The Value of First Impressions:
Leveraging Acquisition Data for Customer Management

(Preliminary version, please do not distribute)

Nicolas Padilla
Eva Ascarza[†]

January 2019

[†]Nicolas Padilla is a doctoral candidate in Marketing, Columbia Business School (email: npadilla19@gsb.columbia.edu). Eva Ascarza is the Jakurski Family Associate Professor of Business Administration, Harvard Business School (email: eascarza@hbs.edu). The authors are grateful to the Wharton Customer Analytics Initiative (WCAI) for providing the data used in the empirical application. The authors thank Bruce Hardie, Kamel Jedidi, 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, and the audiences of the 2018 Marketing Science conference and the WCAI symposium for their comments.

Abstract

The Value of First Impressions: Leveraging Acquisition Data for Customer Management

Firms increasingly have access to richer customer data. What a decade ago was merely a transaction added to a customer database has become a collection of behaviors that a customer engages in while she is making a purchase (e.g., whether her purchase was online or offline, whether she used a tablet or computer, whether she bought a new product or an old best-seller). We posit that these customer decisions carry valuable information for the firm as they could potentially explain a large proportion of the heterogeneity in future behavior, both in terms of what the customer is expected to do (i.e., her lifetime value) and how responsive she will be to marketing actions. The latter point is especially relevant for contexts in which firms do not observe many purchases per individual (e.g., retail) or where targeting occurs soon after customer acquisition. In these contexts, traditional models that only rely on unobserved sources of heterogeneity are unable to help the marketer target customers with precision.

To overcome this limitation, we propose a model that allows marketers to form customers' "first impressions" by leveraging the data collected at the acquisition moment — data that already reside in the firm's database and are available for every customer. The main aspect of the model is that it captures latent dimensions that impact both the variety of behaviors collected at acquisition, as well as future propensities to buy and to respond to marketing actions. Using probabilistic machine learning, we combine deep exponential families with the demand model, flexibly, relating first impressions with consequent customer behavior. We first demonstrate that such a model is flexible enough to capture a wide range of heterogeneity structures (both linear and non-linear), thus being applicable to a variety of behaviors and contexts. We also demonstrate the model's ability to handle large amounts of data while overcoming commonly faced challenges such as data redundancy, missing data, and the presence of irrelevant information. We then applying the model to data from a retail context and illustrate how the focal firm could form customers' first impressions by merely using their transaction database. We show that the focal firm would significantly improve the return on their marketing actions if it targeted just-acquired customers based on their first impressions.

Keywords: Customer Management, Targeting, Deep Exponential Families, Probabilistic Machine Learning

My first impressions of people are invariably right.

- Oscar Wilde

1 INTRODUCTION

Customers are different, not only in their preferences for products and services, but also in the way they respond to marketing actions. Understanding customer heterogeneity is crucial for a variety of marketing problems — from obtaining accurate estimates of the value of current and future customers, to deciding which customers should be targeted in the next marketing campaign. Over the last three decades, the marketing literature has provided researchers and analysts with methods to empirically identify unobserved differences across customers — e.g., customers with higher versus lower propensity to buy (e.g., Schmittlein et al. 1987; Fader et al. 2005, 2010), those who are less sensitive to a price increase (e.g., Rossi et al. 1996; Allenby and Rossi 1998), or those who are more receptive to marketing communications (e.g., Ansari and Mela 2003). However, despite the advances in technological capabilities that allow firms to track customers over long periods of time, many firms face the challenge of not having enough observations per customer to precisely estimate these differences across customers, particularly for customers that were recently acquired. This is especially the case for retail, hospitality services, and non-profit fund raising among other sectors, where a very large proportion of customers are one-time buyers (Fader and Hardie 2002).[1] The lack of repeated observations presents a structural challenge for estimating unobserved differences across customers, precluding firms from leveraging such heterogeneity to differentiate recently acquired customers based on their expected lifetime value and to target marketing actions precisely, which is crucial to secure a second or a third purchase.

Firms in these sectors have traditionally relied on demographics (e.g., age, gender) and/or recency metrics (e.g., how many weeks since your last transaction) to target marketing efforts with limited data (Shaffer and Zhang 1995). These approaches, however, face practical limitations: Recency metrics, for example, do not differentiate among recently acquired customers (as they all were acquired at the same time), and relevant personal information is generally hard to collect. 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 trans-

[1] A recent study among retailers found that between 50% and 80% of customers were one-time shoppers (with an average of 65%), a phenomenon commonly known as the curse of the one-time shopper. This problem is aggravated in online channels, where the average for one-time shoppers is 80%(Zodiac-Metrics 2016).

action added to a customer base is now a collection of behaviors that a customer incurs while she is making such a transaction (e.g., is the purchase online or offline, did she use tablet or computer, did she buy a new product or an old best-seller, did she buy on discount or at full price). While some of these factors may be purely coincidental with the moment in which the customer made her first purchase, others may carry important information as they reflect latent customer preferences/attitudes. Thus, whereas firms in the aforementioned contexts might not observe customers in many occasions, they now have many more cues to form an impression of who these customers are. This "first impression" can then be used to understand heterogeneity across customers, as well as to better target customers right after they are acquired.

In this research we propose a model that allows firms to form first impressions about their customers. The goal of the model is to extract relevant information from the data gathered when a customer made her first purchase, allowing the firm to make inferences about how that particular customer will behave (e.g., respond to marketing interventions) in the future. Drawing from the literature on personality and first impressions formation, the model is based on the premise that the behaviors/choices observed in newly-acquired customers can be informative about underlying traits — traits that will also be predictive of how customers behave in the future. We operationalize these customer traits via a finite set of latent factors and assume that those traits drive, at least partially, the behaviors observed from the customers both at the moment of acquisition as well as in the future. Using probabilistic machine learning methods — in particular deep exponential families — we model those latent traits in a set of interdependent (hidden) layers, allowing the model to reduce the dimensionality of the data while extracting the relevant customers' traits. Such an approach is not only consistent with behavioral theories about purchase motivations, but also allows firms to overcome challenges that usually arise when analyzing multiple sources of customer behavior namely data redundancy, noise, missing data, and the presence of irrelevant information. Furthermore, this modeling framework is extremely flexible — can be incorporated into any demand model — and is able to capture a wide range of relationships (e.g., non-linear) among the observed behaviors.

We first investigate the performance of the proposed Deep Exponential Family Model (DEFM) via simulations. To do so, we benchmark our approach against other models commonly used in marketing, namely a demand model with interaction terms and a full hierarchical Bayesian specification. First, we demonstrate that the proposed first impression model infers customers'

preferences more accurately than any of the benchmarks. We show how, unlike those other models, our approach allows for flexible relationships among the relevant behaviors, enabling the model to accurately infer customers' preferences when those relationships are unknown to the firm/researcher. We further demonstrate how the DEFM is easily scalable as it handles redundant and irrelevant data by reducing dimensionality via those latent factors.

We then apply our model to a retail context. Using transactional data from a retailer as well as information about its product offering and email marketing campaigns, we demonstrate that first impressions are insightful for the firm, not only in terms of how valuable customers will be in the future, but also on their responsiveness to the firm's marketing actions and on their sensitivity to seasonalities on demand. Furthermore, we analyze the value of forming first impressions for newly acquired customers. We show how the proposed approach can better identify heavy spenders (separately from those who are expected to bring less value to the firm) after just one transaction; that is, without the need of observing the customer in multiple occasions. Moreover, we quantify the incremental value (in terms of increase in transactions) that our focal firm would obtain by using the formed first impressions to target just-acquired customers more effectively. For example, we find that a firm's email campaign sent to newly-acquired customers would have increased its ROI by 39.7% had the firm targeted those new customers based on their first impressions.

All in all, in this research we develop a model that links customers' early observed behaviors/choices with future purchase behavior, allowing firms to form "first impressions" of customers right after they have been acquired. At an abstract level, these first impressions capture unobserved customer traits that drive (at least partially) customers behavior, both in the past and in the future. Consequently, by forming customers' first impressions, firms become more accurate at predicting how customers will behave in the future as well as anticipating how they will react to marketing interventions. Conceptually, we build on the literature or personality and first impressions. While not the focus of this research, one could use our model to explore the dimensions that motivate purchase behaviors across industries. Methodologically, our paper contributes to the Customer Relationship Management (CRM) literature by being the first to incorporate the information obtained at the moment of acquisition — generally discarded by firms due to inability to use it effectively — alleviating the curse of dimensionality and redundancy of variables, while allowing for a wide range of (flexible) relationships among behaviors.

Substantively, our research is relevant to marketers facing the challenge of managing customers soon after acquisition. We show how a model of first impressions allows the firm to (1) identify high-value customers, (2) understand heterogeneity in responsiveness to marketing actions, and, ultimately, (3) decide which newly-acquired customers — for whom the firm has only observed one purchase occasion — should be contacted in the next marketing campaign. While some of the behavioral insights might be specific to the context we chose to investigate (i.e., retail), the main finding of showing that first impressions contain valuable information is relevant and generalizable to other contexts as well.

More generally, our model proposes a solution to the so-called "cold start" problem. Because first impressions are the realization of decisions made by a customer (when making her first transaction), they should be informative as to how that customer will behave in the future. In practice, most of what firms know about a newly acquired customer is the information collected when she made her first transaction. Our research highlights that, by not leveraging this information — already available in their databases — firms are leaving money on the table.

2 FIRST IMPRESSIONS

Theories about first impressions originated in the social psychology literature investigating the tendency of humans at creating "first impressions" of others (Asch 1946; Kelley 1950). The main idea behind forming first impressions is to *derive judgements* that can inform social interactions when *limited information is available* (Funder 1987). For example, people commonly form first impressions by incorporating the information from facial appearance or handshaking to *draw inferences of personality traits* (Chaplin et al. 2000; Bar et al. 2006). It has been shown that these impressions can be formed quickly with reasonable levels of accuracy (Ambady and Rosenthal 1992; Ambady et al. 2000; Willis and Todorov 2006; Carlson et al. 2010).

In this research we posit that, analogously, a firm can form first impressions of customers by looking at the behaviors incurred in their first encounter. By integrating all the data obtained at the moment of acquisition, a firm can draw inferences about individual customers' traits — traits that will be predictive of how these customers will behave in the future. Therefore, what is a first impression in our context? It is an *inference* (based on the observed behaviors at the moment of acquisition) that a firm makes about its customers, in order to learn about customers' traits. Should firms draw inferences about *any* customers' traits? Not necessarily, in most cases the firm

will be mainly interested in learning the traits that are related to what a customer might do in the future, both in terms of whether s/he is likely to purchase again and how s/he will respond to the marketing actions. Our approach integrates the (multi-layered hidden) trait model within a demand specification allowing each firm to form first impressions about the traits that relate to its behaviors of interest.

Note that while our approach is conceptually motivated by the above-mentioned stream of literature, the model we propose does not mimic the exact process by which humans form first impressions. Rather, we construct a model to form first impressions that *maximizes the accuracy* at inferring customers traits/preferences. In other words, our approach is free from so-called cognitive biases namely: recency effect, confirmatory bias, order effects, or the impact of pre-specified stereotypes on forming first impressions (e.g., DiGirolamo and Hintzman 1997; Rabin and Schrag 1999; Christopher and Schlenker 2000).

Furthermore, the premise that firms can form first impressions to explain heterogeneity in future behavior is consistent with empirical findings from the CRM literature, more specifically, from the work on customer acquisition that has investigated the relationship between acquisition-related information — e.g., channel of acquisition — and subsequent customer lifetime value (e.g., Verhoef and Donkers 2005; Lewis 2006; Villanueva et al. 2008; Chan et al. 2011; Steffes et al. 2011; Schmitt et al. 2011; Uncles et al. 2013; Datta et al. 2015). Our work similarly investigates relationships between acquisition-related variables and subsequent customer behavior. However, we differentiate from the aforementioned work in two main ways. First and foremost, our end goal is to assist decisions related to the management of already acquired customers (e.g., who should I target in the next campaign?), rather than designing optimal strategies for customer acquisition (e.g., should I offer free trials to increase customer acquisition?). In other words, when forming first impressions the objective is to extract as much observed heterogeneity as possible from initial behaviors while controlling for the firms' acquisition activities, rather than estimating the casual impact of these acquisition variables on future behavior. Second, this literature suggests that customers are inherently different depending on how they have been acquired. We broaden the range of acquisition-related behaviors by looking at not only *how* a customer was acquired (e.g., online vs. offline, trial vs. regular), but also *what* s/he did when s/he was acquired (e.g., what kind of product did she buy on her first purchase? how much did she pay?), hence extracting more information from the initial transaction. The latter aspect is especially relevant for managers/analysts of large

retailers and hospitality businesses, among others, as such information is not only easily observed but typically already exists in the firm's database.

3 A MODEL OF FIRST IMPRESSIONS

We now turn our attention to the specification of the model, first outlining the intuition behind the model before describing the formal specification for each of its components. Then, following a brief discussion about the estimation procedure and parameter identification, we describe how the model is used to form first impressions about customers.

3.1 Model development

We assume that customers have certain traits that are reflected in their behavior, both when they were first acquired as well as in future periods. For example, one of these traits could capture the extent to which a customer is attracted by cheap or heavily discounted products; high levels of this particular trait will likely be reflected in a customer purchasing low price items when s/he was acquired, but also later if/when she purchases again, that customer will likely be influenced by heavy price discounts. Another trait could capture how prone a customer is to use the internet for shopping-related activities; a customer with a high value for that trait not only is more likely to have been acquired online (compared with customers whose trait is low), but also might be more receptive to promotional emails sent from the firm, once the customer has been acquired. We assume these traits to be unobserved to the firm, and stable over time. The purpose of the model will be estimate these customers' traits as well as their influence both in the acquisition variables as well as in future demand. That way, right after a customer has been acquired, the firm can use the inferred customer traits (inferred from her/his acquisition behaviors) to form a first impression about the customer — that is, to predict how the customer will behave in the future.

