected at acquisition, $\boldsymbol{\beta}_{ip}^a$ is an individual-level parameter that reflects tendency to observe such a behavior when customer i is acquired, $\boldsymbol{\sigma}_p^a$ denotes a vector of parameters that are common across customers, and $\mathbf{x}_{m(i)\tau(i)}^a$ 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 $f_p^a(\cdot)$ is the pdf/pmf of a distribution to model acquisition behavior p . Note that some of these behaviors will likely be binary⁹ (e.g., whether the customer was acquired online), in which case we specify $\boldsymbol{\sigma}_p^a = [\mathbf{b}_p^a]$ and model p as

$$p(\mathbf{a}_{ip} = 1) = \text{logit}^{-1}[\boldsymbol{\beta}_{ip}^a + \mathbf{x}_{m(i)\tau(i)}^a \times \mathbf{b}_p^a]. \quad (3)$$

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

$$p(\mathbf{a}_{ip}) = \mathcal{N}[\mathbf{a}_{ip}|\boldsymbol{\beta}_{ip}^a + \mathbf{x}_{m(i)\tau(i)}^a \times \mathbf{b}_p^a, \boldsymbol{\sigma}_p^a], \quad (4)$$

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

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 Equation 1 as a state-space model (e.g., a hidden Markov model) with state variable s_{it} and $p(y_{it}, s_{it}|\mathbf{y}_{1:t-1}, s_{1:t-1}) = p(y_{it}|s_{it}) \times p(s_{it}|s_{it-1})$. 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 Equation 11 for $p(y_{it}, s_{it}|\mathbf{y}_{1:t-1}, s_{1:t-1}, \mathbf{x}_{it}^y, \boldsymbol{\beta}_i^y, \boldsymbol{\sigma}^y) = p(y_{it}|s_{it}, \mathbf{x}_{it}^y, \boldsymbol{\beta}_i^{yq}, \boldsymbol{\sigma}^{yq}) \times p(s_{it}|s_{it-1}, \mathbf{x}_{it}^y, \boldsymbol{\beta}_i^{ye}, \boldsymbol{\sigma}^{ye})$, where $\boldsymbol{\beta}_i^y = [\boldsymbol{\beta}_i^{yq}, \boldsymbol{\beta}_i^{ye}]$ and $\boldsymbol{\sigma}^y = [\boldsymbol{\sigma}^{yq}, \boldsymbol{\sigma}^{ye}]$ are the parameters of the demand model.

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

All of these types of variables are easily incorporated by adjusting the acquisition model accordingly. We define $\beta_i^a = [\beta_{i1}^a \dots \beta_{ip}^a]$ and $\sigma^a = [\sigma_1^a \dots \sigma_p^a]$ as the full set of individual- and population-level vectors of acquisition parameters, respectively.

Note that we have only one observation per individual and behavior. Thus, in theory, having an individual-level parameter β_{ip}^a could completely capture the residual variance of a_{ip} , which 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_{ip}^a 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 will become clearer when we specify the relationship between the individual-level demand and acquisition parameters, β_i^y and β_i^a , as we do in the next section.

Finally, the term $\mathbf{x}_{m(i)\tau(i)}^a$ 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 driven by not only 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.

Linking acquisition and future demand: deep probabilistic model. We use a DEF component (Ranganath et al. 2015) to relate demand and acquisition parameters hierarchically, through hidden layers. We chose this specification because of its hierarchical nature, which allows 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 have the flexibility of probabilistic models, allowing them to be easily incorporated in

more complex model structures, as we do in this research. (For details on DEFs, see Web Appendix B.)

Turning our attention to our modeling challenge, the primary goal of our model is to infer the individual-level parameters β_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

$$\beta_i^y = \mu^y + \mathbf{W}^y \times \mathbf{z}_i^1, \quad (5)$$

$$\beta_i^a = \mu^a + \mathbf{W}^a \times \mathbf{z}_i^1, \quad (6)$$

such that the individual-level parameters, β_i^y and β_i^a , are a (deterministic) function of mean parameters, μ^y and μ^a ; individual deviations from this mean 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 PCA model (Bishop 1999), the vector \mathbf{z}_i^1 captures the individual-level traits that jointly explain 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, \mathbf{z}_{ik}^1 , is Gaussian-distributed with mean $g(-\mathbf{w}_k^{1'} \times \mathbf{z}_i^2)$, and variance 1,

$$p(\mathbf{z}_{i,k}^1 | \mathbf{z}_i^2, \mathbf{W}^1) = \mathcal{N}[\mathbf{z}_{i,k}^1 | g(-\mathbf{w}_k^{1'} \times \mathbf{z}_i^2), 1] \quad k \in \{1, \dots, N_1\}, \quad (7)$$

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 $g(x) = \log \{ \log [1 + \exp(x)] \}$ 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 nonlinear 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 Web Appendix C). Moreover, the nonlinear relationships are captured by the nonlinear 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,

$$p(\mathbf{z}_{i,k}^2) = \mathcal{N}[\mathbf{z}_{i,k}^2 | 0, 1] \quad k \in \{1, \dots, N_2\}. \quad (8)$$

In summary, we link the individual-level demand and acquisition parameters using a DEF component of two Gaussian

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.

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 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(x)`.

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 three layers, $L \leq 3$, in their empirical applications).¹³

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 1, the model would learn only a single trait to describe the variation across all parameters, which fails to capture the heterogeneity in the demand parameters and their (potentially nonlinear) relationships with acquisition characteristics. Similarly, if the number of components of the higher layer, N_2 , is set to 0, the model would be stripped away from the nonlinear function $g(\cdot)$ that allows the model to capture nonlinear 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. This exercise is required when using maximum likelihood estimation, as more flexibility in a model leads to overfitting following the classical bias-variance trade-off and, therefore, poor performance in holdout samples. However, when using Bayesian inference, this exercise would be not only computationally very costly but also unnecessary, provided that adequate priors such as spike-and-slab or sparse-gamma (Karaletsos and Rättsch 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 manifests in a particular variable only 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 activates only the relevant traits, thus avoiding overfitting the data. Specifically, we use sparse Gamma priors for \mathbf{W}^1 and hierarchical Gaussian ARD priors for \mathbf{W}^y and \mathbf{W}^a , both of which are spike-and-slab-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 and would not affect the model's performance. (For details about these priors, see Web Appendix C.1.)

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) is 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 no specific parameter captures the relevance of the upper-layer traits, we can compute a pseudo- α_m^1 for each upper trait m using the components of the weight matrix \mathbf{W}^1 that map to relevant lower-level traits (for details, see Web Appendix D). Finally, examining the posterior estimates of α^1 and pseudo- α_m^1 —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 data set.

These insights are further developed in Web Appendix D.7, where we explore the dimensionality of the DEF component by analyzing the results of estimating the FIM on simulated data for which we know how many traits are needed. There, we show how the model's performance 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_m^1) are diagnostic of relevant and nonrelevant traits.¹⁴

In summary, we take a hybrid approach to model selection, in which we ensure that the number of prespecified 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.

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 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 HB priors (e.g., multivariate Gaussian). Such an approach would assume that all individual-level parameters (β_i^y and β_i^a) are distributed jointly according to a flexible multivariate distribution in which parameters capture all the potential correlations among the variables. However, this full hierarchical approach would require the model to estimate a very high-dimensional correlation matrix, which can become computationally expensive, especially as the number of acquisition variables increases. In contrast,

¹³ We follow the specifications from Ranganath et al. (2015), where the model is estimated using, at most, three layers ($L \leq 3$). In that paper, the model is trained on two large text corpora (5,900 and 8,000 terms), two matrix factorization tasks on a movie ratings data set (50,000 users and 17,700 movies), and a click data set (18,000 users and 20,000 documents). All of these data sets are considerably larger than our data (both in the simulations and in the empirical application). Furthermore, Tables 3 and 4 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 nonlinear relationships.

