WEB APPENDICES

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

Nicolas Padilla
Eva Ascarza

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

Contents

A Augmenting the acquisition characteristics via product embeddings

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

A.1 Data processing

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

A.2 Word2vec algorithm

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

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

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

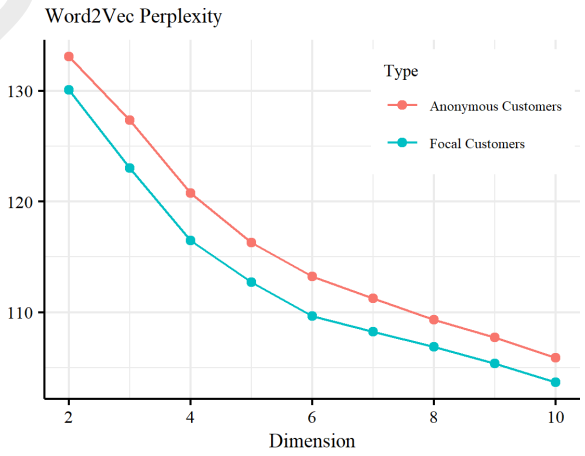
$$L = \sum_{T_i \in T} \sum_{q_i \in Q} \sum_{1<=j<=M, j \neq i} \log P(q_j | q_i), \tag{A.12}$$

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

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

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

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



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

A.3 Interpreting the product dimensions

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

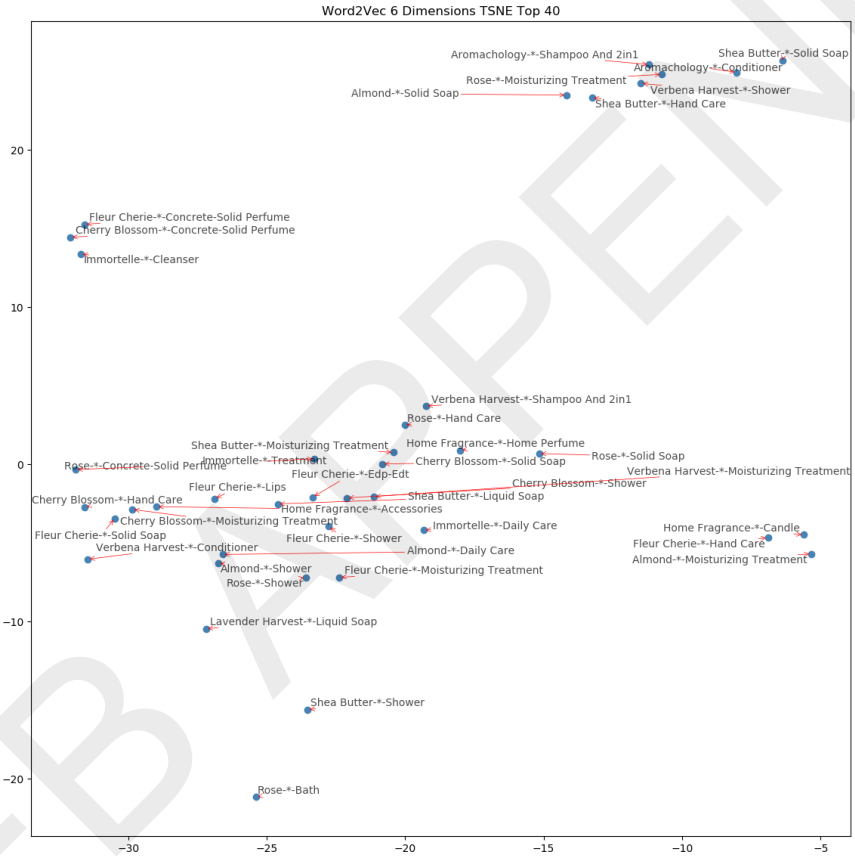
Table A.1: Top 5 products per dimension of the product embeddings.

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

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

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

Figure A.2: Visual representation of the product embeddings



A.4 Product mapping for first purchase data

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

B Brief description of DEFs

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

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

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

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

C Model priors and automatic relevance determination component

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

C.1 Automatic relevance determination

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

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

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

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

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

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

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

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

distribution:

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

C.2 Model priors

We model the prior distribution of the set of parameters using

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

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

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

D Further details about the simulation analyses

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

D.1 Simulation design

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

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

D.2 Data generation process

Generate individual-level parameters

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

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

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

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

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

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

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

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

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

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

- Scenario 1: Linear

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

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

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

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

- Scenario 2: Quadratic/interactions

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

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

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

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

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

- Scenario 3: Positive part

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

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

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

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

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

Simulate individual-level behaviors

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

a Gaussian distribution for all behaviors,

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

$$y_{it} \sim N(\mathbf{x}^{y\,\prime}_{it} \cdot \boldsymbol{\beta}^y_i, \sigma^y), \qquad t = 1, \ldots, 20. \tag{D.24}$$