Naturally, customer behavior is influenced by other factors such as market conditions and seasonalities (e.g., products available in a particular market, holidays), marketing actions targeted to specific customers (e.g., promotional emails), and other (random) variables unobserved to the firm at the moment of purchase. We account for all of these effects by modeling each customer decision — specifically all behaviors observed in the acquisition moment as well as customer demand in future periods — as an *outcome* driven by both the customer latent traits and all other market-dependent, time-varying factors. Modeling the acquisition behaviors as an outcome (rather than as

an input to the model) not only allows us to control for the time-varying factors that shift demand at the moment of acquisition, but also allows for a flexible modeling specification for the latent traits that overcomes challenges such as redundancy, irrelevance of variables, and missing data commonly encounter in the firm's database. (We discuss these challenges in Section 3.1.3 when specifying the model that links the parameters of the acquisition and the demand models).

To sum, our model comprises three main components: (1) The *demand model*, main outcome of interest to the firm, which could include customers transactions, purchase volume, etc., (2) the *acquisition model*, capturing all customers outcomes that are observed to the firm at the moment of acquisition, and (3) the *probabilistic model* that links the underlying customer parameters influencing these two types of behaviors through hidden traits.

3.1.1 Demand model

We start by assuming a general model for demand, for customer i at period t

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

where I represents the total number of customers, T_i denotes the number of periods since the customer was acquired, $\boldsymbol{\beta}_i^{y}$ is a vector containing customer i's effects of covariates $\mathbf{x}_{it}^{y}$, and the vector $\boldsymbol{\sigma}^{y}$ contains the parameters that are common across customers. For example, if the behavior of interest is purchase incidence, one could specify $f^{y}(\,\cdot\,|)$ as a logistic regression, if the behavior modeled is a quantity or a continuous outcome, one could use a Poisson regression or a Gaussian specification, respectively.

To simplify notation we assume the demand model to be independent across periods (given the parameters and covariates), and the distribution $f^{y}(\,\cdot\,)$ to depend only on the scalar $\mathbf{x}_{it}^{y\prime} \cdot \boldsymbol{\beta}_i^{y}$. However, the model can easily be extended for dynamic models with dependence across periods.[2] In summary, we specify a flexible demand model suitable for different specifications, which is parametrized using individual level demand $\boldsymbol{\beta}_i^{y}$ and population-level demand parameters $\boldsymbol{\sigma}^{y}$.

[2]For example, a state-space model (e.g., a hidden Markov model) with state variable s_{it}, will have dependency across periods through $p(y_{it}, s_{it}|y_{i1:t-1}, s_{i1:t-1}) = p(y_{it}|s_{it}) \cdot p(s_{it}|s_{it-1})$. We would implement such a model by having two individual level vectors, $\boldsymbol{\beta}_i^{yq}$ and $\boldsymbol{\beta}_i^{ye}$, as well as two population level vectors, $\boldsymbol{\sigma}^{yq}$ and $\boldsymbol{\sigma}^{ye}$, that would govern transitions among the hidden states and emissions in a state, respectively. We would substitute (1) for $p(y_{it}, s_{it}|y_{i1:t-1}, s_{i1:t-1}, \mathbf{x}_{it}^{y}, \boldsymbol{\beta}_i^{y}, \boldsymbol{\sigma}^{y}) = p(y_{it}|s_{it}, \mathbf{x}_{it}^{y\prime} \cdot \boldsymbol{\beta}_i^{yq}, \boldsymbol{\sigma}^{yq}) \cdot p(s_{it}|s_{it-1}, \mathbf{x}_{it}^{y\prime} \cdot \boldsymbol{\beta}_i^{ye}, \boldsymbol{\sigma}^{ye})$, where $\boldsymbol{\beta}_i^{y} = \begin{bmatrix}\boldsymbol{\beta}_i^{yq} & \boldsymbol{\beta}_i^{ye}\end{bmatrix}$, and $\boldsymbol{\sigma}^{y} = \begin{bmatrix}\boldsymbol{\sigma}^{yq} & \boldsymbol{\sigma}^{ye}\end{bmatrix}$ be the parameters of the demand model.

3.1.2 Acquisition model

We denote A_i the vector of behaviors that are collected at the moment of acquisition, and a_{ik} the k'th component/behavior. Some of these behaviors may be discrete (e.g., whether the customer was acquired online), while others are continuous (e.g., total amount spent in the first transaction). Given a set of parameters we model these acquisition behaviors independently using

$$p(a_{ip}|\mathbf{x}^a_{m(i)\tau(i)}, \mathbf{b}^a_k, \beta^a_{ip}, \boldsymbol{\sigma}^a_p) = f^a_p(a_{ip}|\beta^a_{ip} + \mathbf{x}^a_{m(i)\tau(i)}{}' \cdot \mathbf{b}^a_p, \boldsymbol{\sigma}^a_p) \quad i \in \{1, \ldots, I\},\ p \in \{1, \ldots, P\}, \quad (2)$$

where P is the number of different types of behaviors collected at acquisition, β^a_{ip} is an individual level parameter that reflects tendency to observe such a behavior from individual i, and $\boldsymbol{\sigma}^a_p$ denotes a vector of parameters that are common across individuals. We allow market-level variables (e.g., advertising levels, number of stores, promotional activity in a particular market) to affect acquisition behaviors. Specifically, the vector $\mathbf{b}^a_p$ contains the parameters that measure the impact of a set of market-level covariates $\mathbf{x}^a_{m(i)\tau(i)}$ on behavior p, with $m(i)$ indicating the market customer i belongs to, $\tau(i)$ denoting the time period at which the customers was acquired. Finally, $f^a_p(\ \cdot\ |)$ is a proper distribution to model behavior p. For example, if the p'th behavior is binary (e.g., whether customer bought online) we can use a logit specification for $f^a_p(\ \cdot\ |)$, or if the behavior is categorical, we can use a multinomial logit specification for $f^a_p(\ \cdot\ |)$, or a Gaussian specification if the behavior is continuous. We further define $\boldsymbol{\beta}^a_i = \begin{bmatrix} \beta^a_{i1} & \ldots & \beta^a_{iP} \end{bmatrix}$ and $\boldsymbol{\sigma}^a = \begin{bmatrix} \boldsymbol{\sigma}^a_1 & \ldots & \boldsymbol{\sigma}^a_P \end{bmatrix}$ as the individual- and population-level vectors of acquisition parameters, respectively.

Note that we only have one observation per individual and behavior. Hence, in theory, having an individual-level parameter β^a_{ip} could completely capture the residual variance of a_{ip} that it 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 behaviors, as well as fitting demand, with a reduced set of individual factors or traits. Therefore, the individual level parameters β^a_{ip} will not have full flexibility to accommodate perfectly to the behavior a_{ip}. Rather, these parameters will capture the residual variance that is correlated with the rest of the acquisition variables and with the demand model. This remark will become clearer when we specify the relationship between the individual-level demand and acquisition parameters, $\boldsymbol{\beta}^y_i$ and $\boldsymbol{\beta}^a_i$, as we do next.

3.1.3 Deep probabilistic model: Linking acquisition and future demand

Following the notation introduced above, the firm wants to infer $\boldsymbol{\beta}_i^y$ as accurately as possible, when only the acquisition behaviors (A_i and $\mathbf{x}^a$) have been observed. Being able to achieve such a goal depends mainly on two factors: First, it depends on the amount of information that resides in the acquisition behaviors that is predictive of future behavior. Second, it depends on the ability of the model to extract such information in situations where the underlying relationship between acquisition variables and future demand is unknown, the acquisition data is large, redundant and possibly containing irrelevant information.

The former is an empirical challenge and ultimately depends on the amount and richness of the data collected by firms. Recent advances in technology have helped overcoming this challenge, as most firms nowadays have capabilities to collect many behaviors at the moment of purchase. For example, not only firms know when a customer made her/his first transaction, but they also know which channel s/he used, whether s/he bought in discount, how s/he paid, etc., increasing the amount of information that would be extracted from those observed behaviors.[3]

The latter is a methodological challenge that, in turn, has been amplified by the same technological advances that made data collection more straightforward. As firms collect larger amounts of data, they are also populating their databases with data that are redundant and, in many cases, irrelevant. We overcome this methodological challenge by incorporating a flexible probabilistic model that links the individual-level parameters governing the acquisition and demand models. Specifically, we use a deep exponential family (DEF) component (Ranganath et al. 2015), to relate demand and acquisition parameters hierarchically, through hidden layers. The hierarchical nature of this approach allows the model to identify/extract individual-level traits that affect both acquisition and future demand — traits that will be used by the firm to form first impressions of customers. The presence of multiple layers facilitates the reduction of dimensionality, thus handling redundant and irrelevant variables in a more effective way while accommodating a wide range of possible relationships between acquisition and demand variables.

[3]The amount of data collected by firms might be even richer, including data *prior* to the moment of acquisition. For example, e-retailers would generally collect the identity of websites visited before purchase, search data, etc. When available, those data can be included in the exact same fashion as the acquisition behaviors. For simplicity, we will denote "acquisition" data to all information available to the firm at the moment of acquisition.

Next we provide a short description of DEF models, followed by the specific characterization of the DEF component in the first impression model, and a discussion about what each component of the model represents.

3.1.3.1 Deep exponential families (DEFs) DEFs are probabilistic models that describe a set of observations $\mathbf{X}_i$ with latent variables layered following a structure similar to deep neural networks. These latent variables are distributed according to distributions that belong to the exponential family (e.g., Gaussian, Poisson, Gamma). DEFs not only find interesting exploratory structure in large data sets — similar to deep unsupervised feature learning models — but also enjoy the advantages of probabilistic modeling. Through their connection to exponential families, they support many kinds of data, making them a good candidate to model the wide range of data types encountered in the firm's database. For example, DEFs have been applied to textual data (newspaper articles), binary outcomes (clicks) and counts (movie ratings), being found to give better predictive performance than state-of-the-art models (Ranganath et al. 2015). Finally, DEFs also enjoy the flexibility of graphical models, allowing them to be easily incorporated in more complex model structures, as we do in this research.

As in deep neural networks, DEFs have two sets of variables: layer variables ($\mathbf{z}_i^\ell$) and weights matrices ($\mathbf{W}^\ell$) for the ℓ'th layer. Each layer variable $\mathbf{z}_i^\ell$ is distributed according to a distribution in the exponential family with parameters equal to the inner product of the previous layer parameters $\mathbf{z}_i^{\ell+1}$ and the weights $\mathbf{W}^\ell$, by

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

where $z_{i,k}^\ell$ is the k'th component of vector $\mathbf{z}_i^\ell$, $\mathbf{w}_k^\ell$ is the k'th column of weight matrix $\mathbf{W}^\ell$, $EXPFAM_\ell(\cdot)$ is a distribution that belongs to the exponential family and governs the ℓ'th layer, and $g_\ell(\cdot)$ is a link function that maps the inner product to the natural parameter of the distribution, allowing for non-linear relationships between layers. The top layer is purely governed by a hyperparameter η, that is, $p(z_{i,k}^L) = EXPFAM_L\left(z_{i,k}^L|\eta\right)$, whereas the lowest layer describes the observations, $p(\mathbf{X}_i|\mathbf{z}_i^1, \mathbf{W}^0) = f\left(\mathbf{X}_i|\mathbf{W}^{0\prime}\mathbf{z}_i^1\right)$.

3.1.3.2 Linking acquisition and future demand Turning our attention to the general model of first impressions, we link the individual-level demand and acquisition parameters using

a DEF component of two Gaussian layers, $\mathbf{z}_i^1$ and $\mathbf{z}_i^2$.[4] With reference to the notation introduced in (1) and (2), we set the individual level parameters, $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$, as a (deterministic) function of mean parameters, $\boldsymbol{\mu}^y$ and $\boldsymbol{\mu}^a$, and individual deviations from this mean which are a function of the lower layer vector $\mathbf{z}_i^1$, and weight matrices $\mathbf{W}^y$ and $\mathbf{W}^a$,

$$\boldsymbol{\beta}_i^y = \boldsymbol{\mu}^y + \mathbf{W}^y \cdot \mathbf{z}_i^1 \tag{3}$$

$$\boldsymbol{\beta}_i^a = \boldsymbol{\mu}^a + \mathbf{W}^a \cdot \mathbf{z}_i^1. \tag{4}$$

In other words, the lower level of the DEF model ($\mathbf{z}_i^1$) captures the individual-level traits that affect both the acquisition behaviors and future demand, and is assumed to have dimension N_1.

In principle, N_1 could be as large as the sum of the dimensions of $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$, however, such a specification would suffer from an excess of flexibility which will not allow the model to learn meaningful relationships between demand and acquisition behaviors. Specifically, N_1 could potentially be set equal to the number of variables in the acquisition data ($\dim(A_i)$), plus the number of covariates in the demand model ($\dim(\mathbf{x}_{it}^y)$), plus 1 (the demand intercept). This could easily result in N_1 independent latent traits, each of them determining one element of parameters $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$, thus, not relating acquisition and demand behaviors. Rather, we set N_1 to a lower dimension than the sum of all components in each sub-model such that, not only is the model more tractable, but also enables $\mathbf{z}_i^1$ to capture the individual traits that reflect *both* acquisition and future demand.[5] Similarly as in a Bayesian Principal Components Analysis (Bayesian PCA) model (Tipping and Bishop 1999; Bishop 1999), by reducing the number of components of $\mathbf{z}_i^1$, we constrain the model to explain demand and acquisition with fewer dimensions, thus enabling the model to capture the most relevant traits. The weight matrices $\mathbf{W}^y$ and $\mathbf{W}^a$ capture how each one of these traits manifests in both demand and acquisition behaviors respectively.

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

[4] The model could easily accommodate more layers (e.g., Ranganath et al. (2015) use $L \leqslant 3$ in their empirical application). We find that $L = 2$ is appropriate to represent the data both in our simulations and empirical application.