¹⁴ The posterior distribution of α and \mathbf{W}^1 from real-world data sets would not display as clear-cut a distinction between traits that are meaningful and those that are not, compared with our simulation analyses. We return to this point when discussing the specification of the FIM for our empirical application.

because the FIM includes ARD priors for the lower layer of the DEF, the model allows for only “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 with each other.¹⁵

The upper layer of the DEF—and in particular, the nonlinear 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 nonlinear, 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 in forming first impressions the moment the assumption of linearity does not hold, either because acquisition variables relate to demand parameters in a nonlinear way or when two (or more) acquisition variables interact in their relationship with the demand parameters. As we show in the “Model Performance” subsection, our FIM specification (which includes the second layer) captures several forms of relationships (including linear, interaction effects, and maximum function) without the need to specify those relationships a priori. This is a desirable property of the model because managers, researchers, and 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 (β_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 multilevel (hierarchical) model in which β_i^y are a function of the observed A_i and some population distribution (Allenby and Rossi 1998; Ansari and Mela 2003; Chan, Wu, and Xie 2011; Rossi, McCulloch, and Allenby 1996). While a linear model is attractive for its simplicity and ease of interpretation, if the underlying relationships between the acquisition variables are not linear (or do not follow the specified relationship, due to variable transformation), the model will fail at inferring individual-level demand parameters for newly acquired customers with a certain

level of accuracy. Although nonlinearities 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 to handle missing observations. In contrast, our modeling framework does not require those types of techniques because we model acquisition characteristics as an outcome.

In conclusion, 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 submodels are projected into a lower-dimension space via a DEF component. The specification of the demand submodel is general such that the modeling framework can be applied to a wide range of business contexts. The submodel 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 customers’ 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 acquisition variables. Second, the model allows for flexible relationships (e.g., nonlinear relationships) among the model components. Third, the model can incorporate acquisition characteristics with missing observations, as these are modeled as outcomes that are easily handled using a Bayesian estimation framework. These benefits will become clearer in the “Model Performance” subsection and “Empirical Application” section, when we compare the predictive accuracy of the FIM with that of several alternative specifications.

Estimation and Identification

We estimate the model using full Bayesian statistical inference with Markov chain Monte Carlo sampling. We sample the parameters from the posterior distribution, which is proportional to the joint,¹⁶

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

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 β_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 (1) to be robust to the possibility that the residual variance of the acquisition variables that is not explained by the main factors in the first step is predictive of demand behavior and (2) 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.

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 Web Appendix C.2.

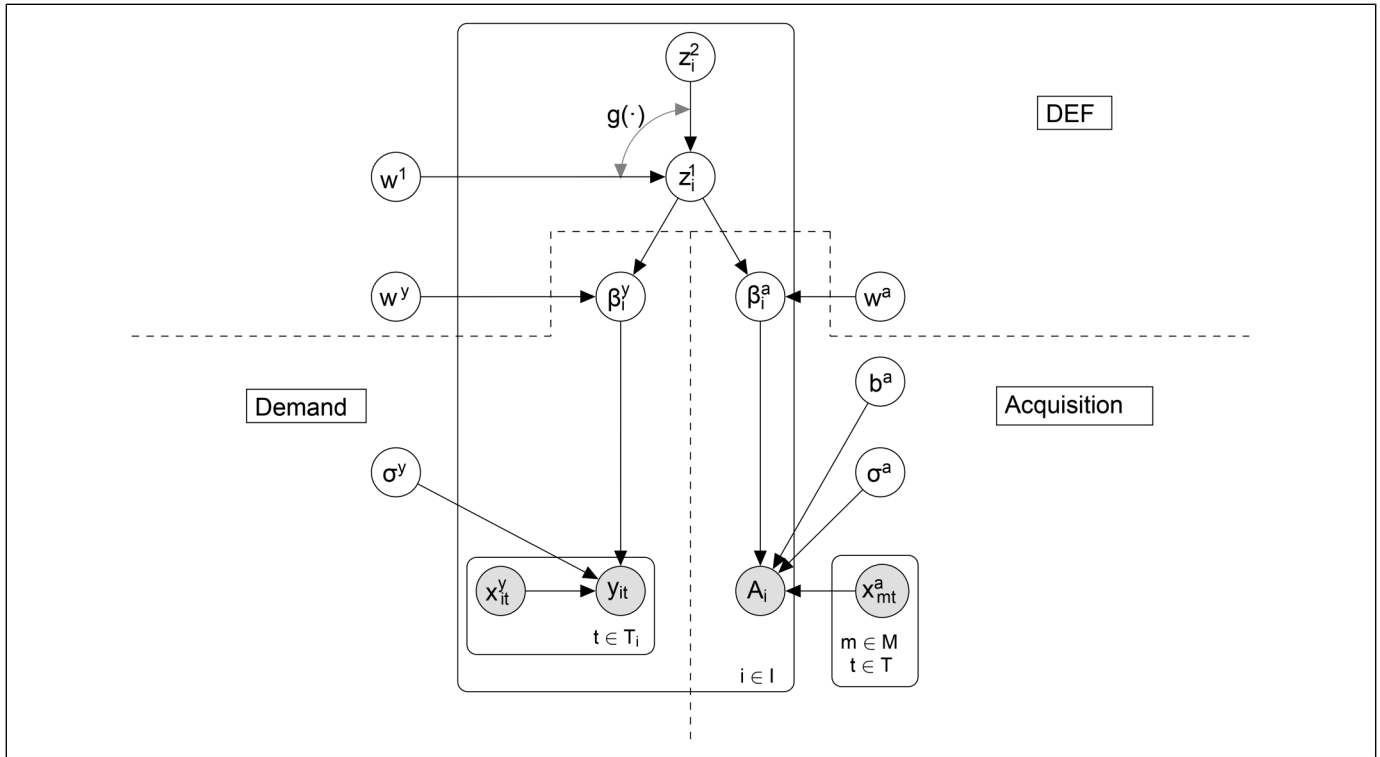


Figure 3. Graphical model of first impressions.

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 (β_i^y , σ^y , β_i^a and σ^a) are identified, provided that the functional forms described in Equations 1 and 2 are well specified. In contrast, not every single parameter of the DEF component is fully identified.

Lower layer. The parameters that link the lower layer of the DEF with β_i^y and β_i^a are identified up to a rotation, similar to a traditional factor analysis model. Specifically, the scales of the lower-layer trait (z_i^1) and weights (w^y and w^a) are identified through the priors scales. Small rotations are identified by the sparsity of the ARD priors (for details, see Web Appendix C)—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 (for details, see Web Appendix E).

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 DNNs, in which the lower layer is a combination 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 (z_i^2) not being identified as different combinations of z_i^2 and w^1 could generate the same value for 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 β_i^y (as corroborated in the “Model Performance” and “Empirical Application” sections), which is the main goal when overcoming the cold start problem.

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 work 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

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.

used to infer the individual-level parameters that will determine their demand (e.g., likelihood that the customer will purchase in the future, their responsiveness to marketing interventions).

More formally, we want to infer $p(\beta_j^y | A_j, \mathcal{D})$ 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 $\Theta = \{\mu^y, \mu^a, \mathbf{W}^y, \mathbf{W}^a, \mathbf{W}^1, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}^a\}$ the population parameters and $\mathbf{Z}_j = \{\mathbf{z}_j^1, \mathbf{z}_j^2\}$, we can write $p(\beta_j^y | A_j, \mathcal{D})$ by both integrating out over the parameters Θ and \mathbf{Z}_j and using the factorization of the joint distribution provided in Equation 9. That is,

$$\begin{aligned}
 p(\beta_j^y | A_j, \mathcal{D}) &= \int p(\beta_j^y, \mathbf{Z}_j, \Theta | A_j, \mathcal{D}) \times d\mathbf{Z}_j \times d\Theta \\
 &= \int p(\beta_j^y | \mathbf{Z}_j, \Theta, A_j) \times p(\mathbf{Z}_j | \Theta, A_j) \times p(\Theta | A_j, \mathcal{D}) \times d\mathbf{Z}_j \times d\Theta \\
 &= \int_{\theta} \left[\int_{\mathbf{Z}_j} p(\beta_j^y | \mathbf{Z}_j, \Theta, A_j) \times p(\mathbf{Z}_j | \Theta, A_j) \times d\mathbf{Z}_j \right] \times p(\Theta | A_j, \mathcal{D}) \times d\Theta \\
 &\approx \int_{\theta} \left[\int_{\mathbf{Z}_j} p(\beta_j^y | \mathbf{Z}_j, \Theta, A_j) \times p(\mathbf{Z}_j | \Theta, A_j) \times d\mathbf{Z}_j \right] \times p(\Theta | \mathcal{D}) \times d\Theta.
 \end{aligned} \tag{10}$$

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 reestimate the model to incorporate A_j . (For a description of the corresponding algorithm, see Web Appendix F.)

