Personalization with HART*

Thomas Wiemann†

July 1, 2025

Please click here for the latest version.

Abstract

Firms personalize prices, advertising, product design, and more to find and serve their—often highly heterogeneous—consumers. When personalizing to *known* consumers, these marketing decisions can be informed by past choice behavior. However, personalization must rely on observed characteristics for *new* consumers with limited or no purchase histories. I propose Bayesian *hierarchical* additive regression trees (HART) to define optimal marketing decisions that adapt to the firm's familiarity with the consumer. HART combines the strengths of supervised machine learning and hierarchical Bayesian models in one framework: First, it flexibly leverages potentially many observed characteristics to personalize to new consumers. Second, it optimally adapts to the consumer's specific preferences as their choices are recorded over time. I develop an efficient Metropolis-within-Gibbs sampler for fully Bayesian inference and apply it in two discrete choice applications. Using data from a canonical conjoint study, I illustrate how HART discovers marketing opportunities for product design in new markets. In a CPG scanner data application, HART leverages observed characteristics to improve out-of-sample choice prediction by 60% for new consumers, and raises profits by 13% and 2% compared to conventional personalization approaches for new and known consumers, respectively.

Keywords. Cold start problem, optimal targeting, demand estimation, hierarchical Bayes, Bayesian additive regression trees, probabilistic machine learning

*I thank Stéphane Bonhomme, Giovanni Compiani, Max Farrell, Christian Hansen, Ali Hortaçsu, Sid Kankanala, Sanjog Misra, and Alexander Torgovitsky for valuable comments and discussions, along with participants at the University of Chicago Econometrics and IO advising groups, and the Booth Quantitative Marketing brown bag. Researcher's own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

†University of Chicago, wiemann@uchicago.edu

1 Introduction

Consumers are highly heterogeneous both in their observed characteristics and their unobserved preferences. Understanding and catering to these differences has been a core marketing challenge for several decades (e.g., Murthi and Sarkar, 2003; Rafieian and Yoganarasimhan, 2023). Canonical examples include targeted product design (Allenby and Ginter, 1995), couponing (Rossi et al., 1996), and e-mail marketing (Ansari and Mela, 2003). Recent personalization examples include churn outreach (Ascarza, 2018), digital advertising (Rafieian and Yoganarasimhan, 2021; Danaher, 2023), promotions (Simester et al., 2020a; Simester et al., 2020b), pricing (Morozov et al., 2021; Dubé and Misra, 2023; Liu, 2023; Smith et al., 2023; Jain et al., 2024), product recommendations (Korganbekova and Zuber, 2023), free trials (Yoganarasimhan et al., 2023), and catalog mailing (Hitsch et al., 2024). In each of these applications, personalization is found to be a valuable device in a firm's success, motivating the large interest in methods for estimating and leveraging consumer heterogeneity.

However, firms seeking to personalize marketing decisions must not only understand heterogeneity among consumers' preferences, they must also make personalization decisions with widely varying *extents of knowledge* about different consumers. For example, Rossi et al. (1996) and Smith et al. (2023) emphasize the value of purchase history for optimal couponing for *known* consumers. Yet, for *new* consumers with no or limited historical demand data, firms have no choice but to exploit other data sources for personalization as in Padilla and Ascarza (2021) and Dubé and Misra (2023). While a firm would ideally leverage all available information for every consumer regardless of their relationship's length, existing methods flexibly leverage either choice behavior or observed characteristics. This complicates personalization efforts in applications with known and new consumers, and is further exacerbated when a consumer's choices accumulate without guidance on the transition between methods.

I address this challenge by proposing a hierarchical Bayesian machine learning approach that flexibly leverages potentially many observed characteristics, adapts to purchase histories of varying lengths, and defines optimal personalization decisions regardless of the firm's familiarity with the consumer. The proposed approach combines the strengths of recent supervised machine learning methods and long-established hierarchical Bayesian models: First, like supervised machine learning methods, new consumers' preferences are predicted with a granular "representative consumer" defined as a flexible function of potentially many observed characteristics. Second, like existing hierarchical Bayesian models, consumer-specific preferences adaptively deviate from those of the representative unit as their choices accumulate. The hierarchical framework also quantifies uncertainty in the firm's knowledge about each consumer's preferences, allowing for

(Bayes) optimal personalization decisions at any point in a consumer's journey. To the best of my knowledge, this is the first Bayesian approach to personalization that flexibly leverages potentially many observed characteristics and purchase histories, addressing an open question highlighted by both Allenby and Rossi (2019) and Dew et al. (2024).[1]

Formally, I propose a new hierarchical nonparametric prior for preference parameters applicable to any generic consumer-level demand model. Motivated by the empirical success of Bayesian Additive Regression Trees (BART) of Chipman et al. (2010) for nonparametric regression and heterogeneous treatment effect estimation (e.g., Hahn et al., 2020; Hill et al., 2020), I model the representative consumer as a (stacked and scaled) sum of many regression trees. For ease of discussion, this novel extension of BART to hierarchical models is dubbed HART (hierarchical additive regression trees). HART flexibly leverages potentially many observed consumer characteristics and thus generalizes existing hierarchical models that exclusively model the representative consumer as a *linear* function of a select *few* characteristics (e.g., Allenby and Ginter, 1995; Rossi et al., 1996; Smith et al., 2023). I further propose Dirichlet HART for settings with sparse dependence on consumer characteristics. Dirichlet HART builds on the prior specification of Linero (2018), inducing exact sparsity and accommodating fully Bayesian variable selection.

A key challenge in high-dimensional Bayesian models such as HART is the computational complexity of sampling from the posterior distribution. This is further complicated by the discreteness of the sum-of-trees representative consumer model which prohibits the application of off-the-shelf samplers such as Hamiltonian Monte Carlo that are based on gradient approximations of the posterior.[2] I resolve these challenges with the development of an efficient Metropolis-within-Gibbs sampler for fully Bayesian inference over consumer-specific preferences, the representative consumer, and the distribution of preferences in the population. This is achieved by combining generic Gibbs samplers for hierarchical models as in Rossi et al. (2009) with adapted steps of the "Bayesian backfitting" algorithm of Chipman et al. (2010). The samples of the proposed Metropolis-within-Gibbs algorithm converge to the exact posterior distribution, thus facilitating optimal personalization decisions that fully account for estimation uncertainty (e.g., Green, 1963; Rossi et al., 1996; Allenby and Rossi, 2019).[3]

I illustrate the HART model with two discrete choice applications. In the canonical conjoint study of Allenby and Ginter (1995) on out-of-state credit card design, HART

[1] Padilla and Ascarza (2021) and Yin et al. (2024) also discuss limitations of existing hierarchical Bayesian models for flexible use of observed characteristics needed for personalization for new consumers.

[2] Hamiltonian Monte Carlo samplers are widely used in probabilistic machine learning applications in marketing, see, e.g., Dew et al. (2024).

[3] An implementation of the Metropolis-within-Gibbs sampler for the HART model is provided in the R package `bayesm.HART`. The package builds on the popular `bayesm` package of Rossi (2023), combining familiar high-level R syntax with an efficient C++ implementation of the MCMC algorithm.

estimates rich observed and unobserved heterogeneity in preferences. Unlike linear hierarchical models, HART discovers nonlinear associations between respondents' demographics and their preferences. This has important implications for counterfactual analysis. For example, while linear hierarchical models predict little advantage in targeting different consumer segments, HART clearly differentiates between segments that are likely to respond to different credit card designs. I further highlight the robustness of Dirichlet HART to high-dimensional observed characteristics with a placebo exercise that appends the data with 100 simulated noise variables. In contrast to results based on the linear approach, counterfactuals based on Dirichlet HART are unaffected as its posterior variable selection approach accurately discards the irrelevant characteristics.

Applying the HART model to CPG scanner data on mayonnaise purchases also generates new insights about the association between consumer characteristics and their preferences, and highlights HART's marketing value for personalized coupons. HART improves out-of-sample choice prediction for new consumers by 60% with standard low-dimensional demographics compared to linear hierarchical models. When additional characteristics are included, the linear hierarchical model's prediction error increases while HART's prediction error decreases to an overall improvement of 113%. These improvements stem from both more flexible use of conventional demographics and the ability to leverage more characteristics. For example, Bayesian posterior variable importance measures of the Dirichlet HART model indicate that preference predictions for new consumers load on several characteristics not typically considered in marketing applications, including *owning a dishwasher*. Finally, I assess the marketing value of the proposed models in a counterfactual personalized couponing exercise similar to Smith et al. (2023). Using a double/debiased machine learning estimator to estimate out-of-sample expected counterfactual profits, I show that personalization with HART would substantially improve profits of the focal manufacturer Hellmann's. Compared to existing pricing, personalization with HART results in 40% higher profits for new consumers and 26% higher profits for known consumers. Scaling by the approximate market size of the midwest-US market, this is associated with $1.35 million higher annual profits. Further, personalization with HART improves upon personalization with conventional alternative approaches by 13% and 2% for new and known consumers, respectively. Importantly, the results indicate the importance of flexible use of observed characteristics *and* adjusting to accumulating purchase history, as achieved by the proposed HART models.

Related Literature. This paper draws from and contributes to several strands of literature in marketing and statistics.

First, the paper contributes to the vast literature on personalization in marketing (e.g., Rossi et al., 1996; Ansari and Mela, 2003; Ascarza, 2018; Dubé and Misra, 2023; Rafieian

and Yoganarasimhan, 2023; Smith et al., 2023; Hitsch et al., 2024). Existing research has largely focused on personalization for either known or new consumers. This paper jointly analyzes personalization for consumers with varying purchase history lengths in a coherent framework with both flexible observed and unobserved heterogeneity. I highlight that flexible use of extended consumer characteristics is valuable for targeting, substantially increasing profits in the couponing application. The results further suggest that capturing potential correlations in unobserved heterogeneity is important for effective targeting. Finally, I confirm that observed characteristics available in conventional scanner data panels are complementary to—not a replacement of—historical choice data. For known consumers, leveraging their purchase history is key for successful personalization even when observed characteristics are flexibly incorporated.