[5] The exact value of N_1 can be set a priori, provided the researcher has prior information about the number of latent traits, it can also be learned from the data using cross-validation, or it can be automatically determined by the model using sparse priors to force automatic relevance determination (Bishop 1999), as we do in the empirical application (Section 5).

$$p(z^1_{i,k}|\mathbf{z}^2_i, \mathbf{w}^1) = \mathcal{N}\left(z^1_{i,k}|g\left(\mathbf{w}^{1\prime}_k \cdot \mathbf{z}^2_i\right), 1\right) \qquad k \in \{1, \ldots, N_1\}, \tag{5}$$

where $g(x) = \log(\log(1 + \exp(x)))$ is the log-softmax function (Ranganath et al. 2015). The upper layer captures higher-level traits (resembling the structure of neural networks), while allowing for non-linear correlations between the traits in the lower level $\mathbf{z}^1_i$. The correlations among the lower layers components are induced by reducing the dimensionality of top layer ($\mathbf{z}^2_i$ is a vector of length N_2, with $N_2 < N_1$)[6] whereas the non-linear relationships are captured by the non-linear link function $g(\cdot)$, which relates the higher-level traits with the lower-level traits that manifest in demand and acquisition. Finally, we model the upper layer using a standard Gaussian distribution,

$$p(z^2_{i,k}) = \mathcal{N}\left(z^2_{i,k}|0, 1\right) \qquad k \in \{1, \ldots, N_2\}. \tag{6}$$

At first glance, the choice of the layers dimensions N_1 and N_2 may seem cumbersome. On the one hand, a model with low values for N_1 and N_2 may miss relevant correlations that can benefit the predictions of the model. On the other hand, high values of N_1 and N_2 increase the intractability and instability of the model as well as the computational burden of the inference procedure. Theoretically, one could test all possible combinations of these and choose the values using cross-validation, but this exercise would be computationally very costly, if fully Bayesian sampling is employed. Instead, we follow Ranganath et al. (2015) and use sparse priors on the weight traits ($\mathbf{W}^y$, $\mathbf{W}^a$, and $\mathbf{W}^1$) to induce automatic relevance determination. Specifically, we use Gaussian ARD priors for $\mathbf{W}^y$ and $\mathbf{W}^a$ and sparse Gamma priors for $\mathbf{W}^1$, both of which are spike-and-slap-like priors that have shown to perform well on feature selection (e.g., Bishop 2006; Kucukelbir et al. 2017). See Appendix A.1 for details.

Using these priors, the problem reduces to finding a "large enough" number of traits to ensure that all relevant traits are recovered. The sparse priors favor a selection of only relevant traits by setting the weight parameters $\mathbf{W}^y$ and $\mathbf{W}^a$ to values different from zero for relevant traits, and to values close to zero for other non-relevant traits. In other words, a trait only manifests in a particular variable (i.e., the acquisition or demand parameters in our model) if the improvement in

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

fit is significant. Otherwise, that trait is "shut down" by the prior. If one were to set N_1 too high, several traits would be "shut down", with the value of the weights being very close to zero on all variables. In that case, adding more traits (i.e., increasing N_1) would not improve the model as it would just add irrelevant traits with weights all being close to zero. Conversely, if one were to set N_1 too low, all traits would be relevant to explain the variance across variables and no trait will be "shut down," in which case one would not be able to determine whether the model would benefit from having more traits.

As a result, both in the simulation analysis and in the empirical application we estimate the model increasing N_1 and select the model with the lowest N_1 such that at least one trait is "shut down." By doing so, we ensure that the model recovers all relevant traits, while maintaining the tractability of the model for our inference procedure. This criterion becomes clearer when we summarize the results of our empirical application (Section 5.6). Finally, as the spike-and-slab gamma priors induce strong sparseness in $\mathbf{W}^1$, it is sufficient to set the value of $N_2 = N_1 - 1$ such that it decreases the dimensionality of the lower layer in one unit.

3.1.4 Bringing all together

Figure 1 shows the graphical model for the first impressions model, connecting all the individual components. In essence, the model comprises a demand and an acquisition models, whose individual-level parameters are linked through a two-layered DEF component. The main benefit of using the two-layered DEF to connect both types of individual-level parameters instead of a simpler structure such as multivariate distribution, as in a full Hierarchical model or a latent factor analysis or Bayesian PCA, is that the DEF component reduces the dimensionality of the acquisition variables — avoiding the curse of redundancy and irrelevance of variables — while allowing for flexible relationships (e.g., non-linear relationships) among the model components. An alternative but similar specification for the model could be a two-step approach that first reduces dimensionality among the acquisition variables (i.e., connecting z_i^1 to $\boldsymbol{\beta}_i^a$) and then connects those factors with future demand. We choose to connect the lower level of the DEF model with both components jointly in order to: (1) be robust to the possibility that the residual variance of the acquisition variables not explained by the main factors of the first step is predictive of demand behavior; and (2) to inform the choice of factors that are predictive of demand behavior, as in supervised topic models (Mcauliffe and Blei 2008), and therefore, overcome redundancy and irrelevance of acquisition variables simultaneously.

– Insert Figure 1 here –

Finally, a different approach to model first impressions could be to simply specify the individual-level demand parameters ($\boldsymbol{\beta}_i^y$) as a direct function of the acquisition variables (A_i^y). Such a specification would resemble a typical demand model with interactions, or a multi-level (hierarchical) model in which $\boldsymbol{\beta}_i^y$ are a function of the observed A_i and some population distribution (Rossi et al. 1996; Allenby and Rossi 1998; Ansari and Mela 2003). Our approach not only is more consistent (conceptually) with the theory that individuals have certain latent traits that are reflected in multiple behaviors, but also provides methodological benefits such as scalability and flexibility. In particular, our model specification not only is scalable to accommodate many behaviors at the moment of acquisition (i.e., $\dim(A_i)$ being large) but also outperforms all other specifications at accurately inferring individual level parameters, especially in situations in which the underlying relationship among acquisition variables and the parameters governing future demand could take any form. The benefits of this model specification will become clearer in the next section, when we investigate the model performance and compare it with simpler specifications including a demand model with acquisition variables as covariates, a full Bayesian hierarchical model, and a "supervised" Bayesian PCA.[7]

3.2 Estimation and Identification

We estimate the model using full Bayesian statistical inference with MCMC sampling. We sample the parameters from the posterior distribution which is proportional to the joint,[8]

$$
\begin{aligned}
p\left(\{\mathbf{z}_i^1, \mathbf{z}_i^2\}_{i=1}^I, \mathbf{W}^y, \mathbf{W}^a, \mathbf{W}^1, \boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}_a, \{y_{i1:T}, A_i\}_i\right) = \\
&\left[\prod_{i=1}^{I}\prod_{t=1}^{T_i} p(y_{it} \mid \mathbf{x}_{it}^y, \mathbf{z}_i^1, \mathbf{W}^y, \boldsymbol{\sigma}^y)\right] \cdot \left[\prod_{i=1}^{I} p(A_i \mid \mathbf{x}_i^a, \mathbf{z}_i^1, \mathbf{w}^a, \boldsymbol{\sigma}^a, \mathbf{b}_a)\right] \\
&\cdot \left[\prod_{i=1}^{I} p(\mathbf{z}_i^1 \mid \mathbf{z}_i^2, \mathbf{W}^1)\right] \cdot \left[\prod_{i=1}^{I} p(\mathbf{z}_i^2)\right] \\
&\cdot p(\mathbf{W}^y, \mathbf{W}^a, \mathbf{W}^1, \boldsymbol{\mu}^y, \boldsymbol{\mu}^a, \boldsymbol{\sigma}^y, \boldsymbol{\sigma}^a, \mathbf{b}_a). \qquad (7)
\end{aligned}
$$

[7] Note that if we eliminate the second (top) layer ($\mathbf{z}_i^2$), the model would resemble a "supervised" factor analysis or Bayesian PCA. "Supervised" because the latent traits are not only extracted from the acquisition variables or other forms of covariates, but are also estimated using the information from the demand model. In other words, the main difference between our model and a Bayesian PCA model is given by the upper layer of the DEF component, which allows for non-linear relationship among variables.

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

In particular, we use the Hamiltonian Monte Carlo (HMC) algorithm. Compared to the random-walk type of exploration from traditional Metropolis-Hastings algorithms, HMC methods allows us to efficiently explore the posterior distribution by exploiting the gradient of the log of the posterior probability. Furthermore, given that the parameters from the different acquisition behaviors and demand model are highly correlated in the posterior distribution through the layers, HMC is a more efficient procedure to obtain samples from the posterior distribution. Specifically, we use the No U-Turn Sampling (NUTS), 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.[9]

Regarding the identification of the model parameters, the demand and acquisition components ($\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$) are fully identified, provided the functional forms described in (1) and (2) are well specified. On the other hand, the individual components of the DEF model are not fully identified. Like in a deep neural network, the deep layers of the model are not identified, as different combinations of those parameters can reflect the same value for the lower layer. In our model specification, this translates to the value of the top layer ($\mathbf{z}_i^2$) not being identified as different combinations of $\mathbf{z}_i^2$ and $\mathbf{w}^1$ could generate the same value for $\mathbf{z}_i^1$. Regarding the lower layer, and similar to factor analysis or a Bayesian PCA model, the components of the lower layer trait ($\mathbf{z}_i^1$) are only weakly identified through the prior, up to a rotation. Specifically, the scales of the lower layer trait ($\mathbf{z}_i^1$) and weights ($\mathbf{w}^y$ and $\mathbf{w}^a$) are identified through the priors scales. Small rotations are identified by sparse priors (see Appendix A for details) that induce automatic relevance determination — 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. 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 parameter across draws to maintain a consistent interpretation of these parameters (see Appendix B for details).

Finally, it is important to note that this lack of identification in the DEF component does not preclude the model from uniquely identifying the individual-level parameters ($\boldsymbol{\beta}_i^y$), as it is the goal when forming a first impression. We now describe how this model will be used to form first impressions of newly acquired customers.

[9]The code is available from the authors.

3.3 Forming first impressions

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

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

$$
\begin{aligned}
p(\boldsymbol{\beta}_j^y|A_j, \mathcal{D}) &= \int p(\boldsymbol{\beta}_j^y, \mathbf{Z}_j, \Theta|A_j, \mathcal{D}) \cdot d\mathbf{Z}_j \cdot d\Theta \\
&= \int p(\boldsymbol{\beta}_j^y|\mathbf{Z}_j, \Theta, A_j) \cdot p(\mathbf{Z}_j|\Theta, A_j) \cdot p(\Theta|A_j, \mathcal{D}) \cdot d\mathbf{Z}_j \cdot d\Theta \\
&= \int_\theta \left[\int_{\mathbf{Z}_j} p(\boldsymbol{\beta}_j^y|\mathbf{Z}_j, \Theta, A_j) \cdot p(\mathbf{Z}_j|\Theta, A_j) \cdot d\mathbf{Z}_j \right] \cdot p(\Theta|A_j, \mathcal{D}) \cdot d\Theta \\
&\approx \int_\theta \left[\int_{\mathbf{Z}_j} p(\boldsymbol{\beta}_j^y|\mathbf{Z}_j, \Theta, A_j) \cdot p(\mathbf{Z}_j|\Theta, A_j) \cdot d\mathbf{Z}_j \right] \cdot p(\Theta|\mathcal{D}) \cdot d\Theta. \qquad (8)
\end{aligned}
$$

The last equation suggests that if the number of customers in the calibration data is large, we can approximate 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 (i.e., form a first impression of him/her)

once this customer has been acquired, without the need to re-estimate the model to incorporate A_j. (See Appendix C for a description of the corresponding algorithm.)

4 MODEL PERFORMANCE

Next we investigate the accuracy of the model at forming first impressions using a simulation analysis. Unlike other simulation exercises, the goal of this analysis is *not* to confirm that we can recover the parameters of the proposed model. Rather, we use simulations to demonstrate that the proposed model is able to recover customers' preferences accurately, even when the data generating process is different from the modeling assumptions. In reality, marketers (and researchers) never know what the exact relationship is between acquisition preferences and future demand sensitivities, therefore, having a flexible model that performs well in a variety of contexts is of critical importance.

4.1 Simulation design

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

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

of acquisition variables, we then extend the analysis to incorporate dozens of variables and show how the model performs at a large scale.

4.1.1 Data generation process

First, we simulate seven acquisition parameters for seven corresponding acquisition behaviors. In order to resemble what real data would look like, and to test whether our model can account for redundancy in the acquisition data (e.g., the number of items purchased and total amount spent at acquisition being highly correlated), we make some of these acquisition parameters highly correlated among themselves. We operationalize such a relationship by assuming that six of the seven parameters are driven by two main factors $\mathbf{f}_i = \begin{pmatrix} f_{i1} \\ f_{i2} \end{pmatrix}$, where $\mathbf{f}_i \sim N(0, I_2)$. Furthermore, we set the seventh acquisition parameter to be independent of other acquisition parameters as well as independent to future demand preferences. The rationale behind this structure is to resemble the situation in which the acquisition data includes irrelevant data and therefore test whether the model is robust to random noise. More specifically,

$$\begin{aligned} \beta^a_{ip} &\sim N\left(\mu^a_p + B_{1p} \cdot f_{i1}\ , \sigma^{ba}_p\right), && p = 1, \ldots, 3 \\ \beta^a_{ip} &\sim N\left(\mu^a_p + B_{2p} \cdot f_{i2}\ , \sigma^{ba}_p\right), && p = 4, \ldots, 6 \\ \beta^a_{i7} &\sim N\left(\mu^a_7\ , \sigma^{ba}_p\right), \end{aligned} \tag{9}$$

where β^a_{ip} is the p^{th} component of acquisition vector $\boldsymbol{\beta}^a_i$, μ^a_p is the mean of the p^{th} acquisition parameter; B_{1p} and B_{2p} represent the impact of factors 1 and 2 respectively on the p^{th} acquisition parameter; and σ_{ba} the standard deviation of the uncorrelated variation of the p^{th} acquisition parameter. (See Appendix D for details about the values used in the simulations).

Second, we simulate the individual customer preferences for demand; these are the values that the firm is interested in inferring (i.e., to form a first impression about). We simulate three parameters governing the demand model: an intercept and two covariate effects. We generate these individual demand parameters β^y_{ik} as a function of the acquisition parameters $\boldsymbol{\beta}^a_i$, following a general form

$$\beta^y_{ik} \sim N\left(\mu^y_k + g_k(\boldsymbol{\beta}^a_i|\Omega_k), \sigma^{by}_k\right), \qquad k = 1, \ldots, 3, \tag{10}$$

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

- **Scenario 1: Linear**

$$g_k(\boldsymbol{\beta}_i^a|\Omega_k) = \boldsymbol{\omega}_k^{1\prime} \cdot \boldsymbol{\beta}_i^a \tag{11}$$

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

- **Scenario 2: Quadratic/interactions**

$$g_k(\boldsymbol{\beta}_i^a|\Omega_k) = \boldsymbol{\omega}_k^{1\prime} \cdot \boldsymbol{\beta}_i^a + \boldsymbol{\beta}_i^{a\prime} \cdot \Omega_k^2 \cdot \boldsymbol{\beta}_i^a \tag{12}$$

  This pattern captures situations in which the relationship between an acquisition variable and future demand depends on other acquisition-related preferences, or when such a relationship is quadratic. For example, it is possible that a strong preference for discounted products at the acquisition moment relates to price sensitivity in future demand *only* if the customer was purchasing for herself/himself, or outside the holiday period. In that case, the relationship between demand preferences and acquisition variables will be best represented by an interaction term.