Model Performance

Before applying the new modeling framework to the empirical context, we need to demonstrate the model's accuracy 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 β_j^y and can therefore evaluate the model's ability to recover the true parameters using Equation 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 may differ 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 Web 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: (1) 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); (2) a full hierarchical model, where demand and acquisition parameters are jointly distributed according to a multivariate Gaussian distribution with a flexible covariance matrix; and (3) a Bayesian PCA model. As discussed in the "Bringing It All Together" subsection, 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 previous research has

Table 1. Accuracy of Predictions of Demand Parameters for (Out-of-Sample) Customers.

	Scenario 1		Scenario 2		Scenario 3	
	Linear		Quadratic/Interactions		Positive Part	
	R-Squared	RMSE	R-Squared	RMSE	R-Squared	RMSE
Intercept						
HB demand-only	.001	6.703	.020	7.624	.007	8.514
Linear HB	.988	.734	.711	4.113	.783	4.056
Full hierarchical	.988	.735	.704	4.164	.781	4.091
Bayesian PCA	.988	.736	.706	4.484	.780	4.329
FIM	.988	.738	.888	2.661	.928	2.987
Effect of Covariates						
HB demand-only	.005	2.562	.004	4.589	.001	4.604
Linear HB	.986	.303	.258	3.969	.736	2.363
Full hierarchical	.986	.303	.258	3.970	.733	2.378
Bayesian PCA	.986	.301	.245	4.364	.738	2.752
FIM	.986	.302	.515	3.229	.745	2.325

modeled these types of data. To measure the accuracy of each model, we compare the predicted posterior mean with the actual values for the demand parameters (both the intercept and the effect of the covariates) of the 200 out-of-sample customers. Table 1 includes the results for all models across all scenarios.¹⁹ We also include the results of estimating an 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.

First, under a true linear relationship (Scenario 1), the FIM predicts the individual parameters as accurately as the benchmark models. The root mean square error (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 on the basis of 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 also add standard regularization techniques (e.g., LASSO) to the benchmark models. (For details, see Web Appendix D.8.)

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 ($\hat{\beta}_{j1}^y$) versus actual (β_{j1}^y) individual demand intercepts from each model, which displays the superior performance of the FIM, as detailed in Table 1. 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

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 nonlinear relationship between acquisition and demand parameters. (See full details in Web Appendix D.6.)

In summary, 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

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

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.

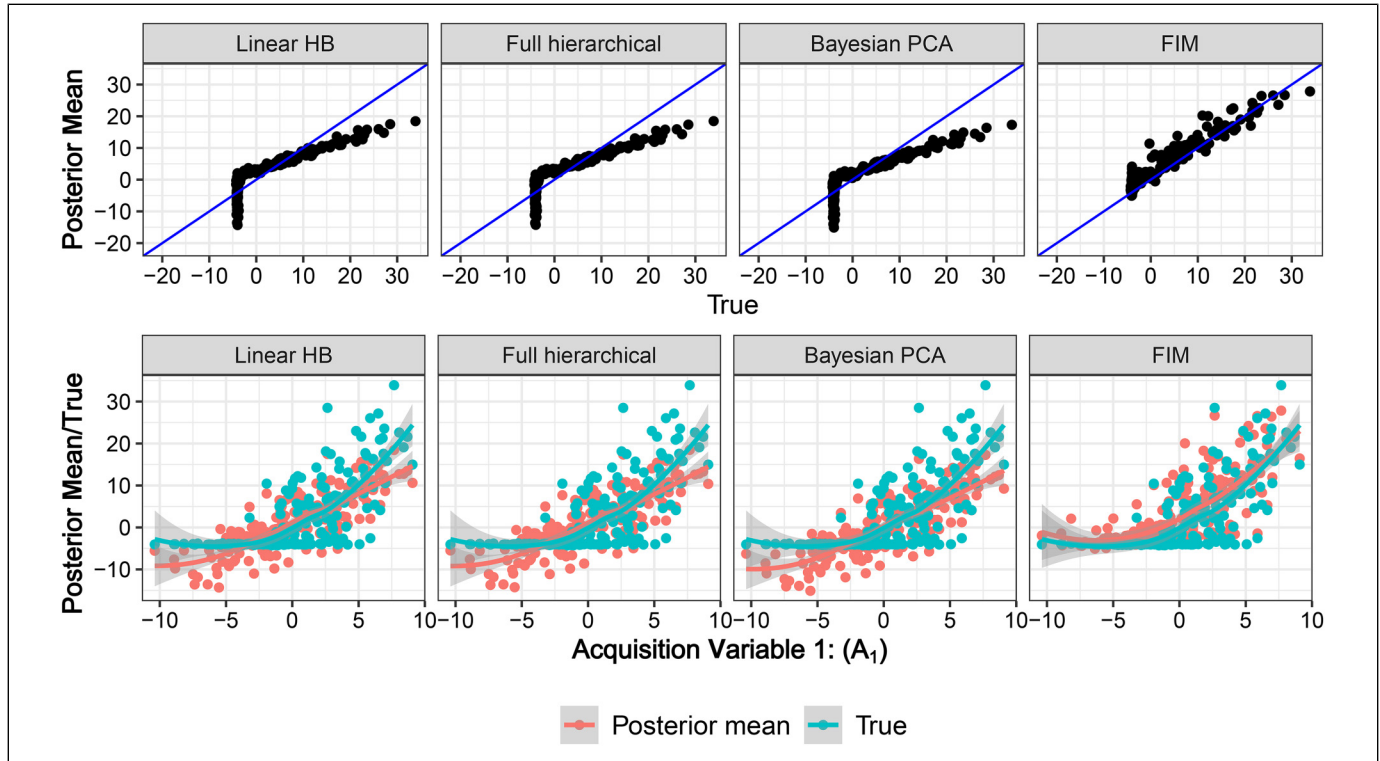


Figure 4. Visualization of model performance for Scenario 3: positive-part individual results of intercept.

Notes: The first row shows the scatter plot of the individual true versus 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).

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

Empirical Application

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 purchase the company's products only via owned stores, either offline (the company owns brick-and-mortar stores across many countries) or online (with one online store per country). Although 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.

Transactional data. We obtain individual-level transactions for registered customers in the six major markets—the United States, the United Kingdom, Germany, France, Italy, and Spain. We observe customers from the moment they make their first purchase (starting in November 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 four years after that date (ending in November 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). Although

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.

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 that 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.

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 Equation 1 is defined as

$$p(y_{it} = 1) = \text{logit}^{-1}[\mathbf{x}_{it}^{y'} \times \boldsymbol{\beta}_i^y + \delta_{\text{rec}} \times \text{Recency}_{it} + \alpha_m], \quad (11)$$

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 $\mathbf{x}_{it}^y = [\mathbf{x}_{it}^y, \text{Recency}_{it}]$ and $\boldsymbol{\sigma}^y = \{\delta_{\text{rec}}, \alpha_1, \dots, \alpha_{M-1}\}$, with M representing the number of markets).

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 United States tends 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.

Although direct marketing 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

Table 2. Summary of Time-Varying Marketing Actions.