Second, HART provides a new approach to the cold start problem, wherein managers are tasked with decision making in environments with little or no historical demand data. The core challenge in this setting is effective extrapolation from existing information to new settings.[4] Recent contributions leverage initial customer behavior to infer shared latent heterogeneity (e.g., Padilla and Ascarza, 2021; Padilla et al., 2024). For example, Padilla and Ascarza (2021) augment a consumer's initial purchase outcome with several outcomes on their initial acquisition. When purchasing and acquisition patterns are related, this provides additional identifying information about a consumer's latent consumer type (see also Ainslie and Rossi (1998)). The proposed HART model instead flexibly leverages potentially many observed consumer characteristics to inform initial personalization. While this supervised learning approach has the benefit of avoiding distributional assumptions imposed on augmented outcomes as in Padilla and Ascarza (2021), HART and outcome-augmentation approaches are complementary. Their combination is a potentially useful avenue for future research.

The idea to use observed characteristics to inform initial predictions in the absence of historical demand data dates back to at least discussions in Lenk and Rao (1990). Allenby and Ginter (1995) leverage consumer characteristics in a hierarchical Bayesian logit model for predicting choices of new conjoint respondents. Yet, subsequent literature has concluded little gain from using demographics (e.g., Rossi et al., 1996; Smith et al., 2023). Results in this paper suggest that this is at least partially due to the inability of existing hierarchical approaches to accommodate rich characteristics in a flexible manner.

The approach of HART to flexibly leverage observed heterogeneity for initial personalization is closely related to recent contributions of Ascarza (2018), Dubé and Misra

[4]The cold start problem is not unique to customer personalization. Managers must also readily make decisions, for example, concerning new products and markets as in Lenk and Rao (1990), Neelamegham and Chintagunta (1999), Neelamegham and Chintagunta (2004). The advantages of HART readily extend to these settings.

(2023), Yoganarasimhan et al. (2023), Hitsch et al. (2024), Yin et al. (2024), and Farrell et al. (2025) on using machine learning to estimate observed heterogeneity.[5] These machine learning approaches offer flexible personalization solutions in modern settings with potentially many observed characteristics (Chintagunta et al., 2016; Bradlow et al., 2017), or where linear functional forms are ill-suited (e.g., using embeddings as in Yin et al. (2024)). The key difference in this paper is that HART embeds the machine learning method within a hierarchical Bayesian framework. This has three key advantages: 1) HART estimates the *distribution* of unobserved heterogeneity, including taste correlations that are often important for marketing counterfactuals (e.g., Allenby and Ginter, 1995). 2) HART's individual-level estimates adapt to a consumer's purchase history, where the learning rate is data-driven and depends on the distribution of unobserved heterogeneity (e.g., Allenby and Rossi, 2019). 3) HART accommodates fully Bayesian inference both on individual-level parameters and flexible preference predictions. Importantly, this provides direct managerial guidance on how to optimally transition from recommendations for a brand new consumer to recommendations for a consumer with accumulating purchase history.

Third, I contribute to the rapidly growing literature on applications of probabilistic machine learning in marketing (Dew et al., 2024). Recent examples include Gaussian process priors (e.g., Dew and Ansari, 2018; Korganbekova and Zuber, 2023; Dew, 2025), deep exponential family components (Padilla and Ascarza, 2021), and Bayesian neural networks (Daviet, 2020). BART developed by Chipman et al. (2010) is widely popular for nonparametric regression applications and heterogeneous treatment effect estimation (e.g., Hill, 2011; Hahn et al., 2020; Hill et al., 2020), including for personalized medical recommendations (Logan et al., 2019). Linero and Yang (2018), Ročková and Van Der Pas (2020), and Jeong and Ročková (2023) characterize conditions for near-optimal posterior concentration rates of variants of BART about complex smooth or discontinuous functions. To the best of my knowledge, this is the first paper to highlight the value of BART in marketing models.

Finally, I contribute to the statistical literature on methodological extensions of BART (Chakraborty, 2016; Linero, 2018; Murray, 2021; Um et al., 2023; Esser et al., 2025; Deshpande et al., 2024). Closely related are Murray (2021) who develops a BART prior for nonparametric logistic and count regression models, and Deshpande et al. (2024) who develop a BART prior for linear varying coefficient models. I complement these recent developments with a hierarchical BART prior on individual-level parameters of generic unit-level likelihoods. The proposed hierarchical prior applies in generic panel applications, including choice and count regression models, frequently considered in marketing

[5]See also Rafieian and Yoganarasimhan (2023) for a recent review.

applications.

The key technical contribution of this paper is the development of a practical Gibbs sampler for posterior inference about the HART parameters. Due to its discrete nature, off-the-shelf gradient-based samplers typically employed for, e.g., Gaussian process models are not applicable to HART. I discuss a general data augmentation and variable transformation strategy that greatly simplifies HART sampling and allows for application of the Bayesian backfitting algorithm of Chipman et al. (1998) and Chipman et al. (2010). Importantly, the proposed sampler is modular and can straightforwardly be combined with existing Gibbs samplers for hierarchical models as in Rossi et al. (2009). Further, the strategies employed for application of Bayesian backfitting are readily amenable to alternative algorithms for supervised Bayesian machine learning methods such as the Bayesian lasso (Park and Casella, 2008), paving a path for future research on alternative hierarchical priors for flexible use of potentially many observed characteristics.

Outline. The rest of the paper proceeds as follows. Section 2 introduces HART for a generic unit-level likelihood model. Section 3 develops the corresponding Metropolis-within-Gibbs sampler. Section 4 then discusses the HART logit model as the leading example. Sections 5 and 6 illustrate the HART logit model in two applications. Section 5 revisits the canonical conjoint study of Allenby and Ginter (1995). Section 6 applies the HART logit model for personalized couponing in a CPG scanner dataset. Section 7 concludes with a discussion.

2 Hierarchical Additive Regression Trees

Hierarchical Bayesian models in marketing conventionally consist of three levels: A unit-level likelihood that models the unit's historical demand data conditional on their consumer-specific parameters, a first-stage prior that associates the consumer-specific parameters with their observed characteristics, and a second-stage prior that restricts these population-level associations (e.g., Rossi and Allenby, 2003). This section introduces hierarchical additive regression trees (HART) as a flexible prior specification that can be used jointly with any unit-level likelihood model. After briefly introducing the generic setting in Section 2.1, Sections 2.2 and 2.3 define the first and second-stage priors, respectively. Subsection 2.4 states the joint posterior distribution with a generic unit-level likelihood.

2.1 Generic setup

I consider a generic hierarchical setting as described in, for example, Allenby and Rossi (2019). A manager observes $\{i \in [n]\}$ units, each associated with historical demand data

$\mathcal{D}_i$.[6] The extent of the available demand data can vary substantially across units and for new units there may be an absence of any observed history. Common examples of units i in marketing applications are consumers or products, with $\mathcal{D}_i$ representing a consumer's purchases or a product's sales across time.

The manager models each unit's demand using a likelihood $L(\mathcal{D}_i|\theta_i)$ where θ_i is a D-dimensional vector of unit-specific parameters. Continuing previous examples, θ_i may represent a consumer's preference parameters (Allenby and Ginter, 1995), a consumer's search propensities (Morozov, 2023), a consumer's receptiveness to advertising (Zantedeschi et al., 2017), or a product's expected demand (Neelamegham and Chintagunta, 1999).

In addition, the manager observes a potentially high-dimensional vector of time-invariant characteristics Z_i for each unit i, including for new units.[7]

For units with historical demand data, the unit-level likelihood identifies the parameters θ_i. For new units with no or little available demand data, inferences about θ_i must carefully exploit other sources of information. A Bayesian approach leverages a first-stage prior.

2.2 First-stage prior

A first-stage prior for θ_i shrinks unit-level parameters towards a *representative unit*. Commonly used priors in the marketing literature consist of two components: 1) the definition of the representative unit that determines *where* the unit-level parameters are shrunk to, and 2) a distribution of unobserved heterogeneity that determines *how* the unit-level parameters are shrunken.

There is a large literature on the choice of distributions, including normals, mixtures of normals, and Dirichlet process mixtures (e.g., McCulloch and Rossi, 1994; Allenby and Lenk, 1994; Allenby et al., 1998; Ansari and Mela, 2003).[8] These allow researchers to specify, for example, whether shrinkage of preferences towards the representative unit should be symmetric. Allowing for non-symmetric shrinkage can be important in settings where preferences exhibit multi-modality or high skewness (e.g., Dubé et al., 2010).

In contrast, the definition of the representative unit as a *linear* projection of θ_i on a select few characteristics has remained unchanged since initial applications of hierarchical Bayesian methods for demand estimation (e.g., Lenk and Rao, 1990; Allenby and Ginter, 1995; Rossi et al., 1996; Smith et al., 2023). This linear approach imposes strin-

[6] Bracket notation $[n] \equiv \{1, \ldots, n\}$ denotes the set of positive integers up to n.

[7] Although missing values in Z_i are not explicitly considered in this paper, HART can readily be extended to these settings by modelling the joint distribution of characteristics (e.g., Padilla and Ascarza, 2021).

[8] See Allenby and Rossi (2019) for a recent review, and Rossi (2014) for a textbook treatment.

gent use of even low-dimensional characteristics as the introduction of interactions and non-linearities quickly becomes intractable (Padilla and Ascarza, 2021). Similarly, despite rich observed characteristics often readily available in modern datasets (Chintagunta et al., 2016; Bradlow et al., 2017), the linear specification cannot take advantage of these potentially high-dimensional characteristics. Both Allenby and Rossi (2019) and Dew et al. (2024) thus identify approaches for flexible dependence on potentially many observed characteristics within hierarchical Bayesian models as a potential avenue for future research.