- **Scenario 3: Positive part**

$$g_k(\boldsymbol{\beta}_i^a|\Omega_k) = \boldsymbol{\omega}_k^{1\prime} \cdot \begin{pmatrix} \max\{\boldsymbol{\beta}_{i1}^a, 0\} \\ \vdots \\ \max\{\boldsymbol{\beta}_{iP}^a, 0\} \end{pmatrix} \tag{13}$$

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

For each scenario, we simulate the intercept (β^y_{i1}) and the effect of the first covariate (β^y_{i2}) according to the functions $g_1(\cdot)$ and $g_2(\cdot)$ as described in equations (11)–(13), while maintaining the effect of the second covariate (β^y_{i2}) to be a linear function of the acquisition variables. Furthermore, to compare parameters in the same scale across scenarios, we scale demand parameters such that the standard deviation across individuals is equal across all scenarios.

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

$$A_{ip} \sim N(\beta^a_{ip} + {\mathbf{x}^a_{m(i)}}' \cdot \mathbf{b}^a_p, \sigma^a_p), \qquad p = 1, \ldots, 7 \tag{14}$$

$$y_{it} \sim N({\mathbf{x}^y_{it}}' \cdot \boldsymbol{\beta}^y_i, \sigma^y), \qquad t = 1, \ldots, 20. \tag{15}$$

4.1.2 Estimated models

In addition to our proposed DEFM model, we use four benchmark models to form first impressions: (1) a hierarchical Bayesian demand-only model in which acquisition variables are not incorporated, (2) a linear model, where individual demand parameters are a linear function of the acquisition behaviors, (3) a full hierarchical model, where individual demand and acquisition parameters are jointly distributed according to a multivariate Gaussian distribution with a flexible covariance matrix, and (4) a Bayesian PCA model, identical to our proposed model, without the higher layer. For all models we assume the same linear demand model as in the data generation process. We describe these models in more detail.

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

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

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

We acknowledge that such a model would fail to provide individual-level demand parameter estimates for customers that are not in the calibration sample. In other words, the best this model

can provide is to draw the estimates from the population distribution. We include this benchmark to illustrate the problem of estimating parameters when only one observation per customer is observed and most importantly, to have a reference of how much error we should obtain if a first impression model only captured random noise.

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

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

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

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

4.1.2.3 Full hierarchical model For the third benchmark, we endogenize the acquisition behaviors by modeling them as an outcome. Similar to our proposed first impression model (described in Section 3.1), the full hierarchical model estimates acquisition and demand parameters jointly, with the difference that these two sets of parameters are modeled using a standard hierarchical model, rather than connected via DEF models. That is, the full hierarchical model assumes that

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

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

Because of the Gaussian specification for $\boldsymbol{\beta}_i$, this model imposes a linear relationship between $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$; this is, the conditional expectation of $\boldsymbol{\beta}_i^y$ given $\boldsymbol{\beta}_i^a$, is linear in $\boldsymbol{\beta}_i^a$. As such, this model is

mathematically equivalent to the linear HB model. However, the full hierarchical model differs from the linear model if acquisition behavior A_i is not linear in $\boldsymbol{\beta}_i^a$ (e.g. logit or log-normal. Moreover, if the number of acquisition behaviors increases, the full hierarchical model becomes more difficult to estimate due to the dimensionality of the covariance matrix. In this simulation exercise we assume a linear (Gaussian) acquisition model and therefore the linear and full hierarchical models should provide equivalent results. Nevertheless, this is not the case in the empirical application as we incorporate binary acquisition behaviors modeled using a logit specification.

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

$$\boldsymbol{\beta}_i^y = \boldsymbol{\mu}^y + \mathbf{W}^y \cdot \mathbf{z}_i^1 \tag{16}$$

$$\boldsymbol{\beta}_i^a = \boldsymbol{\mu}^a + \mathbf{W}^a \cdot \mathbf{z}_i^1. \tag{17}$$

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

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

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

4.1.3 Estimation and assessing model performance

We calibrate each model using acquisition and demand data for $2,000$ customers. First, we corroborate that all models are equally capable of recovering the individual-level parameters for customers in the calibration sample. In particular, we confirm that the in-sample predictions for $\boldsymbol{\beta}_i^y$ are almost perfect for all model specifications and for all scenarios (see Appendix D for the in-sample predictions). In other words, all models are equally capable of accurately estimating individual-level demand parameters for in-sample customers.

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

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

While the procedure described in Section 3.3 is valid for all models, the expectation $E(\boldsymbol{\beta}_j^y | A_j, \mathcal{D})$ can be computed directly for some of the benchmark models, which we do for simplicity. For example, the HB demand-only model, this computation reduces to compute the expectation of individual draws of $\boldsymbol{\beta}_j^y$ from the population mean, which converges to the posterior mean of the population mean $\boldsymbol{\mu}^y$. For the linear HB model, it reduces to use the linear formulation and the posterior mean estimates of $\boldsymbol{\mu}^y$, Γ, and Δ. For the full hierarchical model, the Bayesian PCA model, and our proposed DEFM model, where acquisition is modeled as an outcome, we compute the posterior of $\boldsymbol{\beta}_j^y$ given A_j using HMC as described in Section 3.3.

4.2 Results

Figure 2 shows the scatter plot of the predicted ($\widehat{\beta}_{j1}^y$) versus actual (β_{j1}^y) individual demand intercepts from each model, for each scenario.[10] Not surprisingly, the HB demand-only model that does not incorporate acquisition behavior in the model (top row of Figure 2) cannot distinguish (hold out) individuals from their population mean. Turning our attention to the other model specifications, we start analyzing the scenario in which the relationship between acquisition and demand parameters is linear (left-most column of Figure 2). Under this scenario, all models are equally capable of predicting demand estimates for (hold out) customers using only their acquisition data. This result is not surprising for the benchmark models as their mathematical specification resembles that of the simulated data. However, when the relationship between the acquisition and demand parameters is not perfectly linear (as it is the case in scenarios 2 and 3), all benchmark models struggle to predict these individual-level estimates accurately. On the contrary, our proposed DEFM model is flexible enough to recover these parameters rather accurately. Note that the

[10] For brevity sake, we present the results for one parameter of the demand model (the intercept), but the results hold for all other parameters as well. (See Appendix D for those results.)

flexibility of the DEFM model comes at no overfitting cost; that is, even when the relationship is a simple linear relationship, our model recovers the parameters as well as the benchmark models, which assume a linear relationship by construction.

– Insert Figure 2 here –

We explore the differences in accuracy more systematically in Table 1. We use two different measures of fit: (1) the (squared) correlation between true $\boldsymbol{\beta}_j^y$ and predicted $\widehat{\boldsymbol{\beta}}_j^y$ (i.e., R-squared) — measuring the model's accuracy in sorting customers (e.g., differentiating customers with high vs. low value, more vs. less sensitivity to marketing actions) — and the root mean square error (RMSE) — measuring the accuracy on predicting the value/magnitude of the parameter itself. Table 1 confirms the results from Figure 2. Under a true linear relationship (Scenario 1), the DEFM predicts the individual parameters as good as the benchmark models. The RMSE of the DEFM is comparable to the benchmark models, and the R-squared is equal to the benchmark models. However, when the relationship among the model parameters is not perfectly linear (Scenarios 2 and 3), the DEFM significantly outperforms the benchmark models in all dimensions. In particular, the R-squared of the DEFM is higher than that of the benchmarks, demonstrating that the model is superior at sorting customers based on their demand parameters. Moreover, the RMSE for the DEFM is substantially lower than that of the benchmarks, indicating that the proposed model predicts the exact magnitude of customer preferences (e.g., purchase probability, sensitivity to marketing actions) more accurately than any of the benchmarks.

– Insert Table 1 here –

To sum, we have demonstrated that not only the DEFM performs as well as the model that mimics the data generating process when the relationship among acquisition and demand parameters is linear (Scenario 1), but also outperforms all the benchmarks when underlying relationships are non-linear. In other words, the DEFM is flexible enough to accurately predict customers' preferences even when such a model is not used to simulate the true parameters. This is an important property because the researcher/analyst never knows the true relationships in the data. In the next section we extend our analysis to investigate the performance of the DEFM at a larger scale.

4.3 Model scalability

While the analysis thus far assumed a handful number of acquisition variables, many firms collect a larger quantity of behaviors when a customer makes her first transaction. These firms do not

necessarily know a priori which variables can be most predictive of demand preferences, and if so, what the underlying relationship between these variables would be. Arguably, one could estimate a model with all the acquisition variables (and linear transformation of those) plus the interactions among those variables, and let the model infer these relationships — e.g., estimate the *Linear HB model* including the interactions among all acquisition variables. However, as the number of acquisition variables increases, including non-linear terms becomes computationally unfeasible and even when the model can be estimated, it will most likely overfit to these relationships.

One of main benefits of the proposed model is that, because of the hidden multi-layer structure, the DEFM automatically identifies relevant acquisition variables and relationships among these and the demand parameters. As a result, its performance does not suffer from large dimensionality in the data. To corroborate this claim we extend the simulation analysis, increasing the number of acquisition variables from 7 to 60. Not only do we replicate all the findings from Section 4.2 (i.e., the DEFM performs as well as the benchmarks in the linear case and outperforms them significantly when relationships are not linear), but we also show that models that incorporate all interactions fail to recover demand preferences, even when we estimate those models using regularization techniques (e.g., LASSO). See Appendix D.5 for the simulation details, the estimated models, and the results.

To conclude, we have illustrated the effectiveness of the DEFM at forming customers' first impressions. In particular, we have demonstrated that the DEFM can accurately predict customer preferences using only acquisition data, even when such a model is not used to simulate the true parameters. While the benchmark models fail to form accurate first impressions of newly-acquired customers when the underlying relationships among variables are not perfectly linear, the DEFM 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' preferences.

5 EMPIRICAL APPLICATION

We apply our model in a retail context to show how a firm can form first impressions of newly acquired customers. The firm would do so by calibrating the DEFM using historical data of its

existing customers and form first impressions of newly acquired customers for whom only the acquisition behaviors are observed.

We start by describing the data. We then show preliminary (model-free) evidence of how acquisition behaviors are predictive of future demand. Next, we describe the details of the model specification for this empirical context and show the fit and prediction performance of our model. Finally, we show how the firm can form first impressions of newly acquired customers to identify valuable customers as well as customers that are the most sensitive to marketing actions.

5.1 Description of the data

We obtain data from an international retailer that sells its own brand of beauty and cosmetic products (e.g., skincare, fragrance, haircare).[11] Customers can only purchase the company's products via owned stores, either offline (the company owns "brick and mortar" stores across many countries) or online (with one online store per country). While the company is present in many countries, marketing and operations are very consistent across markets.

We obtain individual-level transactions for registered customers in the six major markets — USA, UK, Germany, France, Italy, and Spain. We observe customers from the moment they make their first purchase (starting on November of 2010). At the point of purchase, customers are asked to provide their name, email, and address so that they can receive promotions and other marketing communications from the firm.[12] We track their behavior up to 4 years after that date (ending on November of 2014). We have $13,473$ customers, with a minimum of 3 and a maximum of 51 periods of individual observations, resulting in $287,584$ observations.[13] During this time, we observe a total of 15,985 repeated transactions.

In addition to the behavior of the $13,473$ registered customers, we collect data on all purchases made by "anonymous" customers in all six markets — i.e., those who never shared their identity with the firm. While their behavior is not included in our main analysis (the firm can neither track their future behavior nor communicate with them via email or mail), we use these transactional

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

[12] A customer can choose not to provide any personal details, in which case the company does not add her to the customer list, nor can we observe her future behavior.

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

data to control for shocks in distribution channels that affect the timing of the introduction of new products in specific markets.

5.1.1 Acquisition behaviors

We extract the acquisition behaviors from each customer's first transaction — data that had been already collected but not leveraged by the focal firm. In particular, we use seven variables to represent a multi-dimensional vector of acquisition behaviors. `Avg.Price` is 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 transaction; `Holiday` is a dummy variable that equals 1 if customer made her first transaction during the winter holiday period and 0 otherwise;[14] `Discount` is a dummy variable that equals 1 if the customer received discounts in the first transaction, and 0 otherwise; `Online` is a dummy variable that equals 1 if the first transaction was made online, and 0 otherwise; and `NewProduct` is a 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.

The variation in the data is very rich (Table 2). For example, 22% of the sample was acquired over the holiday period, and 18% was acquired online. The standard deviations of price, number of items purchased, and amount are large, reflecting the heterogeneous behavior of customers across the six markets. We explore the correlations among those variables (Table 3). We indeed find that some acquisition behaviors are correlated with each other — e.g., customers who purchased many items paid less per item (correlation= -0.330), and those who bought on discount also paid slightly lower than those who paid full price when they were first acquired (correlation= -0.200). However, these correlations are not very large in magnitude, suggesting that while these variables are related, each of them seems to capture different aspects of customers' underlying traits, enabling us to separately identify the importance of each variable when forming a customer's first impression.

– Insert Tables 2 and 3 here –

Nevertheless, while it is to be expected that some of these variables will be correlated, as they capture different behaviors incurred by the same customer (e.g., a customer who buys many

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

items might look for cheaper products), some of these correlations might also arise from the market conditions at the moment in which a customer was acquired (e.g., if the company introduces all of its new products during the holiday. Customers with `Holiday`= 1 will also have `NewProduct`= 1 and vice versa). If not accounted for, the latter case could be potentially problematic because the model would not be able to separate the effect of being a "holiday customer" from 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 biased. We eliminate the risk of these confounds by incorporating the firm's market-level actions that could potentially affect these acquisition behaviors in the acquisition component of the DEFM. Moreover, because we obtained first impressions data during 4 years from 6 different markets, we have substantial longitudinal variation to rule out any firm-related systematic relationship among acquisition behaviors.