Marketing Action	Statistic	Mean	SD	N
Email	Across observations	3.267	4.686	287,584
	Individual average	4.272	3.612	13,473
	Individual SD	3.404	1.790	13,473
	Individual coefficient of variation	1.425	1.082	13,336
Direct marketing	Across observations	1.006	1.889	287,584
	Individual average	1.329	1.018	13,473
	Individual SD	1.731	.769	13,473
	Individual coefficient of variation	2.031	1.205	13,455
Products introduced	Across observations	.923	1.264	287,584
	Individual average	.657	.532	13,473
	Individual SD	.755	.534	13,473
	Individual coefficient of variation	1.354	.478	11,927

product was introduced in a market at the time the first unit of that SKU was sold. We then create a period-/market-level variable representing the number of new products that were introduced in each market in each time period.

Table 2 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 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,

$$\mathbf{x}_{it}^{y'} = [1, \text{Email}_{it}, \text{DM}_{it}, \text{Introd}_{m(i)t}, \text{Season}_{m(i)t}]', \quad (12)$$

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 2), and our model specification, we argue that the potential endogenous nature of the marketing actions is not a main concern in this research (for details, see Web Appendix G.1). 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

²³ As discussed in the “Demand Model” subsection, the proposed FIM can accommodate different demand specifications such as “buy-till-you-die” models or hidden Markov models. 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, Wu and Xie 2011).

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

²⁵ We compute this variable for each market separately because the exact calendar time for the holiday period varies across countries. For example, in the United States, 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.

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, Rossi, and Chintagunta 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.

(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 units bought at the first transaction; Amount is the total amount (in euros) of the ticket at the first transaction²⁶; Discount is a dummy variable that equals 1 if the customer received discounts in the first transaction, and 0 otherwise; and 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 ten-dimensional vector that indicates whether the basket includes a product from a Category, including Body Care, Face Care, Hair Care, Toiletries, and so on, 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 subcategory, and a Travel dummy that equals 1 if the basket includes products on travel size, and 0 otherwise.

Latent representation of shopping baskets: As described previously, 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 six-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:

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

The variation in the acquisition data is very rich (Table 3). For example, 22% of the sample was acquired over the holiday period, 30% of first transactions included at least one discounted product, and 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 Equation 2.

Consistent with the challenges mentioned in the "Modeling Challenges" subsection, some acquisition characteristics are correlated with each other (Table 4)—for example, customers who purchased many items paid less per item (correlation = $-.330$)—and those who bought on discount also paid slightly less than those who paid full price when they were first acquired (correlation = $-.200$). Online first purchases tend to include more items in the basket (correlation = $.411$) and contain products in the face care category (correlation = $.483$). 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 when 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 the "Acquisition Model" subsection, our modeling framework separates these two types of correlations by incorporating firm's market-level actions, $\mathbf{x}_{m(i)\tau(i)}^a$, that potentially affect these acquisition behaviors.

Specifically, we include market-level CRM activities such as number of emails (MarketEmail), direct marketing actions (MarketDM),²⁹ and the number of products introduced by the

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

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

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.

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

Table 3. 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 season	.220	—	13,473
Discount	Whether discounts were applied in transaction	.302	—	13,473
Online	Whether the transaction was online	.176	—	13,473
New product	Whether a new product was purchased	.431	—	13,473
Travel	Whether a travel-size product was purchased	.397	—	13,473
Package size	Average size of products (relative to its subcategory)	1.080	.701	13,352
Avg. BasketDispersion	Average basket dispersion across all dimensions	1.338	.660	9,928
Face Care	Whether a product in the Face Care category was purchased	.352	—	13,473
Hair Care	Whether a product in the Hair Care category was purchased	.120	—	13,473

Notes: For the sake of simplicity, we omit the descriptive statistics for the six BasketNature variables and eight remaining product categories. We also aggregate the BasketDispersion variables by averaging across all dimensions of the word2vec representations. Missing values correspond to first purchases that include products with missing information (and, in the case of BasketDispersion, those with only one item in the basket).

Table 4. Correlations Among Selected Acquisition Characteristics.

	Avg. Price	Quantity	Amount	Size	Holiday	Discount	Online	New Product	Travel	Face care
Avg. price	1.000									
Quantity	-.330	1.000								
Amount	.251	.594	1.000							
Size	.396	-.238	.038	1.000						
Holiday	-.082	.179	.090	-.027	1.000					
Discount	-.200	.285	.184	-.160	.055	1.000				
Online	-.241	.411	.168	-.097	.056	-.049	1.000			
New product	-.036	.250	.248	-.055	.068	.066	.106	1.000		
Travel	-.350	.347	.122	-.348	.088	.289	.009	.149	1.000	
Face care	-.066	.366	.298	-.113	.051	.096	.483	.177	.083	1.000
Hair care	-.124	.261	.121	-.091	-.016	.084	.266	.139	.063	.155

Note: We dropped missing values in pairwise computations only.

firm (Introd) in that period.³⁰ That is,

$$\mathbf{x}_{m(i)\tau(i)}^a = [\text{MarketEmail}_{m(i)\tau(i)}, \text{MarketDM}_{m(i)\tau(i)}, \text{Introd}_{m(i)\tau(i)}]'$$

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.

being acquired; thus, these variables are computed using the set of 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}_{it}^d and $\mathbf{x}_{m(i)\tau(i)}^a$, 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 customer traits and/or a consequence of more products being introduced by the firm when the customer was acquired.

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). 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 who were acquired before April 2014 to use in our

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.

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 Dimensionality of the DEF component, 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 α 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 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 Web 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 the “Model Performance” subsection, we estimate all models (linear HB, Bayesian PCA, and FIM) using No U-Turn Sampling 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 types of data because they explicitly account for latent attrition (e.g., Chan, Wu, and Xie 2011; Schweidel and Knox 2013; Schweidel, Park, and Jamal 2014). For completeness, we test multiple specifications varying the inclusion of time-varying covariates in the transaction process and time-invariant 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 Web Appendix G.3). Finally, we estimate two ML methods widely used for supervised learning (i.e., whether a customer transacts)—namely, a feed-forward DNN and an RF. Both ML models include time-varying covariates, acquisition characteristics, and market-conditions at the moment of acquisition. (For details about the packages used for estimation of the ML methods and related model specifications, see Web Appendix G.4.)

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 nonlinear 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.

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	.051	−3.205	−3.024
	Pop. SD	.364	.086	.245	.549
Email	Pop. mean	.111	.026	.061	.163
	Pop. SD	.167	.031	.110	.235
DM	Pop. mean	.121	.028	.067	.174
	Pop. SD	.137	.023	.094	.182
Product introductions	Pop. mean	−.058	.048	−.164	.024
	Pop. SD	.213	.046	.128	.310
Season	Pop. mean	.361	.072	.235	.502
	Pop. SD	.362	.065	.245	.505

Results

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 ($\beta_{\text{intercept}}^y = -3.110$). Email and direct marketing communications have a positive average impact on purchase ($\beta_{\text{email}}^y = .111$ and $\beta_{\text{dm}}^y = .121$, respectively), whereas product introduction effects are not significant, on average. Finally, customers return to transact more on holiday periods ($\beta_{\text{season}}^y = .361$). Subsequently, we explore the observed heterogeneity in these components (captured by the FIM) as well as the implications for the managers of the firm.

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 Web Appendix G.5.

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_i^y) 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 Web Appendix G.6.) The FIM outperforms all the nested and latent attrition benchmarks in out-of-sample fit (i.e., log-likelihood) as well as at making

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 the “Overcoming the Cold Start Problem” subsection.

Table 6. Comparison with Benchmark Models (Validation Sample).

Model	Log-Like	RMSE		
		Observation	Customer	Period
Linear HB	-2,134.6	.247	1.307	4.570
Latent Attrition w/ Acq.	-2,367.4	.249	1.403	4.951
Latent Attrition w/ Mktg. Actions	-2,194.1	.250	1.361	4.499
Latent Attrition w/ Acq. + Mktg. Actions	-2,384.5	.253	1.421	4.722
Bayesian PCA	-2,010.0	.240	1.184	4.240
Feed-Forward DNN	—	.235	1.095	7.468
RF	—	.236	1.118	6.783
FIM	-1,927.0	.236	1.046	4.058

Notes: Log-likelihood corresponds to the log expected posterior predictive density. Boldfaced cells represent the best model for each metric.

predictions at the observation, customer, and period levels. These results not only corroborate the results presented in the “Model Performance” subsection, now in 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-invariant covariates (e.g., Chan, Wu, and Xie 2011). Not surprisingly, the DNN method provides the most accurate results when looking at observation-level RMSEs, with the FIM performing as well as the RF. However, when looking at customer- and period-level RMSEs, the FIM outperforms all of the models.