I propose to define the representative unit as a flexible function of potentially many observed characteristics via a *sum-of-trees* model. This first-stage prior is given by

$$\theta_i = \Delta(Z_i) + \varepsilon_i, \tag{1}$$

where ε_i is a mean-zero random variable and $\Delta(Z_i)$ denotes a sum-of-trees model of the observed characteristics Z_i.[9] For ease of exposition, I consider the simple normal model of heterogeneity where $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$, but emphasize that the approach can readily be extended to richer models of heterogeneity.

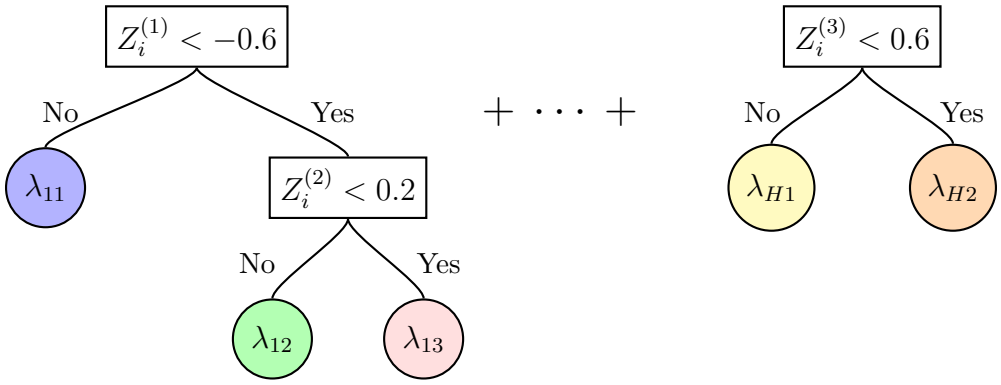
In nonparametric regression of a scalar-valued outcome onto covariates, sum-of-tree models are widely used, for example, as frequentist random forests (Breiman, 2001) or as Bayesian Additive Regression Trees (BART) (Chipman et al., 2010). The sum-of-tree model considered here differs in that the dependent variable is a vector-valued and unobserved parameter θ_i. For illustration, consider first the simplest case where θ_i is scalar-valued, mean-zero, and $\varepsilon_i \sim \mathcal{N}(0, 1)$. If θ_i were observed directly, this corresponds to a simplified setting of the nonparametric regression considered in Chipman et al. (2010). The representative unit is then defined as a sum of H regression trees, where by default $H = 200$. Each tree $h \in [H]$ is characterized by 1) a sequence of nodes that partition the support of Z_i into G_h terminal leaves via splitting rules of the form $Z_i^{(k)} < c$, where c is a constant and $Z_i^{(k)}$ is the k-th component of Z_i, and 2) corresponding coefficients λ_{hg} for all terminal leaves $g \in [G_h]$. Let (R_h, Λ_h) denote the collection of splitting rules and terminal leaf parameters of the hth tree, respectively. A sum-of-trees model, parametrized by the collection of trees and leaf coefficients $\{(R_h, \Lambda_h)\}_{h\in[H]}$, is then given by

$$\delta(Z_i; \{(R_h, \Lambda_h)\}_{h\in[H]}) = \sum_{h=1}^{H} \sum_{g=1}^{G_h} \lambda_{hg} \mathbb{1}\{Z_i \in \mathcal{Z}_{hg}\}, \tag{2}$$

[9] Alternatively, the first-stage prior can be formulated for known transformations of (components of) θ_i. For example, to enforce a negative own-price elasticity in a discrete choice setting, the price coefficient could be reparametrized to $-\exp(\check{\theta}_i^{\text{price}})$. The first-stage prior (1) would then instead be placed on $\check{\theta}_i^{\text{price}}$. Note that this also requires adjustments to the prior variance Σ to account for the log-scale of the price coefficient (Allenby and Rossi, 2019).

where $\mathcal{Z}_{hg}$ denotes the partition corresponding to the gth terminal leaf of the hth tree. The example in Figure 1 illustrates this simple sum-of-trees model.

Figure 1: Example of a Sum-of-Trees Model



Notes: Illustrative example of a sum-of-trees model showing the first and the Hth tree. The splitting rules are based on three observed characteristics $Z_i^{(1)}$, $Z_i^{(2)}$, and $Z_i^{(3)}$. The terminal leaves of the first and Hth tree correspond to coefficients $(\lambda_{11}, \lambda_{12}, \lambda_{13})$, and $(\lambda_{H1}, \lambda_{H2})$, respectively.

Despite their seemingly simple construction as additions of independent step-functions, sum-of-tree models provide accurate approximations to nonlinear (and linear) functions of observed characteristics Z_i. This feature—paired with appropriate statistical tools to estimate the model—has made them a staple of nonparametric regression.[10] It also suggests they are well-suited for granularly-defined representative units in marketing applications.

Of course, this most simple standard normal scalar-valued case bears little resemblance to actual marketing applications where the demand parameters θ_i are nearly always vector-valued. Empirical research also highlights potential correlations between different components of θ_i, for example, consumers' responsiveness to different marketing mix variables and their implications for marketing counterfactuals (e.g., Allenby and Ginter, 1995). In these settings where $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$, I define the representative unit as a scaled vector of $D \equiv \dim(\theta_i)$ mutually distinct sum-of-trees models—that is,

$$\Delta(Z_i; \mu, \Sigma, \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}) \equiv \mu + \Sigma^{1/2} \begin{bmatrix} \delta(Z_i; \{(R_{1h}, \Lambda_{1h})\}_{h\in[H]}) \\ \vdots \\ \delta(Z_i; \{(R_{Dh}, \Lambda_{Dh})\}_{h\in[H]}) \end{bmatrix} \tag{3}$$

where $\Sigma^{1/2}$ is the Cholesky decomposition of Σ and μ is a D-dimensional vector of unconditional means.[11] Note that each sum-of-trees model targets a main component of the

[10] See, for example, Mullainathan and Spiess (2017) and Athey and Imbens (2019) for a discussion of random forests targeted at applied researchers, and Hahn et al. (2020) and Hill et al. (2020) for a discussion of BART.

[11] Note that while the representative consumer $\Delta(\cdot)$ and Σ are identified as the conditional expectation and covariance of θ_i, the unconditional mean μ and the sum-of-tree parameters are not identified. For example, one might add a constant c to μ and subtract $\Sigma^{-1/2}\frac{c}{H}$ from all leaf parameters. Similarly, one can arbitrarily order the individual trees within a factor. Researchers should thus be careful to interpret

representative unit $\Delta(\cdot)$ but also contributes to other correlated components. In this regard, the scaled sum-of-trees model (3) resembles a factor model with loadings determined by the covariance of unit-level unobserved heterogeneity.[12]

The first-stage prior (1) using the sum-of-trees factor model (3) allows for shrinking demand parameters to a granularly-defined representative unit—one that is representative of the population of units with similar observed characteristics Z_i. As in conventional hierarchical models, a smaller variance Σ corresponds to shrinking unit-specific parameters θ_i towards the corresponding representative unit $\Delta(Z_i)$ more tightly.

While flexible dependence on potentially many observed characteristics of the representative unit $\Delta(\cdot)$ is the key benefit of the proposed approach, the representative unit cannot be arbitrarily granular without also losing all informational value. For example, even with a single continuous Z_i, no two units share the same characteristics and hence the only arbitrarily representative unit would be the unit itself. Similarly, the variance Σ should be sufficiently concentrated to impose meaningful shrinkage (Lenk and Orme, 2009). Careful consideration of the choice of second-stage prior over the parameters $\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]}$, μ, and Σ is thus crucial for any practical application.

2.3 Second-stage prior

The second-stage prior defines a probability distribution over the first-stage prior parameters. This is done straightforwardly for the unconditional mean μ and the prior variance Σ where I consider the standard independent conjugate priors. That is, the unconditional mean μ has a normal second-stage prior $\mu \sim \mathcal{N}(\overline{\mu}, A^{-1})$ with mean $\overline{\mu}$ and covariance A^{-1}, and Σ has an inverse-Wishart second-stage prior $\Sigma \sim \mathcal{IW}(\nu, \Psi)$ with degrees of freedom ν and scale matrix Ψ. Because μ and Σ are parametric and the cross-section of units n is typically large in marketing applications, these second-stage priors on (μ, Σ) are often dominated by the data.[13]

Greater care is needed for the sum-of-trees parameters $\{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}$, where R_{dh} and Λ_{dh} are the tree structure and leaf coefficients of the hth tree in the dth sum-of-tree model. Due to their high dimension, naive applications to even large cross-sections can result in overly granular (and thus non-informative) models. As usual, the Bayesian approach introduces proper probability distributions over all parameters—the variables

only the identified first-stage parameters $\Delta(\cdot)$ and Σ.

[12]This factor construction of vector-valued sums-of-trees model is new also in the larger literature on multivariate BART with *observed* outcomes (e.g., Chakraborty, 2016; Um et al., 2023; Esser et al., 2025). As illustrated in Section 3.2, the proposed sum-of-trees factor model (3) substantially simplifies posterior sampling by allowing parallel (rather than sequential) draws across the D sets of parameters $\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]}$.