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

D.3 Estimated models

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

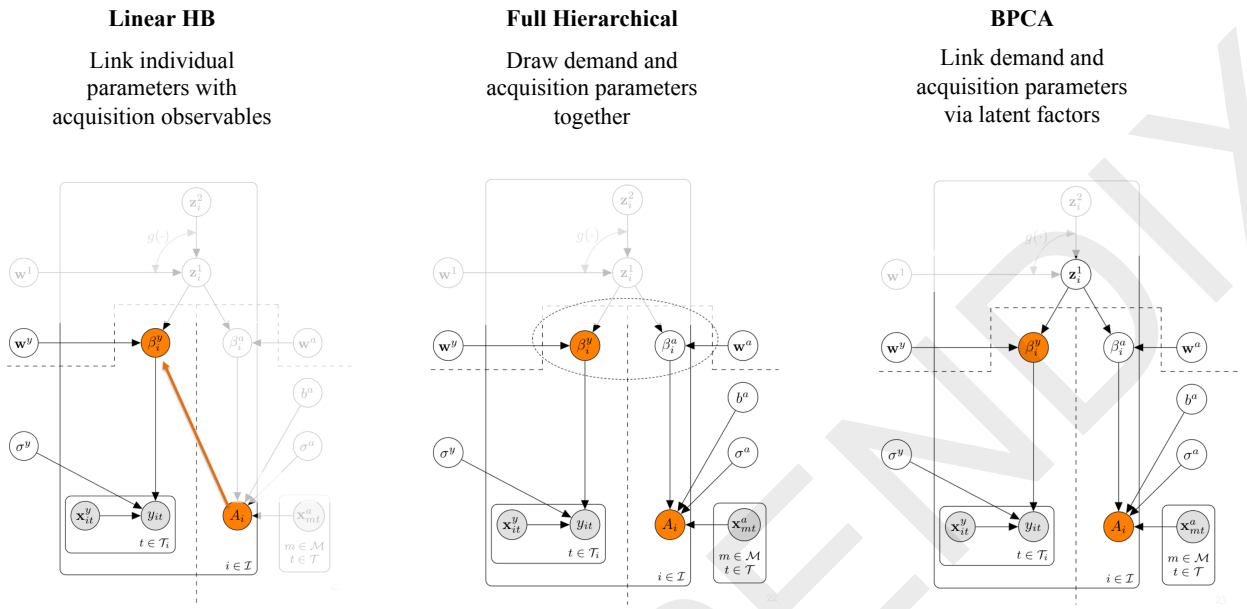
Note that like the linear HB and full hierarchical specifications, the PCA also imposes a linear relationship between $\boldsymbol{\beta}_i^y$ and $\boldsymbol{\beta}_i^a$. However this approach is different from those because it allows for data dimensionality reduction via the latent factors. Similarly,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

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

D.4 Assessing model performance

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

Figure D.3: Visualization of the benchmark models



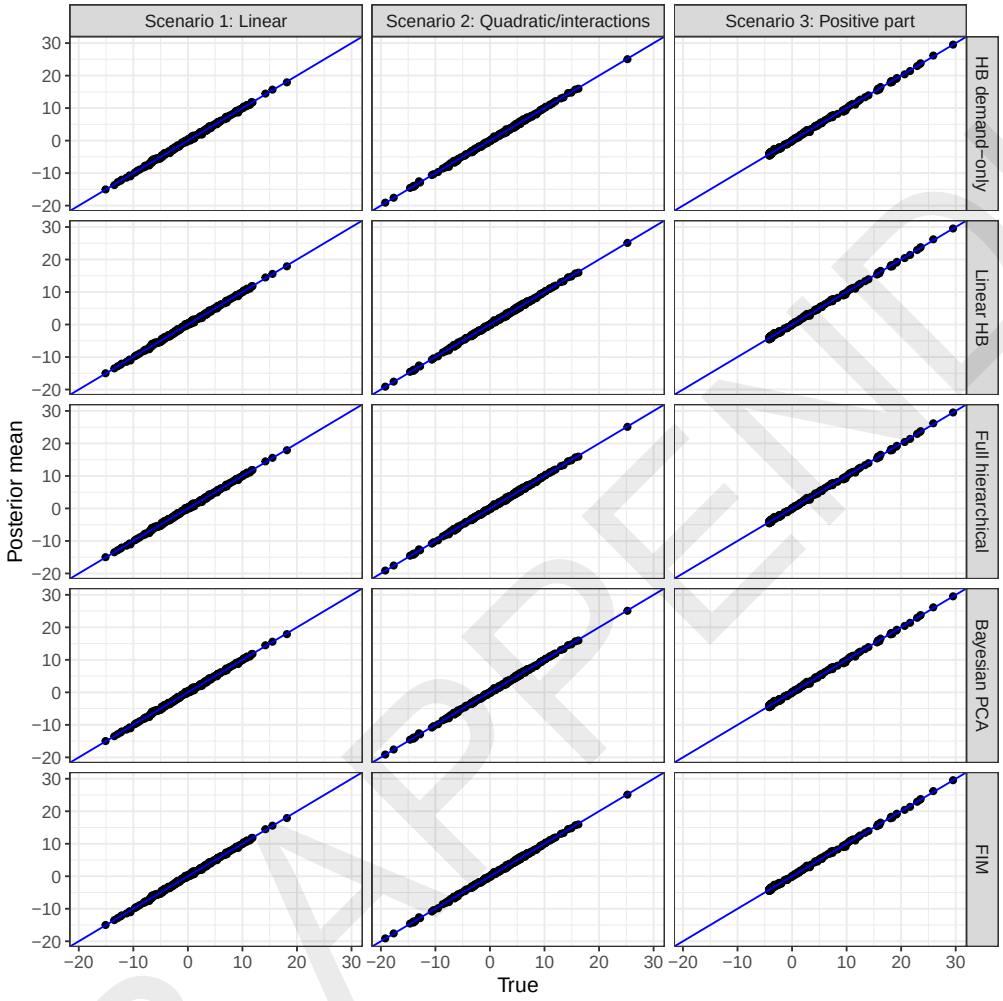
individual-level parameters for customers in the calibration sample. In particular, we confirm that the in-sample predictions for $\boldsymbol{\beta}_i^y$ are almost perfect for all model specifications and for all scenarios (see Figure D.4 for the in-sample predictions). In other words, all models are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