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 is the case for many empirical applications. Next, we illustrate the managerial value of these predictions and discuss other insights (provided by the model) that are of managerial relevance.

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 Web Appendix G.7) to predict each individual’s expected number of transactions. We then compare these inferences with customers’ actual behavior using two sets of prediction metrics

Table 7. Identifying Valuable Customers Using Test Customers.

Model	RMSE	% Customers Correctly Classified	
		Top 10%	Top 20%
Linear HB	.157	.151	.253
Latent Attrition w/ Acq.	.520	.113	.207
Latent Attrition w/ Mktg. Actions	.303	.213	.248
Latent Attrition w/ Acq. + Mktg. Actions	.242	.090	.191
Bayesian PCA	.138	.208	.313
Feed-Forward DNN	.098	.349	.450
RF	.106	.193	.310
FIM	.131	.401	.477
Baseline (random)	—	.100	.200
	—	(.067, .127)	(.170, .230)

Notes: 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. Boldfaced cells represent the best model for each metric.

(Table 7). First, we compute the RMSE on the individual-level average number of transactions per period.³⁶ Second, drawing 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 of 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).

As Table 7 shows, the FIM can predict the value of customers reasonably well: the FIM has a lower RMSE than the linear HB and the Bayesian PCA models and is outperformed only by the RF and the DNN. Moreover, linear HB and Bayesian PCA models 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, correctly identifying 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 nonlinearities in the relationship between acquisition characteristics and future demand parameters, it does an excellent job—significantly better than the benchmarks—at sorting customers on the basis of 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 would be instrumental in increasing

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% central posterior interval (CPI).

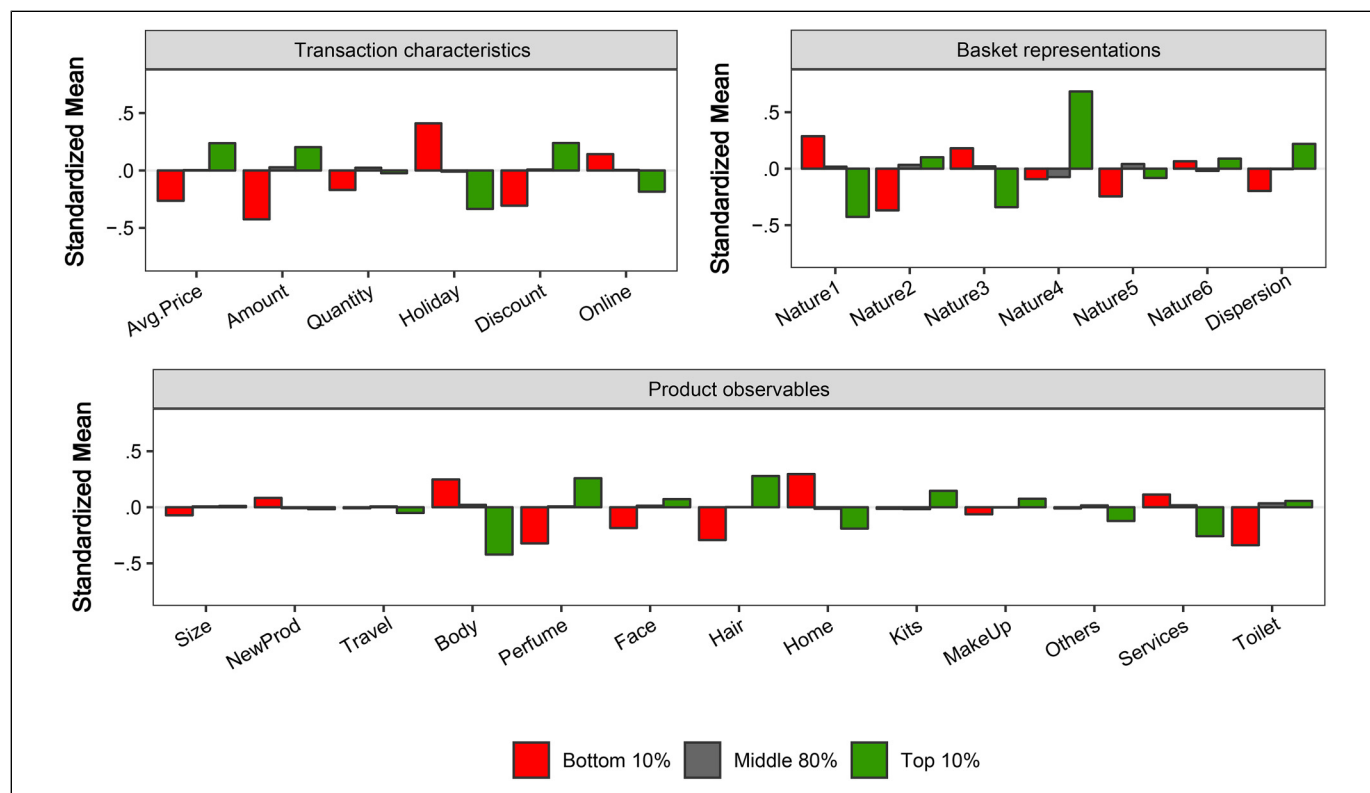


Figure 5. Acquisition characteristics for customers with top/middle/low CLV.

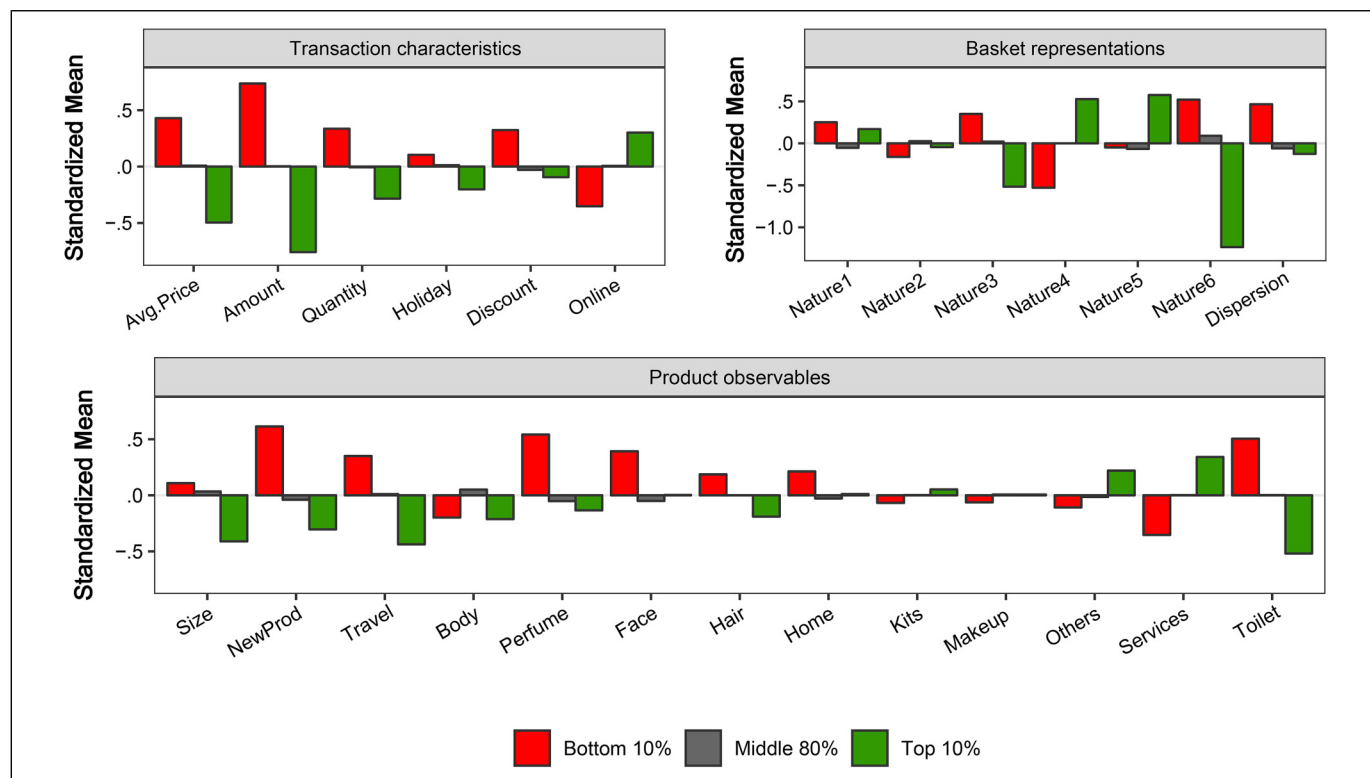


Figure 6. Acquisition characteristics for customers with top/middle/low sensitivity to email.