[13]Appendix A provides default values for the second-stage hyperparameters (ν, Ψ), as well as hyperparameters that define the prior over the sum-of-trees parameters.

to split on, the split cutoffs, and the terminal leaf coefficients—to regularize the representative unit.[14] I impose independent regularizing priors over the parameters of each of the $d \in [D]$ sum-of-trees models $\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]}$, that is,

$$\pi\left(\{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}\right) = \prod_{d=1}^{D} \pi(\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]}),$$

where $\pi(\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]})$ is constructed by adapting the BART prior of Chipman et al. (2010). This prior further simplifies the specification via independence of the $h \in [H]$ trees and $g \in [G_{dh}]$ terminal leaves

$$\pi(\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]}) = \prod_{h=1}^{H} \pi((R_{dh}, \Lambda_{dh})) = \prod_{h=1}^{H} \left[\prod_{g=1}^{G_{dh}} \pi(\lambda_{dhg}|R_{dh})\right] \pi(R_{dh}),\ \forall d \in [D].$$

The prior over the tree structure R_{dh} is constructed following Chipman et al. (1998) who propose an iterative scheme: First, at a node of depth $q(= 0, 1, 2, \ldots)$, the tree splits into two nodes with probability

$$\alpha(1+q)^{-\beta}, \quad \alpha \in (0,1),\ \beta \in [0,\infty),$$

where α controls the base rate of splits and β controls decay of the probability of splits as the depth of the node increases. At the default values $\alpha = 0.95$ and $\beta = 2$ proposed by Chipman et al. (2010), this prior strongly prefers shallower trees with two or three terminal nodes. For example, for the first tree in Figure 1 with three terminal nodes, the prior probability of increasing its number of terminal nodes is 0.12. The prior probability of a tree with 5 or more terminal nodes is 0.03. Second, conditional on splitting a node, the prior split rule selects the kth characteristic $Z_i^{(k)}$ with probability $1/K$ and the split point c uniformly over its range.[15]

Conditional on the tree structure R_{dh}, the prior over the terminal leaf coefficients λ_{dhg} is $\mathcal{N}(0, \sigma_\lambda^2)$. To set σ_λ^2, Chipman et al. (2010) propose to pre-process the dependent variable to lie within ± 0.5 and then choose a prior variance that assigns large prior probability of the sum-of-trees model to this range. Since I employ sum-of-tree models within a hierarchical framework where the coefficients θ_i are themselves unobserved, this simple pre-processing approach does not apply. Instead, I set σ_λ^2 such that the prior representative unit $\Delta(\cdot)$ and the coefficients θ_i have largely overlapping distributions. The

[14]Note that this approach to regularization of Bayesian sums-of-trees models as fully probabilistic models contrasts their frequentist analogues. Frequentist sum-of-trees models like random forests are typically defined as loss-minimizing deterministic functions of the data, that are regularized through greedy and constrained loss-minimization.

[15]Unordered categorical characteristics are commonly accommodated via one-hot encoding, but other approaches are also possible.

first-stage prior construction (1) is convenient for this purpose: Because each sum-of-trees factor in (3) is shifted by μ and has loadings $\Sigma^{1/2}$, prior overlapping distributions can be achieved by setting σ_λ^2 such that any single sum-of-trees factor (2) assigns approximately 95% probability to the range of a standard normal random variable. With H independent trees in each sum, I thus set a default value of $\sigma_\lambda = \frac{\tau}{\sqrt{H}}$ with $\tau = 1$. Given a default number of trees of $H = 200$, this prior strongly regularizes individual terminal leaf coefficients towards zero. Regularization of the terminal leaf coefficients can readily be adjusted by setting τ to a different value.

Remark 2.1 (Adaptively sparse representative units)**.** *Linero (2018) propose a modification of the prior over the tree structure R_{dh} to facilitate nonparametric regression with many predictors. Instead of fixed equal prior probabilities of selecting a characteristic $Z_i^{(k)}$ for the split rule, variable selection is modelled as a multinomial random variable with a Dirichlet prior over its selection probabilities. Linero (2018) shows that this prior construction induces exact sparsity in the dependence of the sum-of-trees model on the observed characteristics. In empirical applications with many characteristics, this approach appears highly successful compared to standard BART and alternative nonparametric regression estimators such as random forests or gaussian processes.*

Since HART builds on top of individual sum-of-trees models, incorporating the Dirichlet prior of Linero (2018) for variable selection is straightforward. It also does not noticeably impede computation as the multinomial-Dirichlet update is conjugate. I apply Dirichlet HART in Section 5.4 to illustrate benefits of sparsity-inducing priors for characterizing the representative unit.

2.4 Posterior distribution

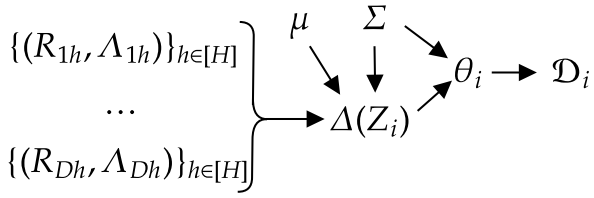
The target parameters of a HART likelihood model are the unit-specific parameters θ_i, the covariance of unobserved heterogeneity Σ, the unconditional mean μ, and the parameters of the stacked sum-of-trees model $\{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}$ that define the representative unit $\Delta(\cdot)$. The joint posterior distribution over these target parameters is given by

$$\begin{aligned} &\pi\left(\{\theta_i\}, \mu, \Sigma, \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]} \mid \mathcal{D}\right) \\ \propto &\left(\prod_{i=1}^{n} L(\mathcal{D}_i|\theta_i)\pi\left(\theta_i|Z_i, \mu, \Sigma, \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}\right)\right) \\ &\times \left[\prod_{d=1}^{D} \pi\left(\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]}\right)\right] \pi(\mu)\pi(\Sigma), \end{aligned} \tag{4}$$

where $\mathcal{D}$ denotes the full dataset. The corresponding DAG is shown in Figure 2.

Posteriors as in (4) are generally not available in an analytically convenient form.

Figure 2: DAG of a Generic HART Model



In the next section, I propose a Markov Chain Monte Carlo (MCMC) algorithm whose samples converge in distribution to (4) to facilitate fully Bayesian inference.

3 Inference via MCMC

This section describes an MCMC algorithm targeting the posterior distribution (4) for a generic HART likelihood model. The key difficulty in efficient sampling lies in the high-dimensionality of the sum-of-trees parameters $\{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}$ that render naive sampling approaches practically infeasible. I illustrate how existing Gibbs-like MCMC samplers of hierarchical Bayes models with unit-level parameters $\{\theta_i\}_{i\in[n]}$ can be augmented for HART sampling by leveraging the Bayesian backfitting algorithm of Chipman et al. (2010). Subsection 3.1 outlines the proposed Metropolis-within-Gibbs algorithm at a high level. The sampling step to generate draws for the representative unit is described in additional detail in Subsection 3.2.

3.1 Metropolis-within-Gibbs sampler

The proposed sampler alternates between sampling from the full conditional distributions of the unit-level parameters, the unconditional mean, the covariance of unobserved heterogeneity, and the sum-of-trees parameters. Computation is substantially simplified by first augmenting the posterior distribution (4) with a *partial* representative unit

$$\Delta^*(Z_i; \Sigma, \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}) \equiv \Delta(Z_i; \mu, \Sigma, \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}) - \mu. \tag{5}$$

After simplifying the full conditionals by factoring the joint posterior distribution and using (5), this suggests iterating through the following steps:

1. $\pi(\{\theta_i\}_{i\in[n]} \mid \Delta(\cdot), \Sigma, \mathcal{D})$
2. $\pi(\mu \mid \{\theta_i\}_{i\in[n]}, \Delta^*(\cdot), \Sigma)$
3. $\pi(\Sigma \mid \{\theta_i\}_{i\in[n]}, \Delta(\cdot))$
4. $\pi(\Delta^*(\cdot), \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]} \mid \{\theta_i\}_{i\in[n]}, \mu, \Sigma)$

Step 1 draws the unit-level parameters. This step takes as given the representative unit, the covariance of unobserved heterogeneity, and the data. Importantly, whether the representative unit is defined as a conventional linear projection or a sum-of-trees factor model—or *any* other model of the representative unit—does not affect this sampling step. The sampler thus readily accommodates existing approaches for sampling unit-level parameters. For example, Rossi et al. (2009) discuss an "improved random walk" Metropolis step applicable in many marketing models. This step constructs i-specific proposal distributions for θ_i based on the Hessian of a weighted average of the pooled and individual-specific log-likelihoods.[16] Recent advances in efficient sampling of hierarchical models as in Bumbaca et al. (2020) are also readily applicable here.

Steps 2 and 3 are the usual conjugate updates for the mean and covariance matrix of a multivariate normal with outcomes constructed as $\theta_i - \Delta^*(Z_i)$ and as $\theta_i - \Delta(Z_i)$, respectively. Note that Step 3 is simple due to the augmentation with the partial representative unit $\Delta^*(\cdot)$, as the conjugate updates would otherwise be complicated by the dependence induced by the loadings in (3).

Finally, Step 4 updates the partial representative unit and the sum-of-trees parameters. Due to their high dimension, it is the most computationally intensive step. In addition to their high number, an additional potential complication is that their dimension changes as trees grow or are pruned. For a single sum-of-trees model with an observed normally distributed scalar-valued outcome, Chipman et al. (2010) propose a "Bayesian backfitting" algorithm that effectively samples from the high dimensional posterior without the need for a transdimensional transition kernel (e.g., Green, 1995). In the next subsection, I outline how this algorithm can be applied within the hierarchical model considered here.

3.2 Bayesian backfitting for the sum-of-trees parameters

To sample from the full conditional distribution of the partial representative unit $\Delta^*(\cdot)$ and the sum-of-trees parameters $\{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}$, Step 4 is further split into two substeps:

$$
\begin{aligned}
4.\quad & \pi(\Delta^*(\cdot), \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]} \mid \{\theta_i\}_{i\in[n]}, \mu, \Sigma) \\
4.1\quad & \pi(\{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]} \mid \{\theta_i\}_{i\in[n]}, \mu, \Sigma) \\
4.2\quad & \pi(\Delta^*(\cdot) \mid \{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]}, \Sigma)
\end{aligned}
$$

This split marginalizes out the representative unit in Step 4.1, which—given that $\Delta^*(\cdot)$ is a deterministic function of the sum-of-trees parameters and Σ—is both immediate and

[16] The improved random walk Metropolis step is used in `bayesm` (Rossi, 2023) and thus widely adopted. See Rossi et al. (2009) Chapter 5.3 for a detailed description.

makes Step 4.2 computationally straightforward. The core challenge thus lies in Step 4.1.