5.1.2 Marketing actions

The firm regularly sends emails and direct marketing to registered customers. The content of these promotional activities is set globally (i.e., the same promotional materials are used across countries, translated to the local language), though their intensity is set by market (e.g., the US tend to send more emails than France).[15] 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.

– Insert Table 4 here –

While direct and email marketing are observed at the individual level (we denote them by `DM` and `Email`, respectively), the availability of new products is not observed at a granular level. We create a new product introduction variable (`Introd`) by combining point-of-sale data (at the SKU level) with a firm-provided SKU list of new products. Specifically, we obtain the list of all

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

new products introduced during the period of our study. We identify the SKUs for all products in that list and infer inventory in each market from *all* purchases observed in that particular market (including all 304,497 transactions from "anonymous" customers). We assume that a new product was introduced in a market at the time the first unit of that SKU was sold.[16] 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 4 shows the summary statistics for the marketing actions summarized across observations and across individuals. For the latter, we summarize individual averages, individual standard deviations, and the individual coefficient of variation. The variation in these data is also very rich, both across customers (capturing mostly differences across markets) and within customers (capturing differences over time).

5.2 Patterns in the data

We first explore the data to see if there is initial evidence that acquisition variables explain differences in subsequent demand behavior across customers, and therefore can be used to form first impressions. For this analysis, we select the customers for whom we observe at least 15 periods and explore the relationship between the total number of transactions they generate during those first 15 periods (excluding the first transaction) and each one of the acquisition variables (Table 5).

Consistent with common belief in the industry (e.g., Artun 2014; RJMetrics 2016), customers that were acquired during the holiday season are less valuable to the firm. On the other hand, customers who bought using discounts on their first transaction generally buy more during the first 15 periods than customers who did not. A similar pattern exists for customers that used the offline channel and those who bought a new product on their first transaction.

– Insert Table 5 here –

For continuous variables (e.g., average price, quantity, and amount), we group customers into four quartiles.[17] We find that customers that bought more expensive products in their first

[16] For this step we include all purchases observed by the firm during our observation window, including those from customers not in our sample (e.g., customers that had been acquired before our observation window or customers who never registered and remained anonymous to the firm).

[17] Values that are exactly equal to the specific quartile ($q_{0.25}$, $q_{0.50}$ and $q_{0.75}$) are grouped in the lower category, this is, the 51% - 75% quartile contains all customers such that the variable falls in the interval $[q_{0.50}, q_{0.75})$. Therefore, the categories are not necessarily exactly equal in size.

transaction tend to buy more frequently in the subsequent 15 periods. Noteworthy, this relationship is not linear. Customers that paid in average the lowest prices per item (those in the 0% - 25% quartile), tend to buy less frequently in their first 15 periods than all other customers. However, there is no significant differences among the top three quartiles. Not surprisingly, customers that bought more units also buy more frequently during their first 15 periods, but similarly as with price, this difference only exists between customers in the lowest quartile and the rest of customers. Moreover, customers that spent in the upper quartile (76% – 100%), tend to buy significantly more frequently in the next 15 periods than all other customers, but customers that spend in the lowest quartile are also more likely to buy than those in the two middle quarters. Taken together, these results suggest non-linear relationships between acquisition behaviors and demand parameters.

While these results provide preliminary evidence that acquisition behaviors are informative of future behavior, this simple analysis is not without caveats. First, this approach does not separate the predicting power of one acquisition variable from that of another. As discussed earlier, we want to obtain the first impression of a customer given all acquisition variables jointly, as doing so will enable us to predict behavior of newly-acquired customers, even as the correlations among acquisition behaviors (perhaps due to firms actions or to changes in behavior) change over time. Second, the analysis thus far does not shed any light about customers' response to marketing actions. These results indicate that "holiday" customers are less likely to transact again. However, are they more/less sensitive to the firm's communication? Should the firm send them an email or a DM? These insights are crucial for the manager interested in elevating the value of the customers already acquired. Finally, this analysis corresponds to a sub-sample of customers — those for whom we observe 15 periods — in order to have a fair comparison across customers over the same number of periods. Ideally, we would like to include all customers, even those who joined later and therefore were not observed for several years, to better capture all the heterogeneity in our data. The proposed model (presented in Section 3) addresses all these caveats.

5.3 Model specification and estimation

Section 3 presented the model in its most general form. Here we briefly outline the model details — for the demand and acquisition components — specific for this empirical context.

5.3.1 Demand model

With reference to (1), we assume a binary demand model of purchase incidence, where $y_{it} = 1$ if customer i transacts at period t, and $y_{it} = 0$ otherwise. We model purchase incidence using a logistic regression model,

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

where we control for latent attrition using recency as a covariate (Neslin et al. 2013) and include market-level fixed effects to capture differences in purchase frequencies across countries. We define the vector of demand time-variant covariates $\mathbf{x}_{it}^y$ as the intercept, firm-initiated marketing actions, and seasonal factors such as holiday periods,

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

where `Email`, `DM`, and `Introd` are the marketing actions as described in Section 5.1.2, and `Season` is a dummy variable that equals 1 for the winter holiday (analogously as the acquisition behavior `Holiday`).

Like most demand models including firm's marketing actions, we face the risk of introducing endogenous variables in our model, potentially preventing us from obtaining unbiased estimates of customers' preferences. This might be the case if, for example, the firm introduced products or run specific campaigns only when periods of lower/higher level of demands were expected, or if the firm sent individual emails or DMs only to customers who are more/less likely to purchase in a particular period. After discussions with the managers of the focal firm and given our model specification, we argue that that is not likely to occur in our data.

First, we observe a rich variation on promotional campaigns and product introductions across periods as well as across markets (Section 5.1). In order to bias our estimates, the omitted variables that drive these firm's decisions would need to affect both demand and firm's actions simultaneously *in all markets*. Other than the holiday periods, which are already accounted for in the model, the firm was not aware of any other systematic factor driving their promotional and product launch decisions. Second, because we include customers' recency and unobserved heterogeneity on purchase frequency (first component of $\boldsymbol{\beta}_i^y$) in our demand model, the identification of the individual-level sensitivity to promotional activity comes mainly from individual differences across periods, allevi-

ating endogeneity concerns arising from potential correlation between the firm's targeting policies in a particular market and customers' level of activity. Nevertheless, in other applications where these conditions do not hold (due to different strategic behavior by the firm), the demand model could be extended to account for the firm's targeting decisions using individual-level responsiveness (Manchanda et al. 2004) or adding correlations between firm decisions and unobserved demand shocks through copulas (Park and Gupta 2012), depending on how these actions are determined by the firm. Those changes would only affect the demand (sub)model and not the overall specification of the DEFM.

5.3.2 Acquisition model

With reference to (2) we model the acquisition variables using linear and logistic regressions. Specifically, each continuous variable p (with $p \in \{\texttt{Avg.Price}, \texttt{Quantity}, \texttt{Amount}\}$),[18] is modeled as

$$a_{ip} = \beta_{ip}^{p} + \mathbf{x}_{m(i)\tau(i)}^{a\prime} \cdot \mathbf{b}_{p}^{a} + \varepsilon_{ip}, \quad \varepsilon_{ip} \sim \mathcal{N}(0, \sigma_{p}^{a}),$$

whereas each discrete variable p (with $p \in \{\texttt{Holiday}, \texttt{Discount}, \texttt{Online}, \texttt{NewProduct}\}$) is modeled as

$$p(a_{ip} = 1) = \text{logit}^{-1}\left[\beta_{ip}^{p} + \mathbf{x}_{m(i)\tau(i)}^{a\prime} \cdot \mathbf{b}_{p}^{a}\right].$$

To control for the overall marketing intensity that a yet-to-be-acquired customer might have been exposed in a particular market at the moment of acquisition, we include market-level number of emails (`MarketEmail`), DMs (`MarketDM`),[19] and the number of products introduced by the firm (`Introd`) in that period.[20] That is,

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

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

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

[20] Note that the number of products introduced in a particular period enters both the demand and the acquisition model ($\mathbf{x}_{it}^{y}$ 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 her first transaction could be a signal of customers traits, and/or a consequence of more products being introduced by the firm when the customer was acquired.

5.3.3 Estimation

We restrict our analysis to periods in which the firm was engaging in marketing activities, which span from October 2012 to November 2014 ($N = 8,985$ customers). In order to mimic the problem faced by the firm, we estimate the model with the transactional behavior of (existing) customers up to April 2014 and use those estimates to form first impressions for customers acquired after April 2014, using only their acquisition variables.[21] Specifically, we split all customers into three groups: *Training*, *Validation*, and *Test*. We randomly select customers that were acquired before April 2014 to use in our *Training* sample ($N = 5,000$) and use their behavior prior to April 2014 to train the models. 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 combined them with those acquired after April 2014 to form our *Test* sample, which we will use to form first impressions to identify valuable customers and to inform our targeting policy ($N = 2,985$).[22] Similarly as in Section 4, we estimate all models (HB demand-only, linear HB, Bayesian PCA and DEFM) using HMC in Stan.[23]

5.4 Results

For brevity sake, we focus our attention to the results that are most relevant for this research, namely the performance of the proposed model at forming first impressions in this empirical setting and the managerial value of doing so. The remaining results, including the parameter estimates for the demand model, are presented in the Web Appendix E.

Table 6 shows the performance of all models on the *Training* sample. The first two columns show the in-sample fit for each of the models, for which we compute log-likelihood and Watanabe-Akaike Information Criterion (WAIC) (Watanabe 2010). Columns 3 thought 6 show different measures of out-of-sample prediction accuracy, computed for customers in the training sample, but using the time periods that were not included in the estimation (i.e., periods after April 2014). We compute log-likelihood as well as the root mean square error (RMSE) for behavioral predictions. In

[21] 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.

[22] 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 targeting policies.

[23] We do not show the Full hierarchical model given its similar performance to the other benchmark models.

particular, we compare the predicted and actual number of transactions at the observation level (i.e., at the customer/period level), at the customer level, calculating the total number of transactions per customer (in "future" periods), and at the period level, computing the total number of transactions per period. While the benchmark models fit the in-sample data better than our proposed model, the DEFM outperforms all benchmarks in the out-of-sample predictions. In other words, whereas the hierarchical models are very flexible at capturing heterogeneity in the training data, the DEFM forecasts the behavior of existing customers with greater accuracy.

– Insert Table 6 here –

We now turn to evaluate the models' accuracy at predicting behavior for customers outside the *Training* data. Recall that the main goal of our model is to accurately form first impressions on just-acquired customers, for whom the firm has no data other than that of the first transaction. Accordingly, a more appropriate test for model accuracy is to examine how the estimated models predict the behavior of customers that were not included in the *Training* sample. Table 7 shows the prediction accuracy of each of the models in the *Validation* sample. Our model outperforms all the benchmarks in out-of-sample fit (i.e., Log-Like) as well as at making predictions at the customer and at the customer/period level. Regarding predicting total sales in a particular period, the DEFM comes in a (very close) second position after the model that does not include acquisition data.

– Insert Table 7 here –

To sum, the DEFM is consistently better than the benchmarks at predicting differences across customers. It not only does so for customers for whom the firm had already observed for sometime (i.e., out-of-sample predictions in Table 6) but most importantly, it is able to accurately predict heterogeneity across customers using only their acquisition behaviors (i.e., out-of-sample predictions in Table 7).

5.5 Forming first impressions of newly acquired customers

We turn to analyze the managerial value of forming first impressions of newly acquired customers. We do so by leveraging the information from customers in the *Test* sample — i.e., those who made their first transaction after April 2014.[24] First, we investigate how accurately the firm can identify

[24] This exercise aims to replicate how a firm would leverage the DEFM in practice.

"heavy spenders" soon after acquisition. That is, whether the firm can form an impression about how valuable customers will be in the future, using only the data from the first transaction. Second, we illustrate how the firm can significantly increase the impact of its marketing efforts by targeting just-acquired customers using their first impressions.

5.5.1 Using first impressions to identify high value customers

We want to test the model's ability to identify high value customers (separately from those who are expected to bring less value to the firm) soon after the customers have been acquired. To do so, we use the estimates from each of the models (estimated on the *Training* sample) and the acquisition behaviors for each customer in the *Test* sample, to predict each individual's expected number of transactions right after s/he has been acquired. We then compare these inferences with the actual behavior using two sets of prediction metrics (Table 8). First, we compute the RMSE on the individual-level average number of transactions per period.[25] Second, we use the model to predict whether a customer will belong to the top 10%, 20% and 25% of highest average number of transactions and report the proportion customers correctly identified/classified. For reference, we compare those figures with what a random classifier would predict (shown in the last row).

– Insert Table 8 here –

As Table 8 shows, the DEFM can better predict the value of customers: the DEFM has a lower RMSE than the Linear HB and the Bayesian PCA models. Moreover, Linear HB and Bayesian models are significantly better than the baseline at identifying valuable customers, which proves that acquisition behaviors carry valuable information to predict the value of customers. Nevertheless, the DEFM improves significantly the identification of valuable customers over the Linear HB and Bayesian PCA models, suggesting that a model that can capture non-linear relationships between acquisition behaviors and demand parameters is better at identifying valuable customers. Across all metrics, we have demonstrated that in this empirical context the DEFM outperforms all the benchmarks at identifying which customers will me more valuable to the firm.

5.5.2 Using first impressions to identify high sensitive customers

The next question we ask is, can a firm identify which customers should (or should not) be targeted in the next campaign based on their acquisition data? We investigate this question by conducting

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

a set of "what-if" analyses that implement different targeting policies on customers in the *Test* sample. We explore three different targeting rules: (1) *No marketing actions*, assuming the company does not send email campaigns to these customers, (2) *Average effect*, assuming the company sends email campaigns to a random subset of these customers with similar frequency as we observed in the calibration data, and (3) *Targeting using first impressions*, assuming the company runs the same email campaigns as in the *Average effect* scenario — i.e., reaching the same number of customers, thus incurring the exact same costs — but prioritizing the recipients for these campaigns using their first impressions. Specifically, we sort customers (in descending order) by their sensitivity to email campaigns, inferred by each first impression model, either linear HB, Bayesian PPCA or DEFM.