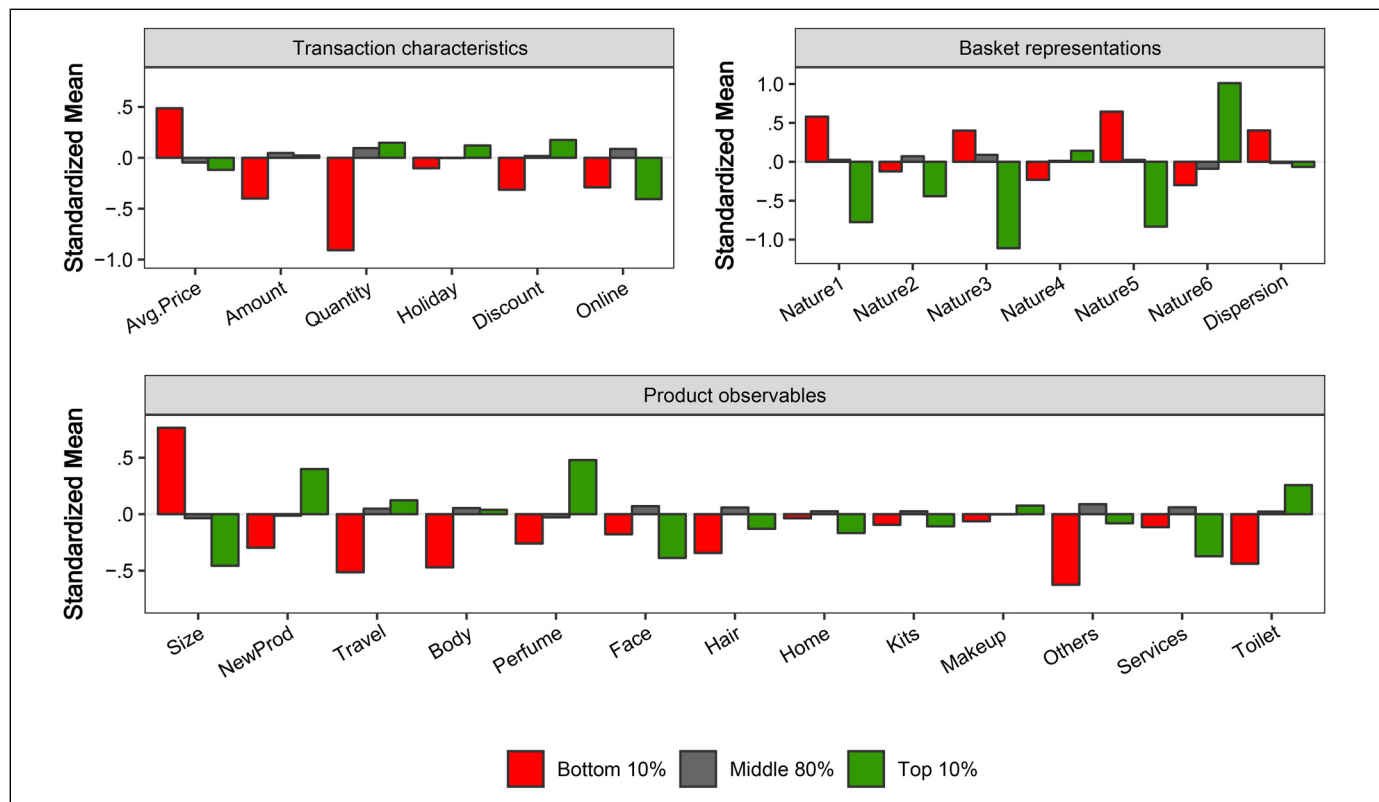


Figure 7. Acquisition characteristics for customers with top/middle/low sensitivity to DM.

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 the “Model Performance” subsection, 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. From 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 we 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 with those in the bottom 10%. They are also characterized to buy certain types of products, as indicated by the high likelihood of including Perfume and Hair products in their first transaction (lower likelihood of including products in the Body Care, Home, and Services categories) as well as by a high score in dimension 4 of the product embeddings.³⁸

We repeat the analysis now sorting customers on the basis of their predicted sensitivity to email (Figure 6) and predicted sensitivity to DM (Figure 7). Consistent with the previous findings, several acquisition characteristics exhibit a nonlinear 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 buy fewer units in their first purchase. With respect to DM, less sensitive customers buy fewer units and more expensive

³⁸ This dimension is related to products such as “Grape Line Showers” and “Olive Harvest Conditioner”; see Table A.1 in Web Appendix A.