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

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

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

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



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

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

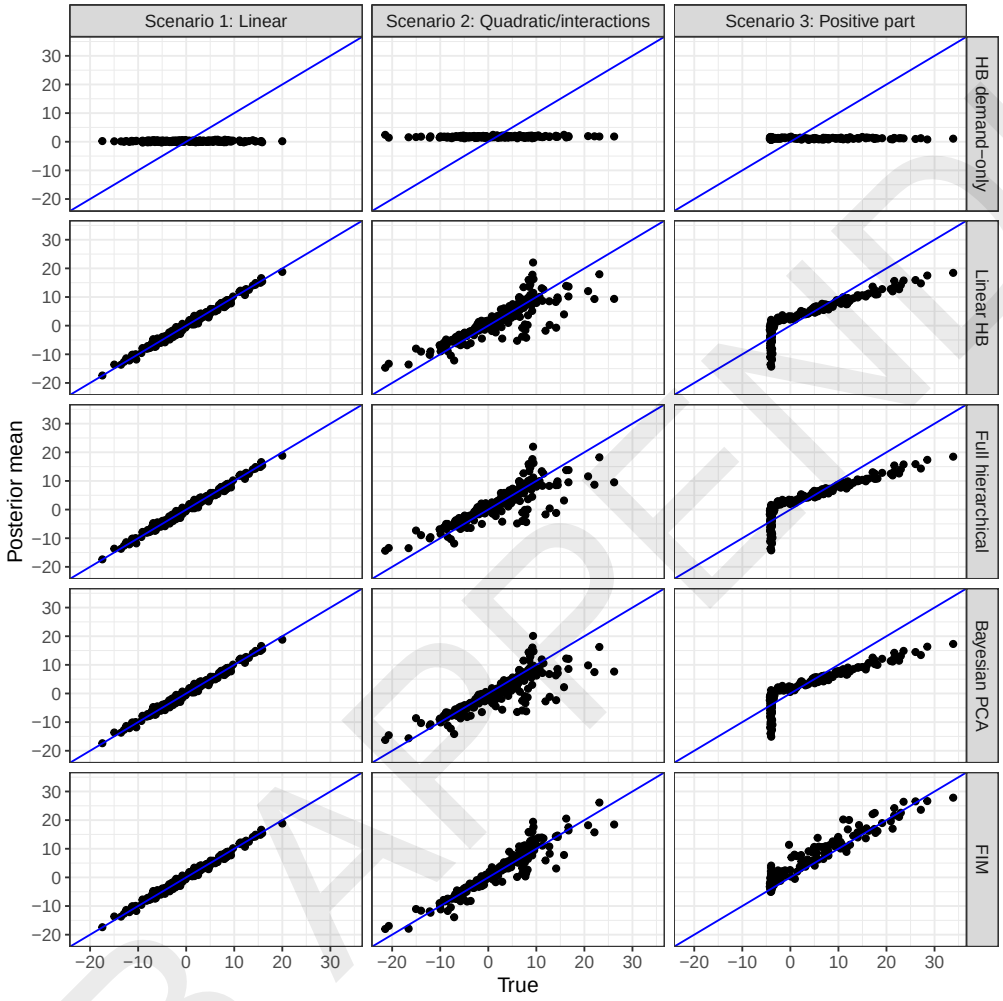
[2]For brevity's sake, we present the results for one parameter of the demand model (the intercept), but the results hold for all other parameters as well.