The simplification of Step 4.1 is due to a reparameterization of the first-stage prior (1). Applying a Mahalanobis transformation by residualizing with the unconditional mean and pre-multiplying with the inverse Cholesky decomposition of Σ yields

$$\begin{aligned} \Sigma^{-1/2}(\theta_i - \mu) &= \Sigma^{-1/2}\Sigma \begin{bmatrix} \delta(Z_i; \{(R_{1h}, \Lambda_{1h})\}_{h\in[H]}) \\ \vdots \\ \delta(Z_i; \{(R_{Dh}, \Lambda_{Dh})\}_{h\in[H]}) \end{bmatrix} + \Sigma^{-1/2}\varepsilon_i \\ \Leftrightarrow \qquad \tilde{\theta}_i &= \begin{bmatrix} \delta(Z_i; \{(R_{1h}, \Lambda_{1h})\}_{h\in[H]}) \\ \vdots \\ \delta(Z_i; \{(R_{Dh}, \Lambda_{Dh})\}_{h\in[H]}) \end{bmatrix} + \tilde{\varepsilon}_i, \end{aligned} \tag{6}$$

where now $\tilde{\varepsilon}_i \sim \mathcal{N}(0, I_D)$ is a vector of independent standard normal random variables. Consequently, the conditional distribution of the sum-of-trees parameters factors into D independent conditional distributions, each depending only on the respective scalar-valued components $\tilde{\theta}_i^{(d)} \equiv (\Sigma^{-1/2}(\theta_i - \mu))^{(d)}$, that is,

$$\pi(\{(R_{dh}, \Lambda_{dh})\}_{d\in[D],h\in[H]} \mid \{\theta_i\}_{i\in[n]}, \mu, \Sigma) = \prod_{d=1}^{D} \pi(\{(R_{dh}, \Lambda_{dh})\}_{h\in[H]} \mid \{\tilde{\theta}_i^{(d)}\}_{i\in[n]}). \tag{7}$$

Step 4.1 is thus further split into D (embarrassingly parallel) steps, each tasked with a single draw from the marginal conditional distributions in (7). Importantly, to generate these single draws, the developed setup allows for the application of a single step of the Bayesian backfitting algorithm of Chipman et al. (2010), where the Mahalanobis-transformed components $\tilde{\theta}_i^{(d)}$—now scalar-valued, mean-zero, and with unit variance—play the role of the observed outcome projected onto the characteristics Z_i.

Within its application in the proposed sampler, I make no modifications to the Bayesian backfitting step of Chipman et al. (2010) itself. Because it is central for the computational feasibility of HART, I nevertheless briefly outline the step for clarity. At its core, the Bayesian backfitting step is a partially collapsed Metropolis-within-Gibbs step, cycling through each of the $h \in [H]$ trees of the sum-of-trees model:

$$\pi((R_{dh}, \Lambda_{dh}) \mid \{\tilde{\theta}_i^{(d)}\}_{i\in[n]}, \{(R_{dh'}, \Lambda_{dh'})\}_{h'\in[H]\setminus\{h\}}), \qquad \forall d \in [D], h \in [H]. \tag{8}$$

This tree-specific sampling step is greatly simplified via the construction of the partial residuals

$$\tilde{\theta}_i^{(d)} - \delta(Z_i; \{(R_{h'}, \Lambda_{h'})\}_{h'\in[H]\setminus\{h\}}) \quad \sim \quad \mathcal{N}\left(\delta(Z_i; (R_h, \Lambda_h)), 1\right), \tag{9}$$

where the normal distribution follows directly from (7) and the additivity of the sum-of-trees model (2). Given conjugacy of (9) and the normal prior over the terminal leaf coefficients $\{\lambda_{dhg}\}_{g\in[G_{dh}]}$ outlined in Section 2.3, this allows for integrating over all terminal leaf coefficients to draw the tree structure R_{dh}. This marginalization is done analytically and thus requires no computationally prohibitive integral evaluation. A new tree structure can then be sampled via a simple Metropolis step by randomly growing or pruning the existing tree using a "birth-death" algorithm proposed by Chipman et al. (1998).[17] After accepting or rejecting the proposed tree structure, the terminal leaf coefficients $\{\lambda_{dhg}\}_{g\in[G_{dh}]}$ are set via the usual conjugate update for each unique partition of the characteristics Z_i. This completes a single tree-specific draw and the next tree is targeted. The efficiency of these Bayesian backfitting steps has contributed substantially to the popularity of BART for nonparametric regression (e.g., Hill et al., 2020), with the motivation for the proposed HART likelihood models being no exception.[18]

4 Example: HART Logit Model

To illustrate HART likelihood models, I apply the proposed flexible hierarchical structure to the canonical setting of static discrete choice with multinomial logit likelihoods as in Allenby and Ginter (1995), Rossi et al. (2009), Smith et al. (2023), and many others. This choice model is widely applied in conjoint analysis and consumer brand choice, and I apply the HART logit model in these contexts in the empirical applications in Sections 5 and 6, respectively.

Consider a sample of $i \in [n]$ consumers each associated with historical demand data $\{(Y_{it}, W_{it})\}_{t=1}^{T_i}$ for T_i purchase occasions, where Y_{it} is a consumer's choice among J options and $W_{it} \equiv (W_{ijt})_{j=1}^{J}$ is a vector of product characteristics such as brand intercepts, prices, and other marketing mix variables. I assume that at every purchase occasion t, a consumer chooses option $j \in [J]$ with probability

$$\mathrm{P}(Y_{it} = j|\theta_i, W_{it}) = \frac{\exp\left(W_{ijt}^\top \theta_i\right)}{\sum_{l=1}^{J} \exp\left(W_{ilt}^\top \theta_i\right)}, \qquad \forall\,(i,j,t),$$

where θ_i is a D-dimensional consumer-specific parameter vector. The model follows from the standard latent utility framework with linear-index deterministic utilities $W_{ijt}^\top \theta_i$ and

[17] Chipman et al. (1998) suggest building a new tree with one of four randomly selected moves: w.p. 0.25 a new terminal node is added, w.p. 0.25 a pair of terminal nodes is pruned, w.p. 0.4 a nonterminal splitting rule is changed, w.p. 0.1 splitting rules between a parent and child node are swapped.

[18] He and Hahn (2023) propose XBART, an accelerated algorithm for sampling stochastic sum-of-tree models. Combining XBART with HART is likely to further improve sampling efficiency while maintaining flexibility of the representative consumer model.

i.i.d. additively separable T1EV latent utility shocks (McFadden, 1974). Under conventional assumptions on static demand (e.g., Rossi et al., 1996), this implies a consumer-specific likelihood given by

$$L(\mathcal{D}_i|\theta_i) \equiv \prod_{t=1}^{T_i} L_t(Y_{it}|\theta_i, W_{it}) = \prod_{t=1}^{T_i}\prod_{j=1}^{J} \mathrm{P}(Y_{it}=j|\theta_i, W_{it})^{\mathbb{1}\{Y_{it}=j\}}, \quad \forall i. \tag{10}$$

The logit likelihood of a consumer's purchase history (10) can be directly substituted into the generic posterior distribution in (4). In this context, $\Delta(\cdot)$ captures a representative consumer—granularly defined using their characteristics Z_i—towards which the preference parameters θ_i are shrunken. This is particularly important for demand estimation in many consumer packaged goods (CPGs) categories where purchases rarely exceed 12 per year (Allenby and Rossi, 1999), and in common conjoint applications where respondents are rarely exposed to more than 16 profiles to prevent survey fatigue (Allenby et al., 2019).

To sample from the HART logit model's posterior, the generic Metropolis-within-Gibbs algorithm of Section 3.1 requires a suitable first step. I use the "improved random walk" of Rossi et al. (2009) and find that it results in efficient sampling in the considered applications of the HART logit model to which I turn next.

5 Application I: Credit Card Conjoint

This section applies the proposed HART logit model to the conjoint dataset considered in Allenby and Ginter (1995, AG95, hereafter) on out-of-state credit card design. The key finding is that HART finds substantively richer heterogeneity associated with observed respondents' characteristics than conventional specifications that define the representative consumer as a linear function of characteristics. I show that knowledge of greater heterogeneity can be exploited by managers for product design targeted to specific consumer segments in the new market.

After a brief overview of the conjoint data of AG95 in Section 5.1, I discuss the main estimation results in Section 5.2. Section 5.3 presents counterfactual estimates for different product designs. Finally, Section 5.4 illustrates robustness of Dirichlet HART to the inclusion of many irrelevant consumer characteristics.

5.1 Data

AG95 consider data from 946 customers of a regional bank collected in a telephone conjoint on credit card attributes. The bank's main motivation stemmed from an effort to expand

to customers in a new market ("out-of-state"). Each respondent offered between 13 and 17 responses to hypothetical choices between two credit cards that were identical in all considered attributes except for two that varied.[19] The authors obfuscate both the identity of the bank and the specific attribute levels—for easy reference, Appendix B.1 replicates their overview of credit card attributes and levels. In total, the dataset contains 14,799 binary responses as well as the respondents' age, income, and gender.

5.2 Demand estimates

I apply both the proposed HART logit model and the conventional hierarchical logit model considered in AG95. Both models use the respondent-level likelihood in Section 4 where W_{ijt} is a 14-dimensional vector of binary credit card attribute levels and the respondent-specific coefficients θ_i are interpreted as the corresponding part-worths. Following AG95, I consider a normal model of heterogeneity in both specifications, using the default inverse-Wishart second-stage prior for Σ. The only difference in the two approaches lies in the specification of the representative consumer as a function of the three demographics: While the proposed HART approach uses a sum-of-trees factor model for the representative consumer $\Delta(Z_i)$, AG95 consider a linear model $\Delta^\top Z_i$. The components of the representative consumer are interpreted as expected part-worths of the respective credit card attribute levels.