Ideally, one would set a field experiment in which the firm can test the new targeting policies in newly-acquired customers. Unfortunately, we could not implement such an experiment with the focal company; thus we have to rely on a model of behavior which we use to simulate the behavior of the *Test* customers under each targeting rule. To that end, we train a demand-only hierarchical model on the *Test* customers and use the estimates to simulate their behavior (for up to 51 periods) under each targeting policy.[26]

Note that we evaluate the scenarios on a different set of customers than those whose behavior was analyzed to determine the targeting policy. This distinction is important for two reasons.First, this approach does not suffer from potential selection bias. By construction, if a policy is derived from insights that "empirically worked" for a sample of customers, evaluating the impact of that policy on the same set of customers would likely overestimate the effect. On the contrary, evaluating the policy on a new set of customers mitigates such a potential bias as we are calculating the "out of sample" impact of such policy. Second, it would be unrealistic to believe that the firm can "test" the new policy on the exact same customers given that their behavior needs to be observed in order to derive the insights. By evaluating the different scenarios on a cross-sectional sample of customers, we aim to simulate a realistic situation in which the firm transforms the already obtained insights into targeting policies that are applied to a new set of just-acquired customers.

– Insert Table 9 here –

[26] We use a demand-only hierarchical model for consistency and simplicity. Note that this model merely aims to replicate how these customers would behave (in lieu of a field experiment), which is different than the first impression models that aim torelate future demand with acquisition variables.

Table 9 shows the posterior mean and posterior standard errors for total simulated sales under each targeting policy. Email campaigns on average have a positive and significant effect on sales (186.657 vs. 182.131; 2.49% increase in transactions), which is consistent with the estimates of the model calibrated on the *Training* sample (presented in Web Appendix E). Interestingly, when targeting based on first impressions using either the Linear HB or the Bayesian PPCA model, the overall effect of the campaign is lower than that of using a random policy. These results suggest that in this empirical context, linear models fail to identify the most sensitive customers based on their acquisition behaviors, incorrectly selecting customers that are less sensitive to email communications. In other words, there are highly-sensitive customers that are not being targeted under this policy which leads to lower sales than the random policy, but still higher than not sending emails altogether. More importantly, targeting customers using the first impressions from our model (DEFM) leads to the highest increase in sales (188.454 vs. 182.131; 3.47% increase in transactions). This increase corresponds to a 39.7% increase in effectiveness of the email campaigns.

Taken together, these results indicate that a firm can form first impressions of customers based merely on their first transaction. However, the relationship between acquisition variables and demand preferences (e.g., email sensitivity) may be non-linear. Therefore, allowing the model to capture these non-linear relationships can enable the firm to increase sales by targeting to just-acquired customers more effectively.

5.6 Additional model insights: Interpreting factors

In addition to enable firms to form first impressions of newly-acquired customers — hence offering a tool for early customer valuation and targeting — exploring the estimates for the (hidden) traits in the DEFM can provide additional insights into customer traits and behaviors. Similarly to the insights provided by a PCA or Factor Analysis model, the estimates of the lower layer in the DEF component can be used to identify which acquisition variables are most related to the behaviors of interest.[27]

Table 10 shows the posterior mean of the weights of each of the rotated trait on each of the acquisition and demand parameters. The first trait is positively correlated with the amount and number of units purchased in the first transaction as well as with whether the purchase occurred

[27] Recall that only the lowest layer of the DEF component is identified (up to a rotation), which is the component that connects acquisition and demand parameters (Figure 1).

in a holiday period. It also correlates with whether it was a discounted purchase, more likely to be an online purchase (than offline), and more likely to buy a recently introduced product than a product that was introduced a long time ago. This trait also correlates with the strength to which the customer responds to future product introductions as well as to email communications. In other words, a customer that scores high on this trait purchased more units and total amount on her first purchase, most likely purchased a recently introduced product, had discounts on that ticket, and there was a higher likelihood that purchase was made online during a holiday period. In addition, this customer would react more positively (or less negatively) to product introductions, and more positively to email communications.

– Insert Table 10 here –

The second trait is positively correlated with the average price paid at acquisition, and negatively correlated with whether the first purchase was made during a holiday period. This second trait also correlates positively with future purchase frequency as well as product introductions' sensitivity, whereas it correlates negatively with direct marketing efforts and email communications. That is, a customer who scores high on this trait most likely purchased more expensive products on her first purchase, with such a purchase most likely not happening during a holiday season. Moreover, this customer would likely be highly valued by the firm as she will have higher purchase frequency in the future, she will react more positively to product introductions, and less strongly to direct marketing and emails.

Finally, the third trait does not weigh heavily in any particular parameter, indicating that, given the observed behaviors investigated in this empirical context, there is very low residual variance left after the first two traits have been taken into account. It is worthwhile to mention that while the interpretation of these factors is specific to the context we investigate, the generalizability of the model is noteworthy. One could employ the DEFM to identify customer traits specific to a particular context as well as generalize it to other types of behavior, beyond those analyzed in this empirical setting.

6 CONCLUSION

We have developed a model that allows firms to form first impressions of newly-acquired customers. The model leverages all the information collected at the moment of acquisition and infers customers'

future behaviors such as propensity to buy and individual response to marketing actions. Our deep exponential family model, or DEFM, connects underlying acquisition and demand preferences using a set of hidden factors that are modeled via deep exponential families. These hidden factors capture the customer traits that drive, at least partially, the observed behaviors both when the customer was first acquired as well as in the future. Similarly to how humans form first impressions of others based on first interactions (e.g., a hand shake, a facial expression), we recommend firms use this model to form customers' first impressions by analyzing the behaviors observed very early on (e.g., what exactly they bought in their first transaction, which channel they used to make a first purchase, what other products they searched for before being acquired).

We have shown how the DEFM can predict a customer's preferences using only the data obtained during her first transaction. This is a very desirable property for firms that face the challenge of not having observed customers in multiple occasions and yet want to target their marketing efforts more effectively. We demonstrated that the DEFM infers customers' preferences accurately, even when the underlying relationship among the acquisition behaviors and the preferences that determine future demand is not linear. Because of the multiple layer structure and flexible relationships among layers, the research/analyst can be agnostic about the (assumed) underlying relationship among variables — we have demonstrated that the proposed model is flexible enough to accommodate different kinds of relationships including both linear and non-linear specifications. Furthermore, the hidden factors automatically extract the relevant information from the existing data — i.e., identify the traits that relate acquisition behaviors with future outcomes — overcoming the challenge (commonly faced by firms) of having a significant amount of redundant and irrelevant data in the customer database.

We have illustrated the benefits of using the DEFM in a retail setting. Using the firm's transactional database, we have shown how using the DEFM enables the focal firm to make individual-level inferences about just-acquired customers. Specifically, we showed how the firm can separate high-value customers (from those who are unlikely to purchase again) as well as identify those who are more sensitive to marketing interventions such as email campaigns. Using a sample of 13,473 customers and the company's marketing activities currently in place, we estimate that the firm could increase the effectiveness of its email campaigns by 39.7% if it targeted customers based on their first impressions. The model can also be used to identify latent traits that characterize customers' behavior in a particular setting as well as identify the acquisition variables that are most

informative to form first impressions. Firms can use these insights to prune their acquisition data and/or to make decisions about what types of variables are worth collecting when customers make the first transaction or visit the company's website in the first place.

All in all, our research highlights that firms are leaving value on the table by not fully leveraging the multiple behaviors observed when a customer makes her first transaction — information already available in most firms' databases. We provide a flexible model that enables managers and analysts to explore and leverage that information, allowing the firms to make better inferences about its customers right after they have been acquired.

While this research highlights the value of using the DEFM to form first impressions, it is also important to acknowledge some limitations of the present research. We have investigated the model performance using linear and logistic specifications for the demand and acquisition models. While the proposed DEFM is very flexible so that it can be adapted to other modelling frameworks, we have not tested how the model would perform empirically in more complex structures. The current model estimation is computationally feasible for datasets with thousands of customers, dozens of time periods, and a handful number of variables (as in our empirical application). While the model scales very easily to situations with more acquisition variables, increasing the sample size to, for example, millions of customers will increase the estimation time substantially, preventing the firm to form customers' first impressions in a timely manner. For those cases, using variational inference might be a better way to estimate and use the model.

A natural extension to this research is to investigate a wider range of acquisition behaviors and its relevance to form customers' first impressions in different contexts. Building on our empirical application, one could further disaggregate the observation from first purchase and incorporate product attributes, or even specific SKUs, that might be informative of customers' future behavior. For example, Anderson et al. (2019) documents the existence of "harbinger products." These are products that, when purchased by a customer in his/her first transaction, are an indicator of the customer being less likely to purchase again. It would be interesting to also see if certain products are predictive of the sensitivity to marketing actions. On a related note, there might be other acquisition behaviors that firms are not currently collecting (e.g., whether the customer visited the store alone and with family) but that could be very valuable in identifying which marketing actions are most likely to increase sales in the future. We encourage future research to continue

investigating these research settings and identifying additional drivers that help forming customers' first impressions.

Finally, we have not formally investigated the latent traits that drive all the observed behaviors. Although the main goal of this work was to provide a flexible model to form customers' first impressions, it would be relevant for researchers and marketers to identify what the individual traits are that characterize shoppers' behavior. To that end, customers' behaviors in a variety of contexts could be measured and estimated in a unifying DEFM model. We hope that this research opens up new avenues for understanding "universal" shopping traits and identifies the behaviors that best relate with those generalizable findings.

References

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

Ambady, N., Bernieri, F. J., and Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. In *Advances in experimental social psychology*, volume 32, pages 201 – 271. Academic Press.

Ambady, N. and Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*.

Anderson, E., Chaoqun, C., Lui, C., and Simester, D. (2019). Do harbinger products signal which new customers will stop purchasing?

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

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

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology*.

Bar, M., Neta, M., and Linz, H. (2006). Very first impressions. *Emotion*.

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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Carlson, E. N., Furr, R. M., and Vazire, S. (2010). Do We Know the First Impressions We Make? Evidence for Idiographic Meta-Accuracy and Calibration of First Impressions. *Social Psychological and Personality Science*.

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

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

Chaplin, W. F., Phillips, J. B., Brown, J. D., Clanton, N. R., and Stein, J. L. (2000). Handshaking, gender, personality, and first impressions. *Journal of Personality and Social Psychology*.

Christopher, A. N. and Schlenker, B. R. (2000). The impact of perceived material wealth and perceiver personality on first impressions. *Journal of Economic Psychology*.

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

DiGirolamo, G. J. and Hintzman, D. L. (1997). First impressions are lasting impressions: A primacy effect in memory for repetitions. *Psychonomic Bulletin and Review*.

Fader, P. S. and Hardie, B. G. S. (2002). A note on an integrated model of customer buying behavior. *European Journal of Operational Research*, 139(3):682–687.

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

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

Funder, D. C. (1987). Errors and Mistakes: Evaluating the Accuracy of Social Judgment.

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

Kelley, H. H. (1950). The Warm-Cold Variable in First Impressions of Persons. *Journal of Personality*, 18(4):431–439.

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

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

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

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

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

Park, S. and Gupta, S. (2012). Handling Endogenous Regressors by Joint Estimation Using Copulas. *Marketing Science*, 31(4):567–586.

Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*.

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

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

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

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

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

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

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

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

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

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

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

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

Willis, J. and Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*.

Zodiac-Metrics (2016). Why you should invest in repeat customers, not one-time shoppers. [Online; accessed 5-February-2017] https://www.zodiacmetrics.com/customer-behavior/repeat-customers-vs-one-time-shoppers/.

7 TABLES AND FIGURES

Table 1: Out of sample predictions of intercept of demand model using acquisition data

	Scenario 1 Linear		Scenario 2 Quadratic/interactions		Scenario 3 Positive part	
	R-squared	RMSE	R-squared	RMSE	R-squared	RMSE
HB demand-only	0.001	6.703	0.007	8.514	0.020	7.624
Linear HB	**0.988**	**0.734**	0.783	4.056	0.711	4.113
Full hierarchical	**0.988**	0.735	0.781	4.091	0.704	4.164
Bayesian PCA	**0.988**	0.736	0.780	4.329	0.706	4.484
DEFM	**0.988**	0.739	**0.872**	**3.306**	**0.891**	**2.616**

Table 2: Summary statistics of acquisition behaviors.

Variable	**Description**	**Mean**	**SD**	**N**
Avg. price (€)	Average price per unit in euros	11.642	10.237	13,473
Quantity	Total number of units bought	4.934	5.298	13,473
Amount (€)	Total ticket amount in euros	39.567	38.433	13,473
Holiday	Whether was acquired during the Holiday	0.220	−−	13,473
Discount	Whether discounts were applied in transaction	0.302	−−	13,473
Online	Whether transaction was online	0.176	−−	13,473
New product	Whether a new product was bought	0.431	−−	13,473

Table 3: Correlations among first impressions.

	Avg. price	Quantity	Amount	Holiday	Discount	Online
Avg. price	1.000					
Quantity	−0.330					
Amount	0.250	0.594				
Holiday	−0.082	0.179	0.089			
Discount	−0.200	0.285	0.184	0.055		
Online	−0.241	0.411	0.168	0.056	−0.049	
New product	−0.036	0.250	0.248	0.068	0.066	0.106

Table 4: Summary of time-varying marketing actions.

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

Table 5: Observed repeated transactions as a function of acquisition behaviors.