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

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

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

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



D.5 Interpreting the model parameters and results

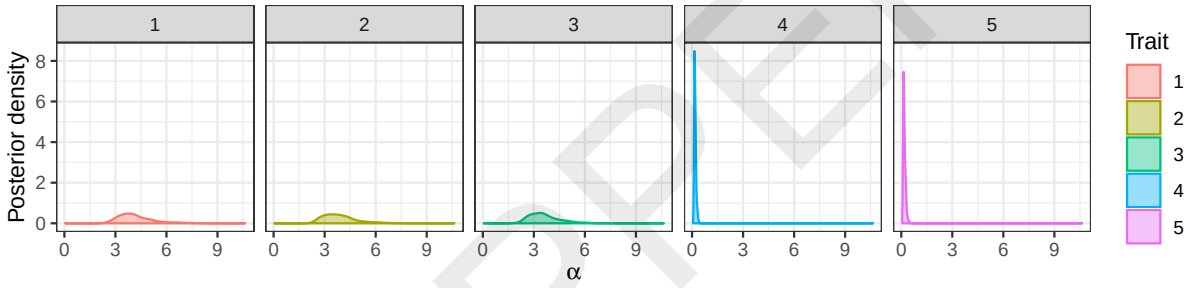
To get a better sense of what the model is doing and what its parameters capture, we explore in detail the model estimates and compare those with the parameters used to simulate the data. We do so for the linear case, as it is the easiest to interpret the relationships among variables. For this particular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

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

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

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

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



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

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

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

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

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

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

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

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

D.6 Why is the model giving superior performance?

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

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

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

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

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

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

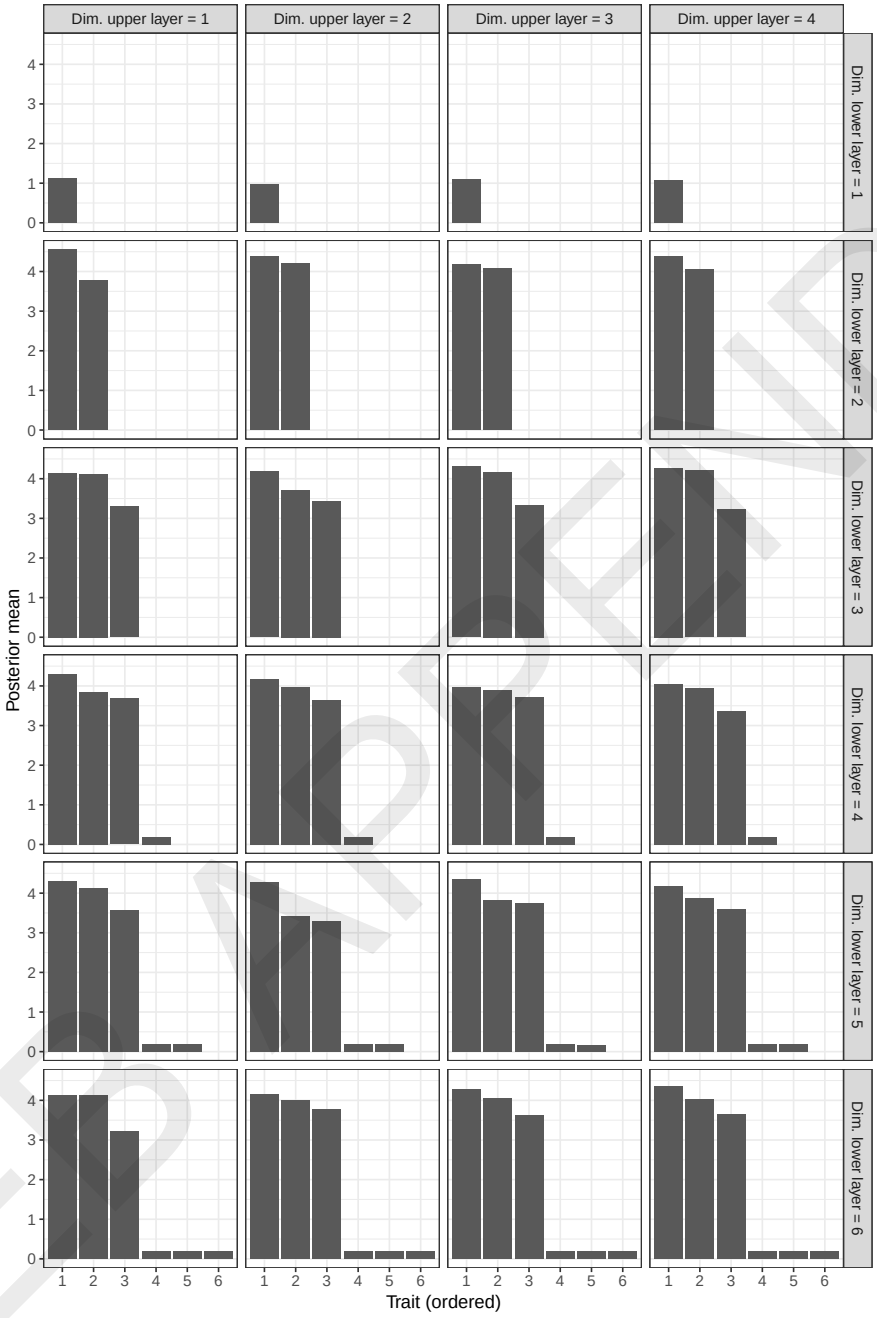
D.7 Exploring the number of dimensions per layer