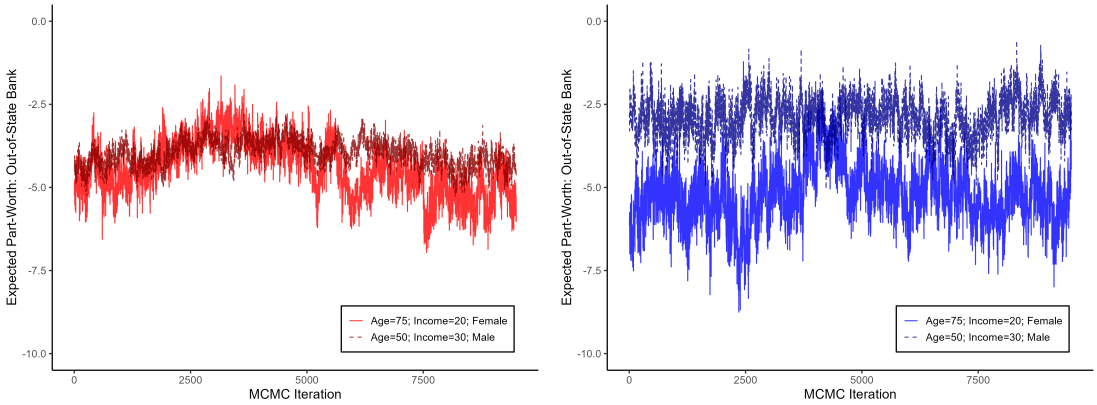
The MCMC algorithms are run for 10,500 iterations to obtain posterior samples (see Remark 5.1 for a runtime comparison). Traceplots of the MCMC draws indicate convergence of the chains after less than 500 iterations, which I discard as burn-in. The remaining 10,000 draws are used for inference. As an example, Figure 3 shows the traceplots of the expected part-worths of the out-of-state bank attribute for two consumer segments. The proposed Metropolis-within-Gibbs sampler shows effective mixing of the chain. Additional traceplots are provided in Appendix B.2.

I begin by highlighting several important aspects in which HART and conventional linear estimates overlap. First, both models find similar *overall* (or: unconditional) expected part-worths. Columns (1)-(2) in Table 1 provide posterior means and standard deviations for selected attribute levels. Both provide evidence that a randomly selected respondent substantially penalizes an out-of-state credit card. The magnitudes of the low interest rate and low annual fee part-worths further suggest that the bank may compensate for the out-of-state penalty by adjusting other credit card attributes.

Similarly, both models find evidence of substantial unobserved heterogeneity. Table 2

[19] Allenby and Ginter (1995, p. 395) provide the following example of a hypothetical choice scenario: "The first card has a medium fixed annual interest rate and a medium annual fee. The second card has a high fixed annual interest rate and a low annual fee."

Figure 3: MCMC Traceplot of Expected Part-Worths for Out-of-State Bank



(a) Linear hierarchical prior (b) HART

Notes: Figures show traceplots of expected part-worths for the out-of-state bank attribute. Panel (a) and (b) show results for the linear hierarchical prior and the HART prior, respectively. Expected part-worths for each model are evaluated for two consumer segments: older women with low income (solid lines) and middle-aged men with moderate income (darker dashed lines).

provides a subset of the posterior mean covariance matrix for the proposed HART logit model. The estimates indicate economically substantial heterogeneity in the out-of-state part-worths. They further suggest that respondents who are less sensitive to the out-of-state attribute also more strongly prefer low annual fees (posterior mean correlation: 0.67). In contrast, the correlation in preferences between out-of-state credit cards and low interest rates is low (posterior mean correlation: 0.11). Posterior covariance estimates based on the conventional linear model show qualitatively similar levels and correlations of preferences (see Appendix B.2).

Despite these similarities, a closer inspection of the representative consumer estimates reveals economically important differences between the approaches. First, where the linear prior approach identifies monotonic associations between respondents' characteristics and their part-worths, HART finds evidence of nonmonotonicity. Figures 4 and 5 show expected part-worths of the linear and HART logit models as functions of age and income. Panel (a) of Figure 4 shows that being middle-aged is associated with little sensitivity to out-of-state credit cards, while both younger and older respondents penalize it more strongly. Similarly, panel (b) of Figure 5 suggests that the strong positive association between income and a respondent's preference for a low interest rate holds primarily for low and middle-income levels but is mitigated for high income levels. The linear specifications average over any such nonlinearities, losing potential insights in the process.

In addition to nonlinear associations between preferences and individual characteristics, the HART logit model also accommodates interactions between respondents' age, income, and gender. These nonlinear interactions have substantial impact on the expected part-worths associated with more granularly-defined consumer segments.

Table 1: Expected Part-Worths Estimates

	Overall		Segment 1: Female; Age=75; Income=20;		Segment 2: Male; Age=50; Income=30;	
	Linear (1)	HART (2)	Linear (3)	HART (4)	Linear (5)	HART (6)
Low Interest	5.01	4.98	4.08	3.17	4.45	5.10
	(0.68)	(0.94)	(0.48)	(0.65)	(0.27)	(0.56)
Low Annual Fee	4.19	4.17	4.78	4.28	3.69	4.21
	(0.70)	(1.03)	(0.52)	(0.74)	(0.30)	(0.58)
Out-of-State Bank	-3.78	-3.78	-4.40	-5.19	-3.96	-2.89
	(0.68)	(1.04)	(0.85)	(0.92)	(0.40)	(0.62)
High Cash Rebate	2.41	2.48	1.71	1.97	1.96	1.71
	(0.62)	(0.85)	(0.50)	(0.62)	(0.29)	(0.50)
High Credit Limit	1.15	1.16	1.16	1.42	1.05	1.41
	(0.32)	(0.60)	(0.33)	(0.46)	(0.18)	(0.37)
Long Grace Period	3.51	3.52	2.48	2.74	2.96	2.88
	(0.67)	(0.69)	(0.36)	(0.51)	(0.23)	(0.39)

Notes: The table shows posterior means of expected part-worths $\Delta(\cdot)$ for selected attribute levels. Posterior standard deviations are in parentheses. Columns (1)-(2) show overall (or: unconditional) expected part-worths. Columns (3)-(4) and columns (5)-(6) show expected part-worths for two consumer segments, older women with low income and middle-aged men with moderate income, respectively. Odd columns correspond to results for the linear hierarchical prior logit model. Even columns correspond to the HART logit model. Results are based on 10,500 MCMC draws with 500 draws discarded as burn-in.

My discussion focuses on two segments. Segment 1 is defined as older female respondents with low income. Segment 2 is defined as middle-aged male respondents with moderate income. These segments are selected for two primary reasons. First, Segment 1 is the most populous segment of female respondents in the data of AG95 (see Appendix B.2) and is chosen because posterior uncertainty of the representative consumer is inversely related to the number of respondents with similar characteristics. Segment 2 is of similar size and thus expected to have posterior uncertainty of approximately the same order. Second, the segments allow for illustration of substantial differences between conventional linear and HART logit models. Importantly, however, the conclusion that the HART logit model finds richer heterogeneity does not depend on the specific segments—see Appendix B.2 for the empirical distribution of posterior means of the expected part-worths for all respondents.

Table 1 presents results for both segments in columns (3)-(6). The conventional linear approach finds only moderate differences in the expected part-worths between the two segments. For example, the difference in the out-of-state part-worths between the two segments is only 0.44, a difference of approx. 10%, with middle-aged male customers only slightly less sensitive to the out-of-state attribute than their older female counterparts. Similarly, the difference in the low interest rate part-worths is -0.37, a difference of approx. -9%, with middle-aged male customers slightly more keen on low interest rates than

Table 2: HART Covariance Matrix Estimates

Low fixed interest	8.68	0.30	0.11	0.38	0.24	0.36
	(1.50)					
Low annual fee	3.52	15.34	0.67	0.47	0.47	0.58
	(1.15)	(1.90)				
Out-of-state bank	1.25	10.07	14.63	0.42	0.46	0.28
	(1.16)	(1.70)	(2.31)			
High cash rebate	3.53	5.68	4.92	9.62	0.48	0.71
	(1.20)	(1.28)	(1.16)	(1.26)		
High credit limit	1.84	4.73	4.47	3.84	6.75	0.47
	(0.83)	(0.76)	(0.89)	(0.63)	(0.85)	
Long grace period	2.49	5.28	2.48	5.12	2.83	5.34
	(1.00)	(0.90)	(0.80)	(0.97)	(0.62)	(0.88)

Notes: The table shows posterior means of the lower-triangular covariance matrix Σ for selected attribute levels. Posterior standard deviations are in parentheses. The upper triangular matrix shows correlations. Results are based on 10,500 MCMC draws of the HART logit model with 500 draws discarded as burn-in.

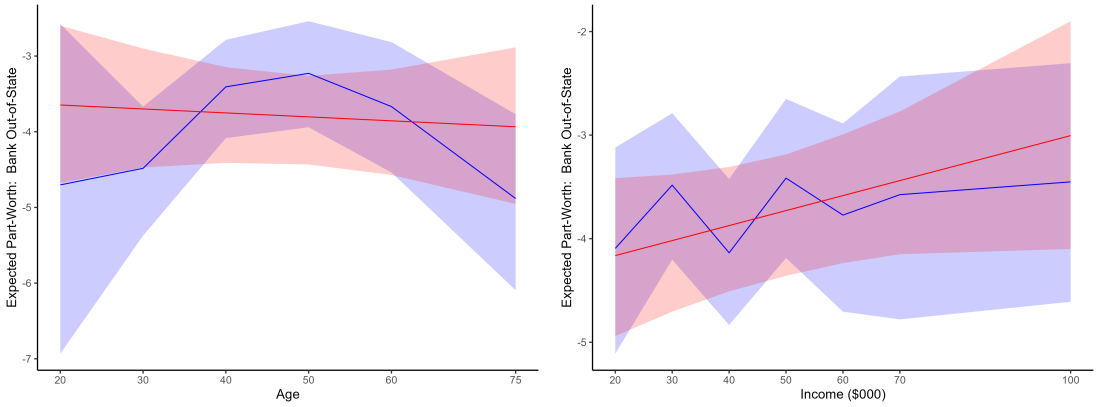
their older female counterparts. In stark contrast, HART finds substantially different preferences for the two segments: The differences between the expected part-worths in the out-of-state and low interest rate attributes are 2.3 (approx. 44%) and -1.93 (approx. -61%), respectively. In the next section, I illustrate how managers can exploit this heterogeneity for targeted product design.