Acquisition variable		Transactions		
		Mean	SE	N
Avg. price	0% - 25%	0.85	0.03	2,135
	26% - 50%	1.03	0.03	2,134
	51% - 75%	1.05	0.03	2,134
	76% - 100%	1.03	0.04	2,134
Quantity	0% - 25%	0.92	0.03	2,242
	26% - 50%	1.02	0.03	2,366
	51% - 75%	1.00	0.04	1,976
	76% - 100%	1.02	0.04	1,953
Amount	0% - 25%	0.95	0.03	2,180
	26% - 50%	0.89	0.03	2,156
	51% - 75%	0.91	0.03	2,067
	76% - 100%	1.21	0.04	2,134
Holiday	No	1.02	0.02	6,978
	Yes	0.85	0.04	1,559
Discount	No	0.96	0.02	6,210
	Yes	1.08	0.03	2,327
Online	No	1.00	0.02	7,112
	Yes	0.93	0.04	1,425
New product	No	0.90	0.02	4,306
	Yes	1.08	0.03	4,231

Table 6: Model fit and prediction accuracy for the *Training* sample

	In-sample		Out-of-sample (future periods)			
	Log-Like	WAIC	Log-Like	RMSE		
Model				Observation	Customer	Period
HB demand-only	−8187.2	17983.1	−5402.4	0.197	0.650	40.9
Linear HB	−8091.0	**17940.0**	−5524.0	0.200	0.693	54.5
Bayesian PPCA	**−8030.8**	17971.1	−5522.8	0.200	0.679	44.9
DEFM	−9136.9	18805.9	**−5237.2**	**0.193**	**0.552**	**35.9**

Table 7: Model fit and prediction accuracy for the *Validation* sample

	Log-Like	RMSE		
Model		Observation	Customer	Period
HB demand-only	−2012.1	0.237	1.142	**3.621**
Linear HB	−2049.2	0.238	1.166	3.957
Bayesian PPCA	−2037.3	0.239	1.156	3.814
DEFM	**−1869.9**	**0.229**	**0.901**	3.636

Table 8: Prediction of valuable customers using *Test* customers.

		% customers correctly classified		
Model	**RMSE**	**Top 10%**	**Top 20%**	**Top 25%**
Linear HB	0.123	0.136	0.236	0.282
Bayesian PPCA	0.120	0.134	0.235	0.283
DEFM	**0.098**	**0.431**	**0.454**	**0.440**
Baseline (random)	–	0.100	0.200	0.250
	–	(0.067,0.127)	(0.170,0.230)	(0.226,0.276)

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

Table 9: Sales under different targeting policies

	Simulated sales	
Targeting policy	Posterior mean	Posterior standard error
No marketing actions	182.131	0.252
Average effect	186.657	0.243
Targeting using first impressions		
Linear HB	183.138	0.242
Bayesian PPCA	183.447	0.243
DEFM	**188.454**	0.248

Table 10: Rotated traits weights' on acquisition and demand variables

Model component	Parameter	Trait 1	2	3
Acquisition ($\mathbf{W}^a$)	Avg. price (log)	0.004	0.281	−0.014
	Amount (log)	0.973	0.088	−0.005
	Quantity (log-log)	1.003	−0.016	−0.004
	Holiday	0.499	−0.295	−0.005
	Discount	0.952	−0.096	−0.008
	Online	0.947	0.031	0.003
	New product	0.986	−0.025	−0.001
Demand ($\mathbf{W}^y$)	Intercept	0.019	0.151	0.021
	Email	0.095	−0.047	0.012
	DM	−0.013	−0.069	0.022
	Product introductions	0.264	0.107	0.011
	Season	0.011	−0.033	−0.014

Figure 1: Graphical model of first impressions

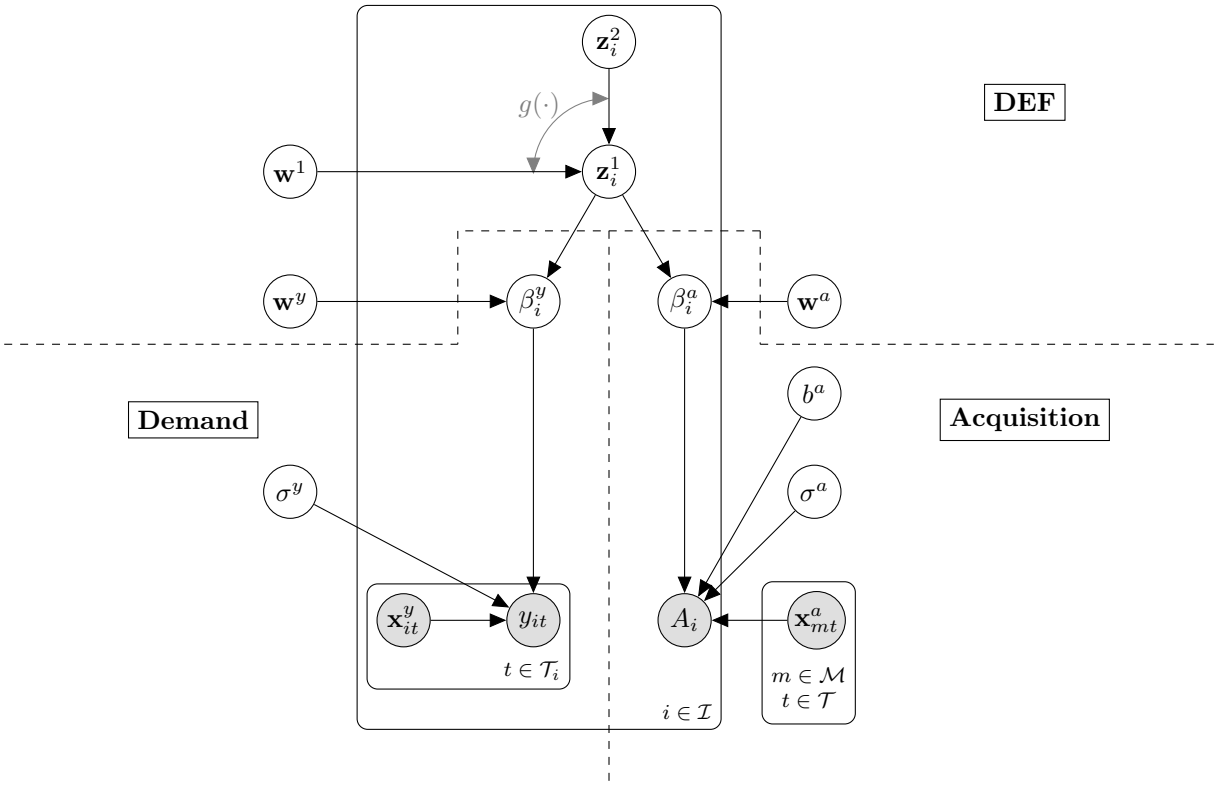
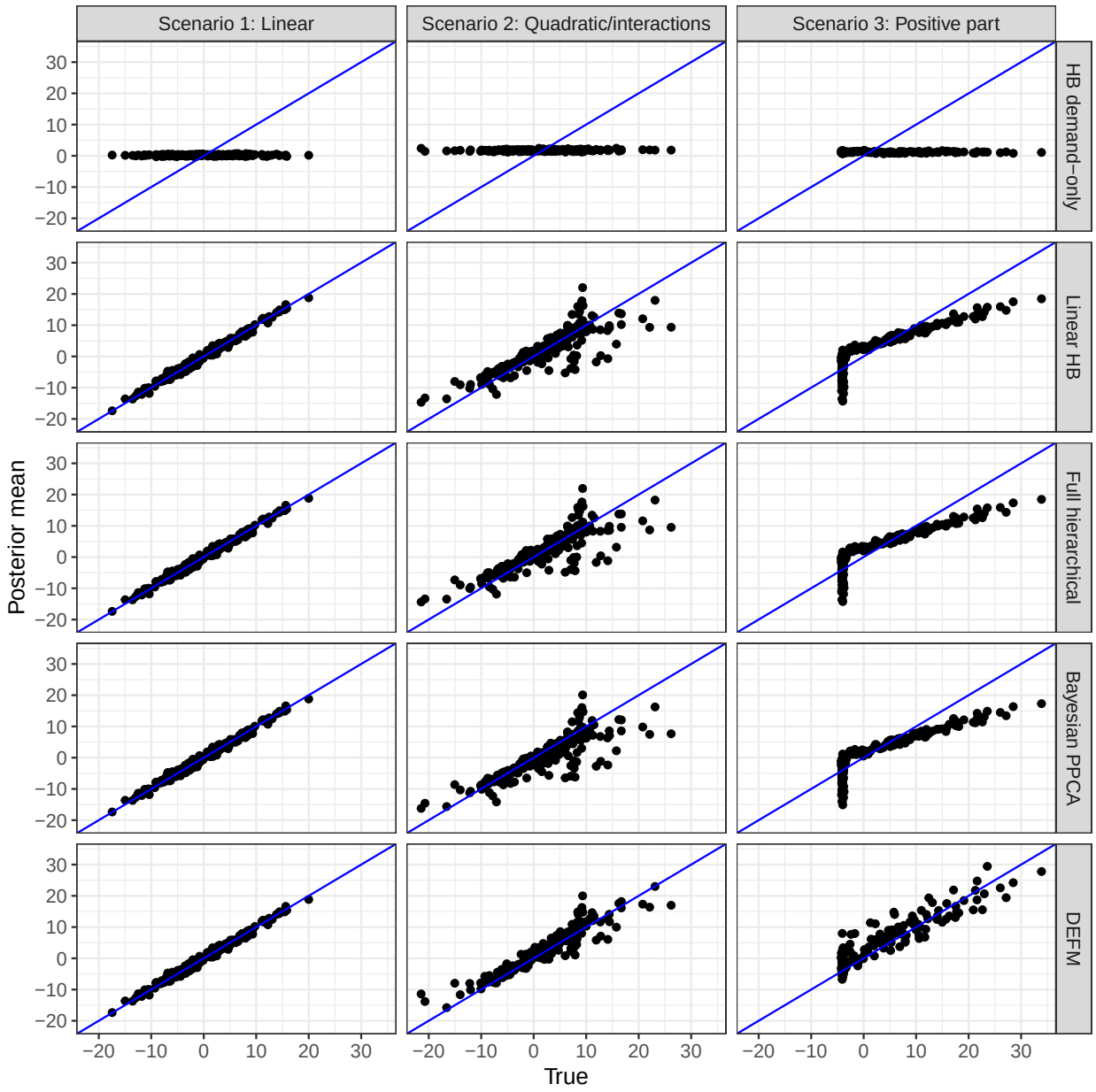


Figure 2: Individual posterior mean vs. true intercepts of the demand model. Each dot represents a customer from the hold out set; i.e., only their acquisition behaviors are used to form first impressions about their individual-level parameters. In blue, the 45 degree line represents perfect predictive power.



WEB APPENDIX

A Model priors and automatic relevance determination component

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

A.1 Automatic relevance determination

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

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

$$p(\mathbf{w}_{nk}^y) = \mathcal{N}\left(\mathbf{w}_{nk}^y | 0, \sigma^y \cdot \alpha_k^1\right) \tag{18}$$

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

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

$$p(\mathbf{w}_{pk}^a) = \begin{cases} \mathcal{N}\left(\mathbf{w}_{pk}^a | 0, \alpha_k^1\right) & \text{if } p \text{ is discrete} \\ \mathcal{N}\left(\mathbf{w}_{pk}^a | 0, \sigma_p^a \cdot \alpha_k^1\right) & \text{if } p \text{ is continuous} \end{cases}, \tag{19}$$

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

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

$$p(\mathbf{w}_{km}^1) = \text{Gamma}\left(\mathbf{w}_{km}^1 | 0.1, 0.3\right) \tag{20}$$

A.2 Model priors

We model the prior distribution of the set of parameters using

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

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

terms by:

$$
\begin{aligned}
p(\boldsymbol{\alpha}^1) &= \prod_{k=1}^{N_1} \text{InverseGamma}(\alpha_k^1|1,1), \\
p(\boldsymbol{\mu}^y) &= \prod_{k=1}^{D_y} \mathcal{N}(\mu_k^y|0,5), \\
p(\boldsymbol{\mu}^a) &= \prod_{p=1}^{P} \mathcal{N}(\mu_p^a|0,5), \\
p(\mathbf{b}^a) &= \prod_{n=1}^{N_{ac}} \prod_{p=1}^{P} \mathcal{N}(b_{np}^a|0,5), \\
p(\sigma^y) &= \log \mathcal{N}(\sigma^y|0,1), \qquad \text{(if demand model is a regression)}, \\
p(\boldsymbol{\sigma}^a) &= \prod_{p=1}^{P_c} \log \mathcal{N}(\sigma_p^a|0,1)
\end{aligned} \tag{21}
$$

B Rotation of traits

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

First, we define the vectors $\boldsymbol{\beta}_i^{ya} = \begin{pmatrix} \boldsymbol{\beta}_i^y \\ \boldsymbol{\beta}_i^a \end{pmatrix}$, and $\boldsymbol{\mu}^{ya} = \begin{pmatrix} \boldsymbol{\mu}^y \\ \boldsymbol{\mu}^a \end{pmatrix}$ of length $(D_y + P)$, and the matrix $\mathbf{W}^{ya} = \begin{bmatrix} \mathbf{W}^y \\ \mathbf{W}^a \end{bmatrix}$ of size $(D_y + P) \times N_1$. Second, we rewrite (3) and (4) as:

$$\boldsymbol{\beta}_i^{ya} = \boldsymbol{\mu}^{ya} + \mathbf{W}^{ya} \cdot z_i^1. \tag{22}$$

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

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

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

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

1. **Fix labels**:

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

   This construction allow us to choose the labels of the traits that explain the most variance in decreasing order, similarly as in Bayesian PCA (Bishop 2006), which are unlikely to switch across posterior samples for well behaved samples of the product $\mathbf{W}_d^{ya} \cdot z_{id}^1$, which is identified

in our model. However, the sign of the traits are not uniquely determined by the SVD. Note that if me multiply by -1 a column of $\mathbf{U}_d$, and we also multiply by -1 the same corresponding row of $\mathbf{V}'_d$, then we would also obtain a valid SVD.[28]

2. **Fix signs**:

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

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

$$n(k) = \underset{n=1,\ldots,(D_y+P)}{\arg\max} \left\{ \frac{1}{D} \sum_{d=1}^{D} \text{abs}\left(\widehat{w}_{nk,d}^{ya}\right) \right\} \tag{23}$$

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

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

   Finally, we construct our sample by:

$$\widetilde{\mathbf{W}}_d^{ya} = \widehat{\mathbf{W}}_d^{ya} \cdot \widetilde{I}_d \qquad \forall d$$

$$\widetilde{z}_{id}^1 = \widetilde{I}_d \cdot \widehat{z}_{id}^1 \qquad \forall i, d$$

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

C Algorithm for newly-acquired customers

With reference to (8), once we have estimated the full model using the calibration data, we can form first impressions of newly acquired customers using the following procedure:

Algorithm 1 Forming first impressions

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

D Details about the simulation analyses

In this appendix we provide further details about the simulation exercise described in Section 4

D.1 Simulated values

D.1.1 Generate acquisition parameters We simulate $\boldsymbol{\beta}_i^a$ using B_1, B_{2p}, and σ_p^{ba}; following (9). We use $\sigma^{ba} = 0.1$. For B_{1p} and B_{2p}, Table D.1 presents the values used in the analysis.