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

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

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

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



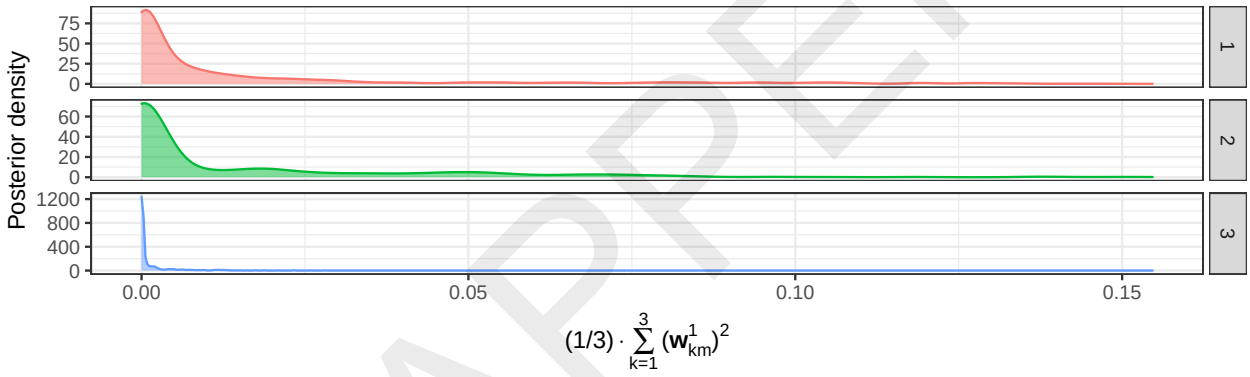
In contrast to the usefulness of α to detect relevant lower traits, our model does not have an analogous parameter to explore how many upper level traits are enough to capture the relevant non-linear interrelations. Instead, each component of the upper weight $\mathbf{W}^1$, $\mathbf{w}^1_{km}$ (for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

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

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

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

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

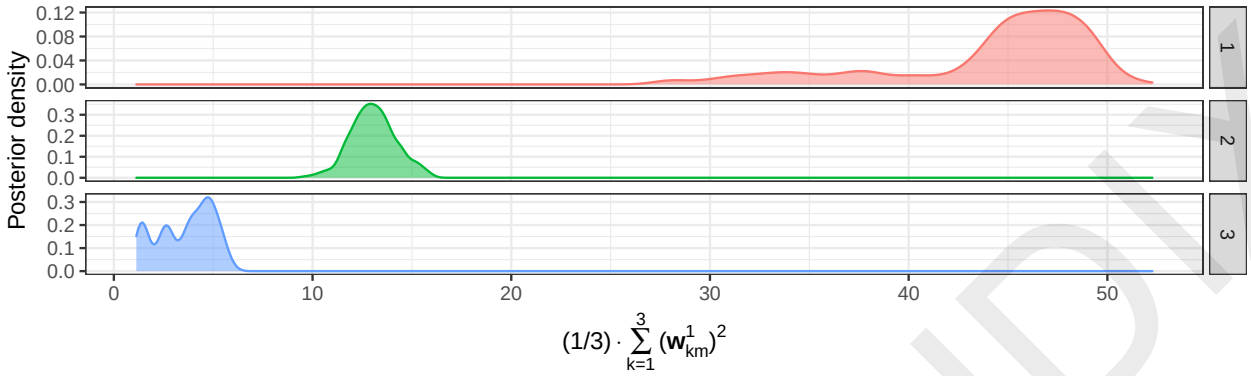


More interestingly, we further explore this quantity using a scenario in which the model requires to capture non-linear relationships, such as the one with Interactions. Figure D.9 shows the posterior of pseudo-α^1 for two FIM specifications with different values of N_2. First, Figure D.9a clearly shows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