5.3 Counterfactual estimates

For the bank considering an expansion of its credit card offering outside its existing operating region, knowledge about consumer segments that are most likely to respond to a new credit card is crucial for both product design and advertising. Indeed, AG95 emphasize (p.392): "To succeed in a competitive environment, organizations must identify which customers are *most* likely to buy new products and services [...]." With its ability to characterize the demand of granular consumer segments using observed characteristics, the proposed HART logit model is ideally suited for this task of focusing attention on key segments.

Following AG95, I consider the design of an out-of-state credit card that is most likely to succeed against existing credit card offerings. Table 3 presents counterfactual choice probabilities for potential offerings against a baseline in-state credit card. Importantly, these counterfactuals characterize the comprehensive demand response capturing both observed preference heterogeneity through the representative consumer $\Delta(\cdot)$ and unobserved preference heterogeneity—including correlated preferences—captured by Σ. Continuing the previous discussion, I focus on compensating for the out-of-state attribute by offering a credit card with low interest rates or a low annual fee and compute the counterfactual

Figure 4: Conditional Part-Worths Estimates (Bank Out-of-State)



(a) Age (b) Income

Notes: Figures show the posterior mean expected part-worth $\Delta(\cdot)$ of the out-of-state bank attribute as functions of respondents' age (panel (a)) and income (panel (b)). Expected part-worths are computed by integrating over the empirical distribution of other characteristics. Shaded areas denote 90% point-wise credible intervals. Expected part-worths of the linear hierarchical prior are shown in red and of the HART model in blue. Results are based on 10,500 MCMC draws with 500 draws discarded as burn-in.

shares in the two segments of older female customers with low income, and of middle-aged male customers with moderate income.

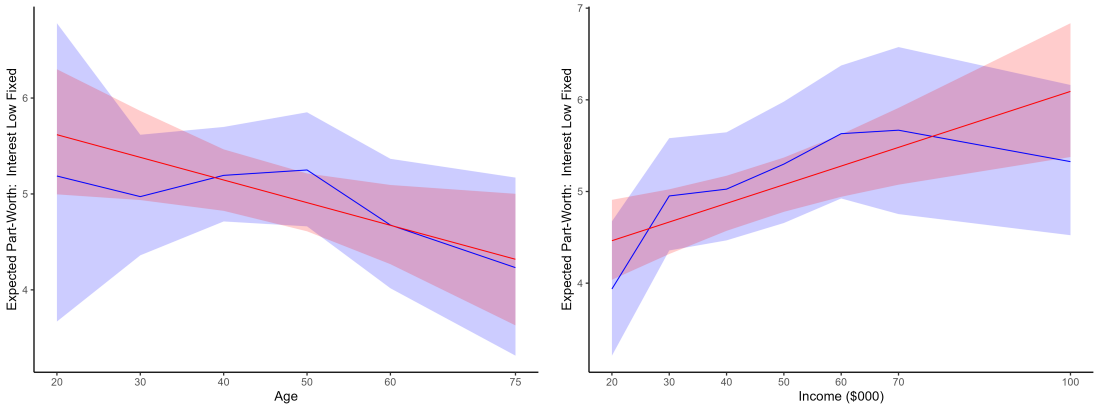
Table 3: Counterfactual Shares of Out-of-State Credit Card Offerings

	Segment 1: Female; Age=75; Income=20;		Segment 2: Male; Age=50; Income=30;	
Credit Card	Linear (1)	HART (2)	Linear (3)	HART (4)
Out-of-State Bank	0.16	0.11	0.18	0.25
	(0.05)	(0.05)	(0.04)	(0.06)
Out-of-State Bank & Low Interest	0.48	0.35	0.53	0.66
	(0.08)	(0.09)	(0.06)	(0.07)
Out-of-State Bank & Low Annual Fee	0.52	0.45	0.48	0.57
	(0.08)	(0.08)	(0.05)	(0.07)

Notes: The table shows posterior means of counterfactual market shares of three out-of-state credit card offerings evaluated against a baseline in-state credit card. Posterior standard deviations are in parentheses. Columns (1)-(2) and columns (3)-(4) show counterfactual market shares for two consumer segments, older women with low income and middle-aged men with moderate income, respectively. Odd columns correspond to results for the linear hierarchical prior logit model. Even columns correspond to the HART logit model. Results are based on 10,500 MCMC draws with 500 draws discarded as burn-in.

Both the HART and conventional linear logit model estimates find low expected demand for out-of-state credit cards when other attributes are at baseline. Similarly, both models suggest that low interest rates and annual fees can increase attractiveness of the offering. Yet, the models starkly contrast in their ability to differentiate between consumer segments that are worth targeting. The linear estimates suggest no substantial differences across both segments and compensating attributes. For example, the counterfactual share of a low annual fee out-of-state credit card among older women with low income is 0.52,

Figure 5: Conditional Part-Worths Estimates (Low Interest)



(a) Age (b) Income

Notes: Figures show the posterior mean expected part-worth $\Delta(\cdot)$ of the low interest rate attribute as functions of respondents' age (panel (a)) and income (panel (b)). See also the notes of Figure 4.

while an offering with a low interest rate among middle-aged men with moderate income is 0.53. With posterior standard deviations around 0.08, these differences are unlikely to inspire confidence in a manager forced to choose which credit card to offer and which potential customers to advertise it to. In contrast, the HART estimates provide clear evidence in favor of targeting a low interest card to the segment of middle-aged men: the corresponding counterfactual share is 0.66 versus 0.45 for the low annual fee card targeted to the segment of older women with a posterior standard deviation of about 0.08.

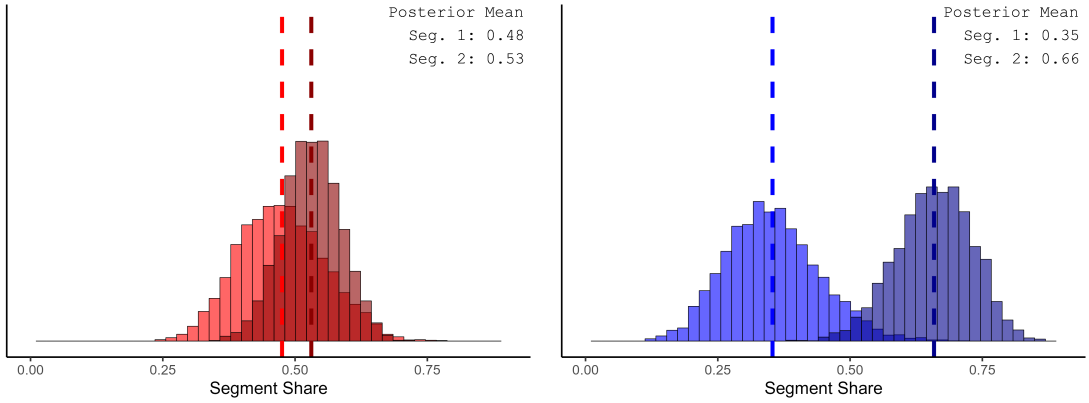
These stark differences in the models' ability to differentiate between segments that are worth targeting is also reflected in the counterfactual shares for the individual credit card offerings. Figure 6 shows the posterior distribution of counterfactual shares of a low interest out-of-state credit card for both segments.[20] The posterior counterfactual distributions of the linear approach in panel (a) overlap substantially, indicating little gain to targeted advertising or other personalized approaches. The analogous HART logit posteriors in panel (b) show economically and statistically significant differences, clearly outlining the bank's opportunity to target a low interest rate card to the segment of middle-aged men with moderate income.

5.4 Robustness to high-dimensional characteristics

In addition to restrictive functional forms, previous research has raised concerns about the intractability of linear representative consumer specifications when the number of observed characteristics is large (e.g., Padilla and Ascarza, 2021). I use the setting of AG95 to illustrate the effects of many characteristics in a placebo exercise where the three demo-

[20] Appendix B.3 provides additional plots of posterior counterfactual shares.

Figure 6: Counterfactual Shares for an Out-of-State Credit Card with Low Interest Rate



(a) Linear hierarchical prior (b) HART

Notes: The figure shows the posterior distribution of expected counterfactual shares of an out-of-state credit card with low interest rate against a baseline in-state credit card. Panel (a) and (b) show results for the linear hierarchical prior and the HART prior, respectively. Expected counterfactual shares for each model are evaluated for two consumer segments: older women with low income (solid) and middle-aged men with moderate income (dark shaded). Results are based on 10,500 MCMC draws with 500 draws discarded as burn-in.

graphic variables are augmented by additional *irrelevant* characteristics. These irrelevant characteristics are 100 simulated standard normal random variables drawn independently for each of the respondents. Any association between these additional characteristics and consumer choices is purely spurious by design, which allows us to evaluate competing models by the robustness of their estimates to the inclusion of these irrelevant characteristics.

In addition to the linear and HART logit specifications, I also consider Dirichlet HART logit which augments the sum-of-trees prior with a Dirichlet prior over the variable selection probabilities. As outlined in Remark 2.1, the Dirichlet HART model is particularly suitable for high-dimensional settings by directly inducing sparsity. All three models use the default second-stage prior settings outlined in Appendix A.

Estimates are based on 10,500 MCMC iterations with the first 500 iterations discarded as burn-in. Efficiency of the MCMC algorithm for both the HART and Dirichlet HART logit models is only mildly affected. In particular, traceplots indicate convergence of the HART and Dirichlet HART logit models after less than 500 iterations as in the case without additional characteristics (see Appendix B.4), and there is no substantial effect on runtime (see Remark 5.1). In contrast, the runtime of the linear model increases from approximately 2 minutes to nearly 8.5 *hours*. Even with this long runtime, traceplots for the linear hierarchical logit model do not indicate convergence.