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

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

D.1.2 Generate demand parameters We simulate $\boldsymbol{\beta}_i^y$ using Ω_k, and σ_p^{by}; following (10). We use $\sigma^{by} = 0.1$. For Ω_k, Tables D.2, D.3, and D.4 present the values used in the analysis for the Linear, Quadratic/Interaction, and Positive part scenarios, respectively. We simulate Gaussian values for ω_k^1 and Ω_k^2. For the positive part scenario, we simulate values only for two variables, whereas we set the effect to zero

Table D.2: Simulated values for $\boldsymbol{\omega}_k^1$ in the Linear scenario

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

Table D.3: Simulated values for $\boldsymbol{\omega}_k^1$ and Ω_k^2 in the Quadratic/Interaction scenario

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

Table D.4: Simulated values for $\boldsymbol{\omega}_k^1$ in the Positive part scenario

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

D.1.3 Generate acquisition and demand behaviors

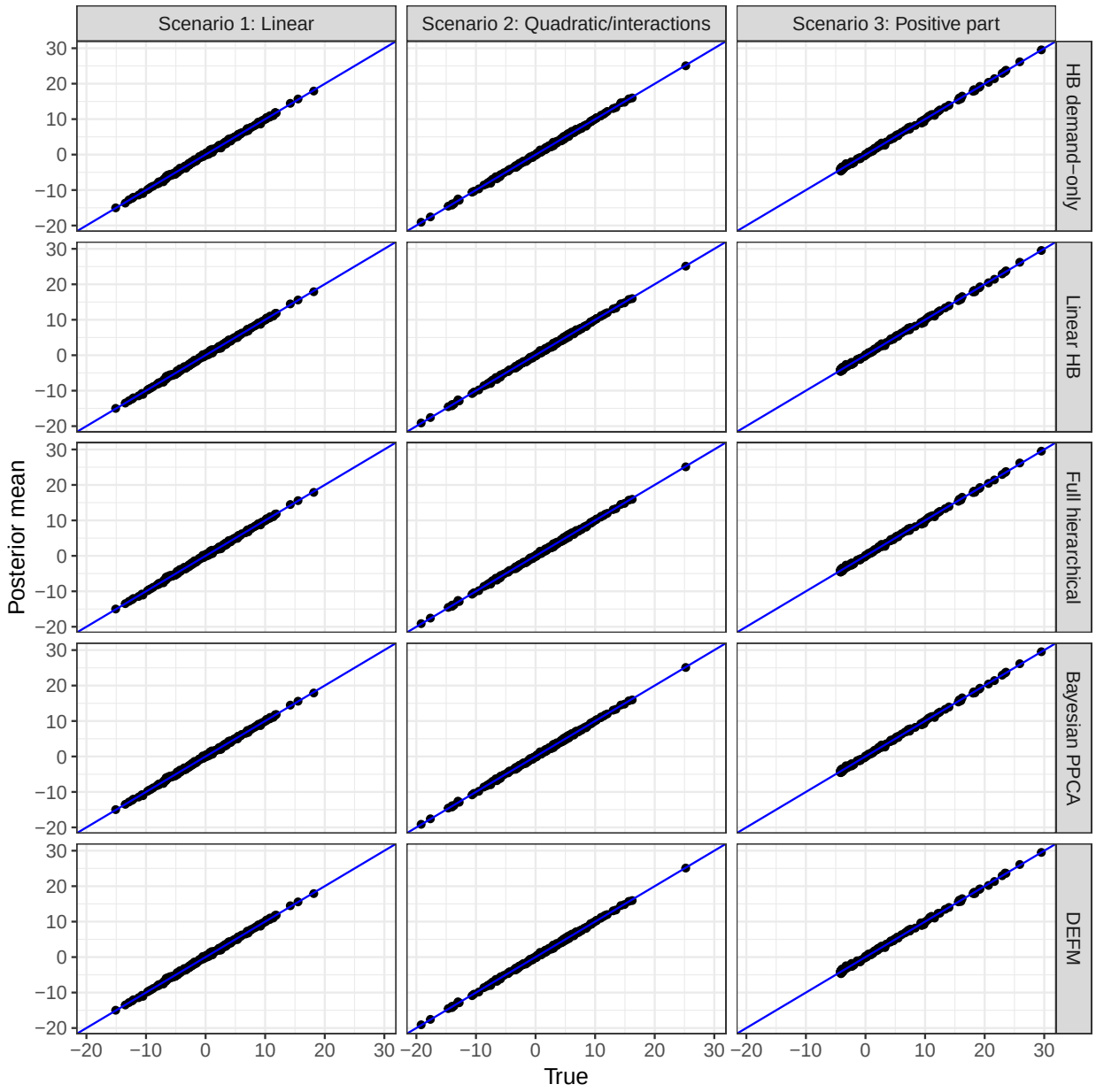
We simulate A_i using $\mathbf{x}_{m(i)}^a$, $\mathbf{b}^a$, and σ_p^a; following (14). We use $\sigma^a = 0.5$, we draw $\mathbf{x}_{m(i)}^a \sim \mathcal{N}(0, 1)$, and $\mathbf{b}^a \sim \mathcal{N}(0, 2)$.

We simulate y_{it} using $\mathbf{x}_{it}^y$, and σ^y; following (15). We use $\sigma^y = 0.5$, and we draw $\mathbf{x}_{it}^y \sim$ Bernoulli(0.5).

D.2 In-sample predictions

We show that all models are equally capable of capturing individual level demand parameters of customers in the calibration sample. We plot in Figure D.1 the individual posterior mean versus the true values across individuals, for each estimated model and each scenario (linear, quadratic/interaction, and positive part). All values lie in the 45 degree line for all models and scenarios.

Figure D.1: Individual posterior mean vs. true intercepts of the demand model. Each dot is a customer from the calibration sample. In blue, the 45 degree line.



D.3 Results for the covariates

Table D.5: Demand parameter recovery for out of sample customers

Model	Intercept		Covariate 1		Covariate 2	
	R-squared	RMSE	R-squared	RMSE	R-squared	RMSE
Linear						
HB demand-only	0.001	6.703	0.005	2.562	0.004	4.944
Linear	0.988	0.734	0.986	0.303	0.988	0.551
Full hierarchical	0.988	0.735	0.986	0.303	0.988	0.550
Bayesian PPCA	0.988	0.736	0.986	0.301	0.988	0.549
DEFM	0.988	0.739	0.986	0.302	0.988	0.551
Quadratic/Interaction						
HB demand-only	0.020	7.624	0.004	4.589	0.021	3.443
Linear	0.711	4.113	0.258	3.969	0.989	0.355
Full hierarchical	0.704	4.164	0.258	3.970	0.989	0.354
Bayesian PPCA	0.706	4.484	0.245	4.364	0.989	0.355
DEFM	0.891	2.616	0.624	2.898	0.989	0.358
Positive part						
HB demand-only	0.007	8.514	0.001	4.604	0.022	6.734
Linear	0.783	4.056	0.736	2.363	0.988	0.731
Full hierarchical	0.781	4.091	0.733	2.378	0.989	0.727
Bayesian PPCA	0.780	4.329	0.738	2.752	0.989	0.726
DEFM	0.872	3.306	0.637	2.842	0.987	0.769

D.4 Relationship between variables

We show in Table D.6 the posterior mean of the rotated weight traits on demand parameters and acquisition parameters, for the Linear scenario. The first two traits capture most of the variance across individuals for demand and acquisition parameters, while the other traits capture residual variance. First, trait 1 captures the correlation among acquisition variables 1 through 3, whereas trait 2 captures the correlation of acquisition variables 4 through 6. Second, both traits capture

relationships with demand: trait 1 is negatively correlated with intercept and positively correlated with both covariates, whereas trait 2 is negatively correlated with intercept and covariate 2 (effect on covariate 1 is not significantly different from zero).

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

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

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

Now, we are interested in comparing these insights with the true values used for the simulation, specifically how these estimated traits relate to the true factors in the data generation process. In the data generation process, demand parameters are generated from acquisition parameters. Instead, the DEFM gives us the overall correlation of the traits with demand parameters, and not the one-to-one relationships between acquisition variables and demand parameters. Therefore, in order to assess whether our model can capture the essence of the insights the "true" effect of factors f_{i1} and f_{i2} on acquisition parameters and demand parameters in Table D.7. For the acquisition parameters, these true effects are B_{1p} and B_{2p} from (9) (whose values are shown in Table D.1). For the demand parameters, these effects can be obtained by replacing (9) in (10), which reduces to $\boldsymbol{\omega}_k^{1\prime} B_1$ and $\boldsymbol{\omega}_k^{1\prime} B_2$ for the effects of factors 1 and 2, respectively.

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

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

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

D.5 Model "at scale"

We show that models that incorporate all interactions fail to recover demand preferences when the number of acquisition variables is large, whereas the DEFM model can accurately infer these non-linear relationships. We maintain a similar simulation structure, where acquisition parameters are driven by factors, but instead we now have 5 factors and 60 acqusition behaviors, where acquisition behavior is driven by one and only one factor, and each factor generates 12 acquisition parameters. We start by describing the simulation details and their differences to the main analysis in Section 4.1. Then, we describe the additional estimated models, specifically those that include interactions. Finally, similarly as in Section 4.2, we show the models' ability to infer demand parameters for out of sample customers.

D.5.1 Simulation details We assume there are 3 demand parameters (intercept and two covariates) and 60 acquisition parameters, for 60 acquisition behaviors. We generate these acquisition parameters as being highly correlated among each other by assuming these parameters are driven by one of five factors $f_{i1}, \ldots, f_{i5}$. Similarly as in Equation (9), we generate acquisition parameters by:

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

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

The rest of the simulation design is identical as the simulation in Section 4.1, with a different set of parameters Ω. In order to incorporate noise and to allow for different acquisition parameters to inform demand parameters, we relate demand parameters only to a subset of acquisition parameters. Specifically, we choose Ω such that demand parameters are only affected by acquisition parameters from three out of the five factors. We achieve this by setting to zero Ω values for the remaining acquisition parameters. The intercept is a function of the acquisition parameters from factors 1, 2 and 3 (i.e., $\Omega_{1p} = 0,\ \forall p = 37, \ldots, 60$). Covariate 1 is a function of the acquisition parameters from factors 1, 2 and 4 (i.e., $\Omega_{2p} = 0,\ \forall p = 25, \ldots, 36, 49, \ldots, 60$). Covariate 2 is a function of the acquisition parameters from factors 2, 3 and 4 (i.e., $\Omega_{3p} = 0,\ \forall p = 1, \ldots, 12, 49, \ldots, 60$). Similarly as in the main simulation analysis, Covariate 2 is always a linear function of acquisition parameters for all scenarios. The values we use for Ω are specific to each scenario:

- **Linear**: Following (11), we define $\omega^1_{kp} \sim \mathcal{N}(0, 2)$ for all non-zero ω^1_{kp}.

- **Quadratic/Interaction**: Following (12), we define $\omega^1_{kp} \sim \mathcal{N}(0, 2)$ for all non-zero ω^1_{kp}; and $\Omega^2_{kpp'} \sim \mathcal{N}(0, 1)$ for all non-zero $\Omega^2_{kpp'}$.
- **Positive part**: To avoid attenuating the effect of the non-linear function by combining a large number of non-linear functions of correlated acquisition parameters, we fix the effect to the intercept and the first covariate as a function of only one acquisition parameter from each of the three factors that determine that demand parameter. Following (13), we define $\omega^1_{3p} \sim \mathcal{N}(0, 2)$ for all non-zero ω^1_{3p}, and:

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

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

D.5.2 Estimated models In addition to all models described in Section 4.1.2, we estimate a Linear HB model where we include all interactions of acquisition parameters,

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

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

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

D.5.3 Results We estimate all models except the Full hierarchical model, which is computationally unstable given that now there are 60 acquisition variables, and therefore we need a 63×63 covariance matrix. Note that in theory, and in practice as we showed in Section 4.2, the full hierarchical model is equivalent to a Linear HB model. Therefore, removing this model from the analysis does not bias our benchmark.

We show in Table D.8 the out of sample prediction of intercept, and the two covariates under all three scenarios for all models. We replicate the main results from Section 4.2. Both the Linear HB and Bayesian PCA models perform well in the Linear scenario. The DEFM performs as good as these models in the Linear scenario, and outperforms these linear models in the Quadratic/Interaction and the Positive part scenarios. More importantly, both models that include all interactions, Linear and Lasso, do not perform well in any scenario.

Table D.8: Model at scale results

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

E Further results from empirical application analyses

We now present the parameter estimates for DEFM demand model. Table E.9 shows the population mean and population variance of each of the demand parameters. Customers in the sample have a low propensity to transact in average ($\beta^y = -3.186$). Email marketing communications have a positive average impact on purchase ($\beta^y = 0.122$), whereas direct marketing communication and product introduction effects are non significant. Finally, customers return to transact more on holiday periods ($\beta^y = 0.324$).

Table E.9: Parameter estimates of Deep exponential family model

Demand parameter		**Posterior statistics** Post. mean	Post. sd	PCI 2.5%	PCI 97.5%
Intercept	Pop. mean	−3.186	0.044	−3.268	−3.107
	Pop. variance	0.700	0.120	0.485	0.942
Email	Pop. mean	0.122	0.030	0.062	0.181
	Pop. variance	0.006	0.006	0.000	0.022
DM	Pop. mean	0.065	0.043	−0.021	0.140
	Pop. variance	0.046	0.029	0.008	0.118
Product introductions	Pop. mean	0.042	0.057	−0.068	0.152
	Pop. variance	0.023	0.022	0.002	0.083
Season	Pop. mean	0.324	0.088	0.158	0.486
	Pop. variance	0.048	0.047	0.002	0.170