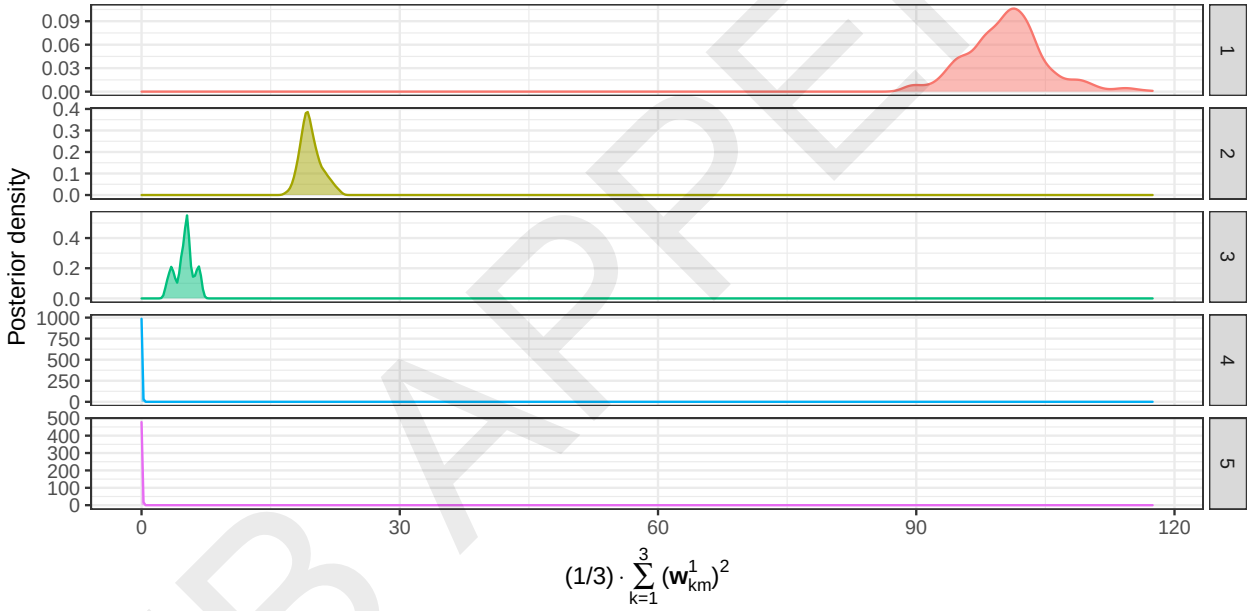
Finally, we leverage the results of multiple estimated FIM specifications over the Interactions scenario and show that once the FIM specification contains the dimensions "needed" by the data, the performance of the model remains the same even if we add dimensions to the DEF component.

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

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



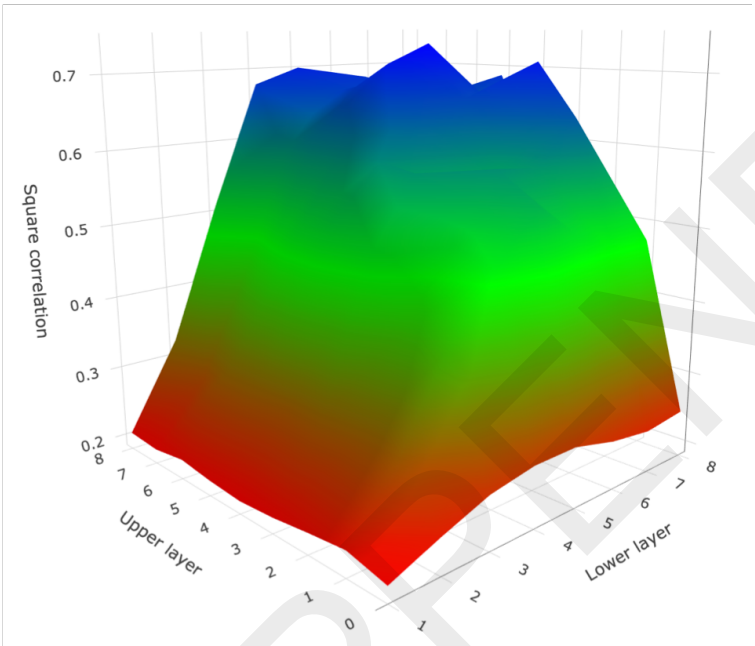
(b) FIM ($N_1 = 5$, $N_2 = 5$).



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

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

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



D.8 Model performance "at scale"

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table D.9: Model at scale results

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

E Rotation of traits

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

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

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

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

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

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

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

1. **Fix labels**:

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

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

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

2. **Fix signs**:

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

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

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

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

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

   Finally, we construct our sample by:

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

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

F Algorithm for newly-acquired customers

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

Algorithm 1 Forming first impressions

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

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

G Empirical application: Additional results

G.1 Possible sources of endogeneity in the model components

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

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

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

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

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

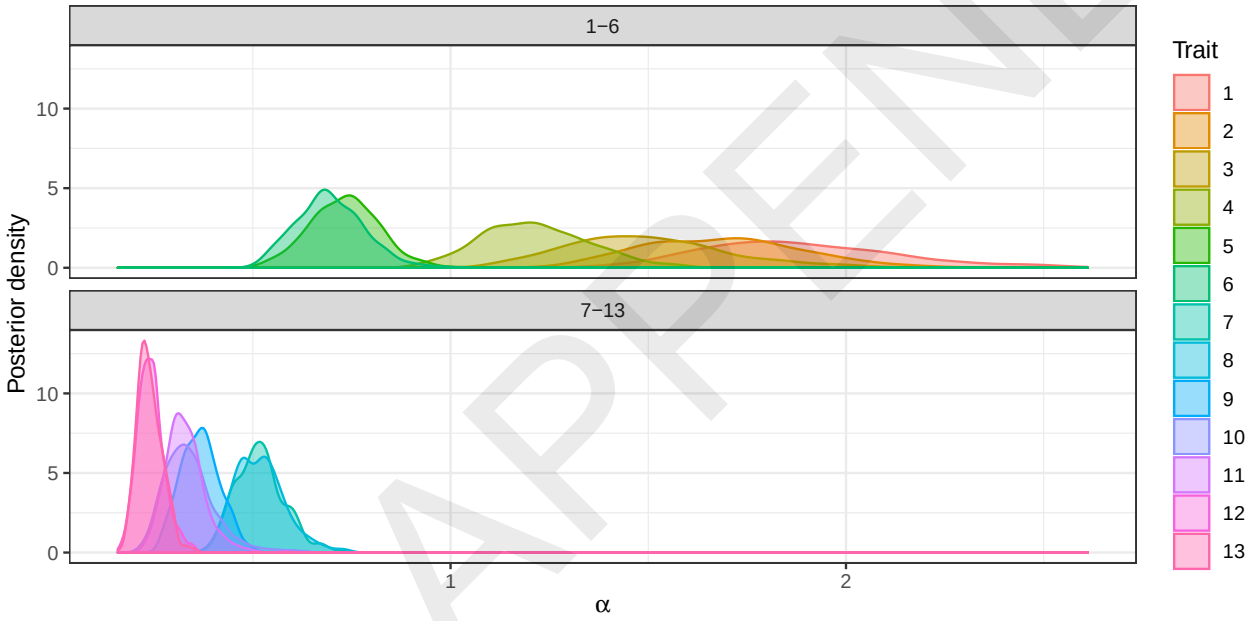
To conclude, given the business nature of our application, the rich variation in our data (Section 5.1.2, Marketing actions), and our model specification, we argue that the potential endogenous nature of the marketing actions is not a main concern in this research. Nevertheless, in situations where these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

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

G.2 Exploring the latent factors

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

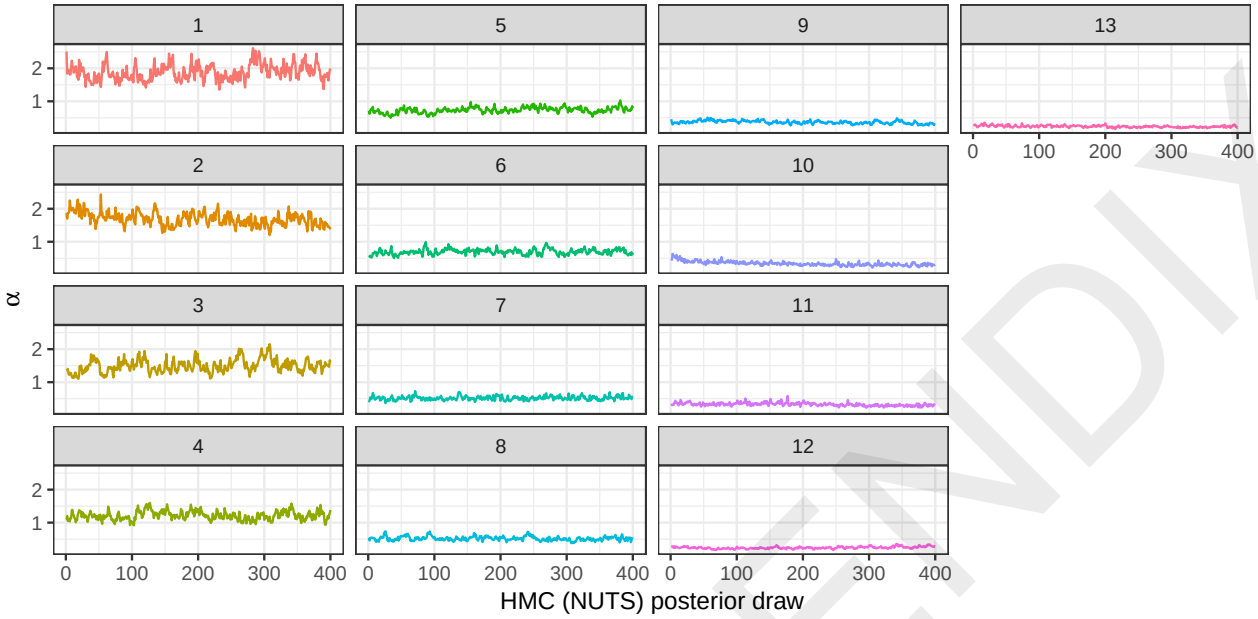
Figure G.11: Posterior distribution of $\boldsymbol{\alpha}$.