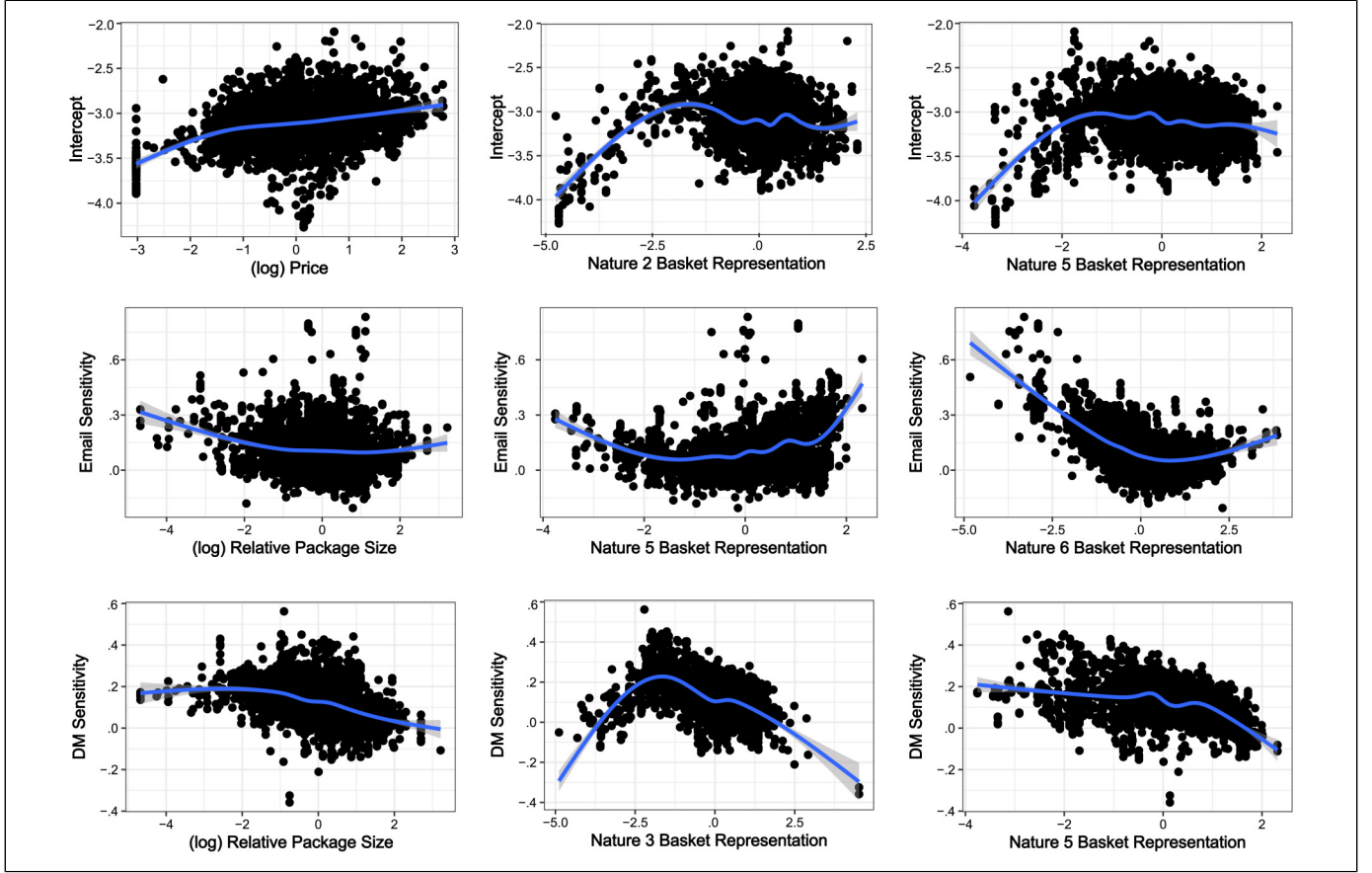


Figure 8. Empirical relationship between the posterior mean and some (continuous) acquisition characteristics.

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.

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 8 shows the individual-level posterior mean of the demand parameter versus the acquisition characteristics for a set of demand parameters and acquisition characteristics. In particular, we find that these plots corroborate that there are nonlinear relationships that the model can uncover.³⁹ Figure 9 explores 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 captures these relationships not only for the intercept but also for other demand parameters. For instance, the holiday season lift is higher for customers who were acquired during a previous holiday season compared with those who were not, but this difference is considerably larger for those who did not purchase a travel-sized product when acquired. In addition, the differences in email sensitivities across customers who received discounts on their first purchase exist only for those who purchased a recently introduced product at acquisition.

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 ML approach, the model connects underlying acquisition and demand parameters using a set of hidden factors modeled via DEFs. The multilayer structure with flexible relationships among layers enables the researcher or analyst to be

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

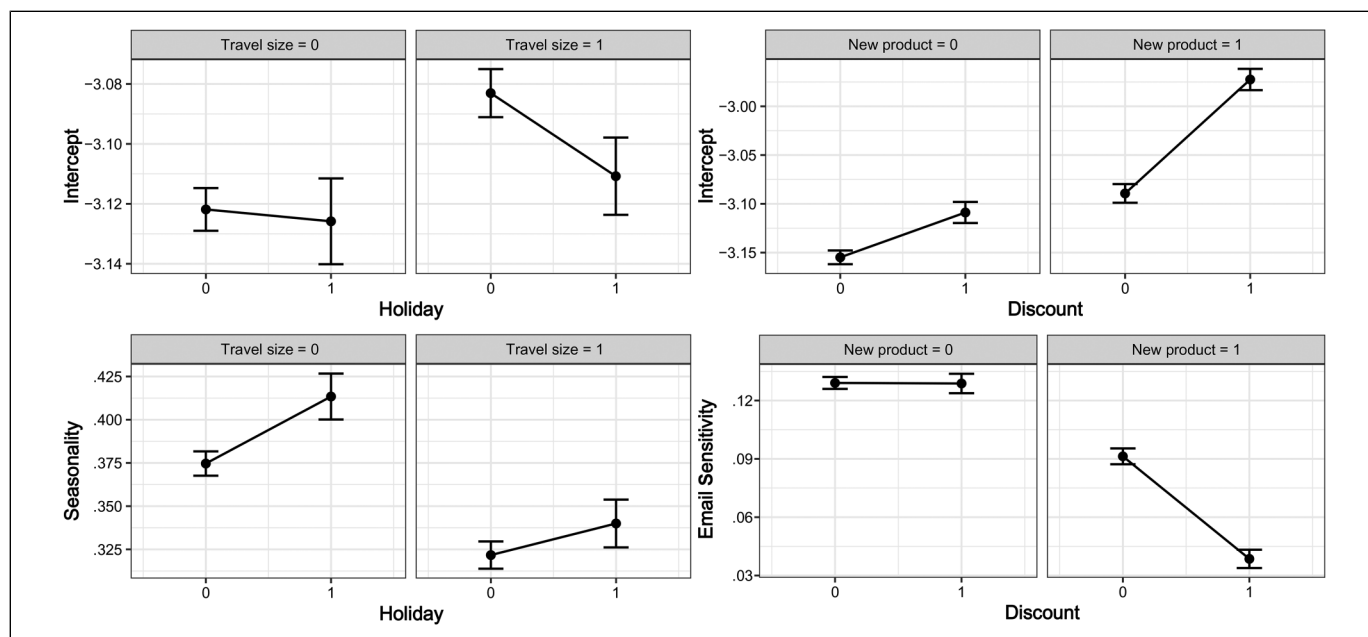


Figure 9. Empirical relationship between the posterior mean and some of the (binary) acquisition characteristics.

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 (e.g., distinguish high-value customers from those unlikely to purchase again) and those most and least sensitive to marketing interventions (e.g., email campaigns, DM). We leverage the model predictions to identify characteristics of first transactions that are predictive of customer behavior in future periods. For example, compared with the rest, top 10% heavy spenders are more likely to be acquired online and their first purchases are more likely to be expensive and contain discounted products; customers identified as most sensitive to email marketing are more likely to be acquired online but buy less expensive products, and their first purchases are more likely to contain fewer units.

These findings suggest that firms can meaningfully categorize customers by drawing on characteristics of their first transactions. We believe this approach to customer segmentation to be promising in that it relies on neither customer-provided data, which is sometimes difficult to obtain (Dubé and Misra 2017), nor external sources of data that could pose privacy concerns. The resulting insights can be used both to prune acquisition data and to inform decisions about the types of variables

worth collecting from customers who 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 observed when a customer makes a first transaction, and it 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 data sets 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 about 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.

Because the main goal of this work is 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.

Acknowledgments

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, the WCAI symposium and the 2020 Virtual Quantitative Marketing Seminar for their comments. The authors are grateful to Hengyu Kuang for excellent research assistantship.

Associate Editor

Fred Feinberg

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

References

- Allenby, Greg M. and Peter E. Rossi (1998), "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89 (1/2), 57–78.
- Anderson, E., Chaoqun Chen, Ayelet Israeli, and Duncan Simester (2021), "Harbinger Categories," working paper.
- Ansari, Asim and Carl F. Mela (2003), "E-customization," *Journal of Marketing Research*, 40 (2), 131–45.
- Artun, Ömer (2014), "What Are Those New Holiday Customers Worth?" Internet Retailer (December 19), <https://www.internetretailer.com/2014/12/19/what-are-those-new-holiday-customers-worth>.
- Ascarza, Eva (2018), "Retention Futility: Targeting High-Risk Customers Might Be Ineffective," *Journal of Marketing Research*, 55 (1), 80–98.
- Bishop, Christopher M. (1999), "Bayesian PCA," in *Advances in Neural Information Processing Systems*, Vol. 11, M. Kearns, S.olla, and D. Cohn, eds. 382–88.
- Bishop, Christopher M. (2006), *Pattern Recognition and Machine Learning*. New York: Springer.
- Braun, Michael, David A. Schweidel, and Eli Stein (2015), "Transaction Attributes and Customer Valuation," *Journal of Marketing Research*, 52 (6), 848–64.
- Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, et al. (2016), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76 (1), 1–29.
- Chan, Tat Y., Chunhua Wu, and Ying Xie (2011), "Measuring the Lifetime Value of Customers Acquired From Google Search Advertising," *Marketing Science*, 30 (5), 837–50.
- Chen, Fanglin, Xiao Liu, Davide Proserpio, and Isamar Troncoso (2020), "Product2vec: Understanding Product-Level Competition Using Representation Learning," SSRN (February 7), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3519358.
- Datta, Hannes, Bram Foubert, and Harald J. Van Heerde (2015), "The Challenge of Retaining Customers Acquired with Free Trials," *Journal of Marketing Research*, 52 (2), 217–34.
- Dew, Ryan and Asim Ansari (2018), "Bayesian Nonparametric Customer Base Analysis with Model-Based Visualizations," *Marketing Science*, 37 (2), 216–35.
- Dew, Ryan, Asim Ansari, and Yang Li (2020), "Modeling Dynamic Heterogeneity Using Gaussian Processes," *Journal of Marketing Research*, 57 (1), 55–77.
- Dubé, Jean-Pierre and Sanjog Misra (2017), "Scalable Price Targeting," technical report, National Bureau of Economic Research.
- Fader, Peter S., Bruce G.S. Hardie, and Kinshuk Jerath (2007), "Estimating CLV Using Aggregated Data: The Tuscan Lifestyles Case Revisited," *Journal of Interactive Marketing*, 21 (3), 55–71.
- Fader, Peter S., Bruce G.S. Hardie, and Ka Lok Lee (2005), "'Counting Your Customers' the Easy Way: An Alternative to the Pareto/NBD Model," *Marketing Science*, 24 (2), 275–84.
- Fader, Peter S., Bruce G.S. Hardie, and Jen Shang (2010), "Customer-Base Analysis in a Discrete-Time Noncontractual Setting," *Marketing Science*, 29 (6), 1086–1108.
- Gopalakrishnan, Arun, Eric T. Bradlow, and Peter S. Fader (2016), "A Cross-Cohort Change-point Model for Customer-Base Analysis," *Marketing Science*, 36 (2), 195–215.
- Hoffman, Matthew D. and Andrew Gelman (2014), "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, 15 (1), 1593–1623.
- Jacobs, Bruno J., Bas Donkers, and Dennis Fok (2016), "Model-Based Purchase Predictions for Large Assortments," *Marketing Science*, 35 (3), 389–404.

- Karaletsos, Theofanis and Gunnar Rätsch (2015), "Automatic Relevance Determination for Deep Generative Models," arXiv preprint, <https://arxiv.org/abs/1505.07765>.
- Knox, George and Rutger van Oest (2014), "Customer Complaints and Recovery Effectiveness: A Customer Base Approach," *Journal of Marketing*, 78 (5), 42–57.
- Kucukelbir, Alp, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei (2017), "Automatic Differentiation Variational Inference," *Journal of Machine Learning Research*, 18 (1), 430–74.
- Kumar, Madhav, Dean Eckles, and Sinan Aral (2020), "Scalable Bundling Via Dense Product Embeddings," arXiv preprint, <https://arxiv.org/abs/2002.00100>.
- Lewis, Michael (2006), "Customer Acquisition Promotions and Customer Asset Value," *Journal of Marketing Research*, 43 (2), 195–203.
- Loupou, Pantelis, Alexandro Nathan, and Moran Cerf (2019), "Starting Cold: The Power of Social Networks in Predicting Non-Contractual Customer Behavior," SSRN, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3001978.
- MacKay, David J.C. (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, Puneet, Peter E. Rossi, and Pradeep K. Chintagunta (2004), "Response Modeling with Nonrandom Marketing-Mix Variables," *Journal of Marketing Research*, 41 (4), 467–78.
- Marr, Bernard (2015), "Big Data: A Game Changer in the Retail sector," *Forbes* (November 10), <https://www.forbes.com/sites/bernardmarr/2015/11/10/big-data-a-game-changer-in-the-retail-sector/>.
- McAuliffe, Jon D. and David M. Blei (2008), "Supervised Topic Models," in *Advances in Neural Information Processing Systems*, 121–28.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013), "Efficient Estimation of Word Representations in Vector Space," arXiv preprint, <https://arxiv.org/abs/1301.3781>.
- Mohamed, Shakir, Zoubin Ghahramani, and Katherine A. Heller (2008), "Bayesian Exponential Family PCA," in *Advances in Neural Information Processing Systems*, Vol. 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds. 1089–96.
- Neal, Radford M. (2012), *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics Series Vol. 118. New York: Springer Science & Business Media.
- Neslin, Scott A., Gail A. Taylor, Kimberly D. Grantham, and Kimberly R. McNeil (2013), "Overcoming the 'Recency Trap' in Customer Relationship Management," *Journal of the Academy of Marketing Science*, 41 (3), 320–37.
- Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, 31 (3), 521–43.
- Park, Sungho and Sachin Gupta (2012), "Handling Endogenous Regressors by Joint Estimation Using Copulas," *Marketing Science*, 31 (4), 567–86.
- Ranganath, Rajesh, Adler Perotte, Noémie Elhadad, and David M. Blei (2016), "Deep Survival Analysis," arXiv preprint, <https://arxiv.org/abs/1608.02158>.
- Ranganath, Rajesh, Linpeng Tang, Laurent Charlin, and David M. Blei (2015), "Deep Exponential Families," in *Artificial Intelligence and Statistics*, 762–71.
- Research and Markets (2016), "Overview & Evolution of the Global Retail Industry," research report (accessed September 23, 2017), <https://www.marketresearch.com/Falke-Information-v4079/Overview-Evolution-Global-Retail-10044025/#:~:text=The%20global%20retail%20sector%20is,of%20people%20throughout%20the%20globe>.
- RJMetrics (2016), "The Ecommerce Holiday Customer Benchmark," (accessed February 5, 2017), <https://rjmetrics.com/resources/reports/the-ecommerce-holiday-customer-benchmark/>.
- Rossi, Peter E., Robert E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15 (4), 321–40.
- Ruiz, Francisco J., Susan Athey, and David M. Blei (2017), "Shopper: A Probabilistic Model of Consumer Choice with Substitutes and Complements," arXiv preprint, <https://arxiv.org/abs/1711.03560>.
- Schmitt, Philipp, Bernd Skiera, and Christophe Van den Bulte (2011), "Referral Programs and Customer Value," *Journal of Marketing*, 75 (1), 46–59.
- Schmittlein, David C., Donald G. Morrison, and Richard Colombo (1987), "Counting Your Customers: Who Are They and What Will They Do Next?" *Management Science*, 33 (1), 1–24.
- Schweidel, David A. and George Knox (2013), "Incorporating Direct Marketing Activity into Latent Attrition Models," *Marketing Science*, 32 (3), 471–87.
- Schweidel, David A., Young-Hoon Park, and Zainab Jamal (2014), "A Multiactivity Latent Attrition Model for Customer Base Analysis," *Marketing Science*, 33 (2), 273–86.
- Shaffer, Greg and Z. John Zhang (1995), "Competitive Coupon Targeting," *Marketing Science*, 14 (4), 395–416.
- Steffes, Erin M., B.P.S. Murthi, and Ram C. Rao (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, Mark D., Robert East, and Wendy Lomax (2013), "Good Customers: The Value of Customers by Mode of Acquisition," *Australasian Marketing Journal*, 21 (2), 119–25.
- Verhoef, Peter C. and Bas Donkers (2005), "The Effect of Acquisition Channels on Customer Loyalty and Cross-Buying," *Journal of Interactive Marketing*, 19 (2), 31–43.
- Villanueva, Julian, Shijin Yoo, and Dominique M. Hanssens (2008), "The Impact of Marketing-Induced Versus Word-of-Mouth Customer Acquisition on Customer Equity Growth," *Journal of Marketing*, 45 (1), 48–59.
- Voigt, Sebastian. and Oliver Hinz (2016), "Making Digital Freemium Business Models a Success: Predicting Customers' Lifetime Value via Initial Purchase Information," *Business & Information Systems Engineering*, 58 (2), 107–18.
- Wang, Yixin and David M. Blei (2019), "The Blessings of Multiple Causes," *Journal of the American Statistical Association*, 114 (528), 1574–96.