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

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

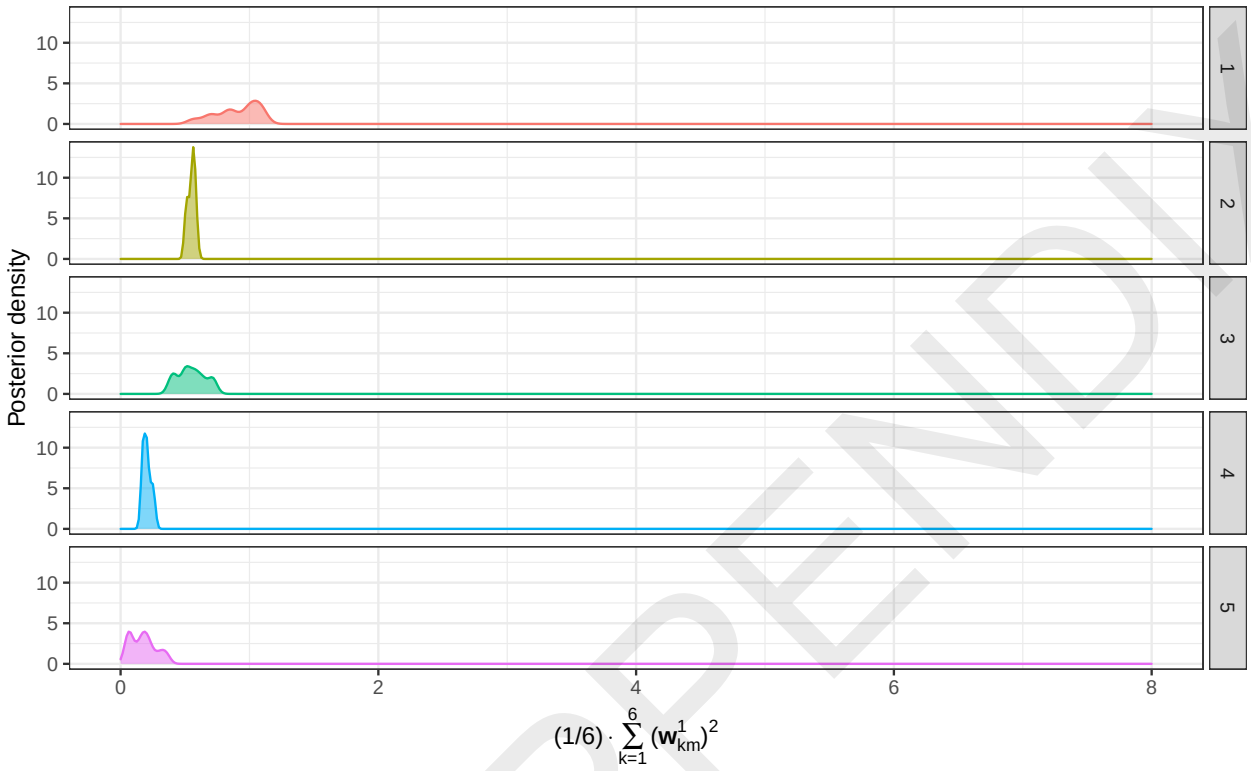
Figure G.12: Convergence of $\boldsymbol{\alpha}$.



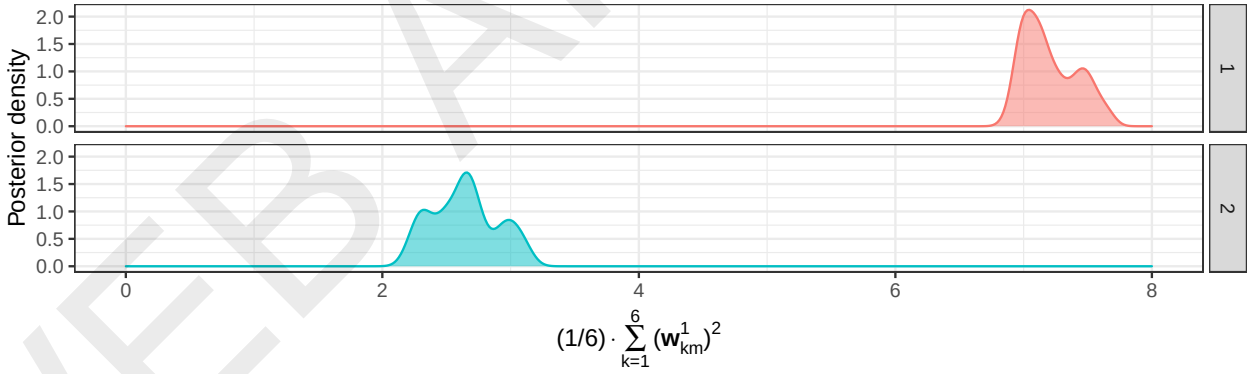
non-linear interrelations present in the data. For robustness, we estimate another FIM specification with $N_2 = 2$ instead, and we find that all upper traits are relevant, suggesting that $N_2 = 2$ may not be enough to capture the non-linear relationships present in the data.

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

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



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



G.3 Latent attrition benchmarks models

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

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

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

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

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

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

Table G.10: Latent attrition benchmarks models.

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

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

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

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

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

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

G.5 Interpreting the latent traits

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

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

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

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

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

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

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

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

G.6 FIM predictive accuracy using in-sample customers

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

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

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

G.7 Population distribution and individual-level posterior distributions

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

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