To assess the robustness of the models to the inclusion of many irrelevant characteristics, I first assess their out-of-sample predictive accuracy. Table 4 reports the root mean squared prediction error (RMSPE) of respondents' binary choices. RMSPEs are estimated by 10-fold cross-validation in which the last $T_i - \tilde{T}$ choice profiles of randomly selected

respondents are assigned to 10 equally-sized subsamples. Here, $\tilde{T}$ denotes the number of observed profiles for the out-of-sample respondents that is varied between 0, 1, 5, 9, and 14 across the columns of the table. For predictions based on $\tilde{T} = 5$ observed profiles, for example, the models are first estimated on the full conjoint data of 9/10ths of the respondents *and* the first 5 choice profiles of the remaining 1/10th. The individual-level posterior predictive distributions are then used to create predictions for the remaining $T_i - \tilde{T}$ choice profiles of the out-of-sample respondents.[21] Varying $\tilde{T}$ in this manner allows for the assessment of model performance for both known and new consumers.

Table 4: Out-of-Sample RMSPE

	# Observed Profiles				
	0	1	5	9	14
3 Demos.					
Linear	0.428	0.425	0.417	0.393	0.367
HART	0.428	0.426	0.415	0.392	0.369
3 Demos. + 100 Noise					
Linear	0.481	0.473	0.461	0.432	0.413
HART	0.429	0.426	0.417	0.394	0.369
Dirichlet HART	0.427	0.426	0.417	0.396	0.366

Notes: The table shows root mean squared prediction error (RMSPE) of respondents' binary choices. RMSPEs are estimated by 10-fold cross-validation where the last $T_i - \tilde{T}$ choice profiles of randomly selected respondents are assigned to 10 equally-sized subsamples. $\tilde{T}$ denotes the number of observed profiles for the out-of-sample respondents that is varied between 0, 1, 5, 9, and 14 across the columns of the table. Choice predictions are computed as posterior means over the choice probabilities. The top panel (3 Demos.) shows results using the three respondent characteristics of AG95. The bottom panel (3 Demos. + 100 Noise) shows results where respondent characteristics are augmented by 100 standard normal irrelevant characteristics.

The RMSPE results show that inclusion of 100 irrelevant characteristics substantially worsens the predictive accuracy of the linear specification but has little effect on the HART and Dirichlet HART logit models. For completely new consumers with 0 observed profiles, for example, the RMSPE of the linear specification is 0.481, compared to 0.429 and 0.427 for the HART and Dirichlet HART logit models, respectively. Further, the irrelevant characteristics cause a lasting drop in the linear first-stage prior's performance: the predictive accuracy of the linear specification with 9 observed profiles is 0.432 when the irrelevant characteristics are included in the first-stage prior. This is worse than its predictive accuracy for entirely new consumers when only the 3 demographics of AG95 are used. The inclusion of 100 irrelevant characteristics thus "costs" the linear specification more than 9 conjoint profiles. No such predictive cost is incurred by the HART and

[21] Posterior choice predictions are computed as posterior means over the choice probabilities and hence correspond to the squared-loss Bayes estimator.

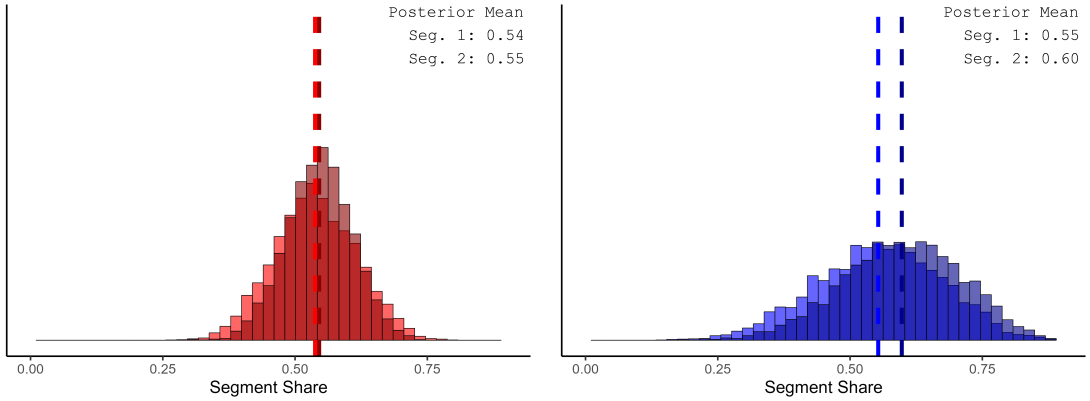
Dirichlet HART logit models.

Of key interest to marketing applications is the robustness of the models' counterfactual estimates. Figure 7 shows the corresponding posterior distribution of counterfactual shares of a low interest out-of-state credit card for the two segments previously considered in Section 5.3. The counterfactual predictions of the linear specifications are nearly entirely overlapping over the two segments showing no qualitative heterogeneity between demand of older women and middle-aged men. This additional reduction in estimated heterogeneity compared to Figure 6 is also reflected in HART estimates. However, while the linear specification shows posterior distributions of nearly equal dispersion as in the low-dimensional case, the HART estimates reflect higher posterior uncertainty over the two segments. In this regard, the HART logit estimates appear to more suitably characterize estimation uncertainty in the high-dimensional setting than the conventional linear specification.

In contrast to both the linear and HART logit models, the Dirichlet HART logit model maintains heterogeneous demand for a low interest out-of-state credit card across the two segments. Akin to the HART estimates of Figure 6, the Dirichlet HART logit model finds that targeting a low interest rate card to the segment of middle-aged men with moderate income is more likely to be successful than targeting a low annual fee card to the segment of older women with low income. Dirichlet HART thus provides the most robust counterfactual estimates of the three models in the presence of many irrelevant characteristics.

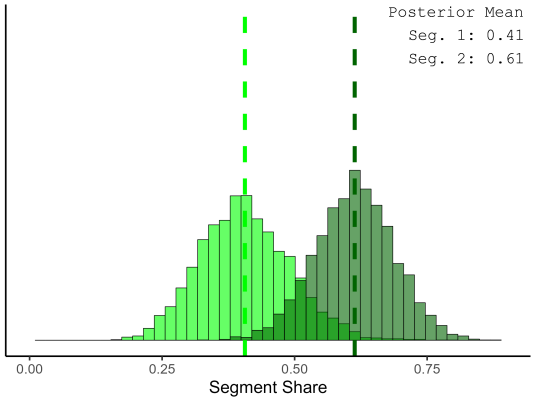
Note that the Dirichlet HART logit model's robustness to the inclusion of many irrelevant characteristics stems from its ability to induce sparsity in the definition of the representative consumer by placing a Dirichlet prior over the variable selection probabilities in the sum-of-trees models. This approach allows for a fully Bayesian characterization of variable importance. Figure 8 shows the posterior distribution of the variable selection probabilities corresponding to the out-of-state sum-of-trees factor. The plot indicates that selection probabilities for both income and gender are significantly different from 5%, while there is greater posterior uncertainty about the importance of respondents' age. Importantly, however, all additional noise characteristics are correctly identified as unimportant with posterior choice probabilities at or near 0%.

Figure 7: Counterfactual Shares for an Out-of-State Credit Card with Low Interest Rate



(a) Linear hierarchical prior

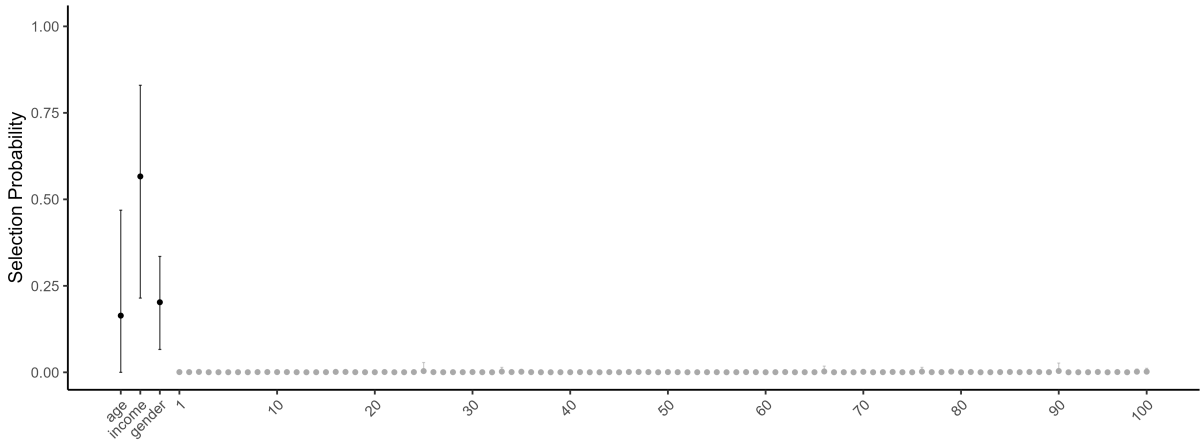
(b) HART



(c) Dirichlet HART

Notes: The figure shows the posterior distribution of expected counterfactual shares of an out-of-state credit card with low interest rate against a baseline in-state credit card. Estimates were computed using both the three demographic variables and 100 irrelevant characteristics. Panel (a), (b), and (c) show results for the linear hierarchical prior, the HART prior, and the Dirichlet HART prior, respectively. Expected counterfactual shares for each model are evaluated for two consumer segments: older women with low income (solid) and middle-aged men with moderate income (dark shaded). Results are based on 10,500 MCMC draws with 500 draws discarded as burn-in.

Figure 8: Posterior Variable Selection Probabilities: Out-of-State Bank



Notes: The figure shows the posterior mean selection probabilities for the out-of-state bank attribute sum-of-trees factor across all respondents' characteristics (3 characteristics of AG95 and 100 irrelevant characteristics). Bars indicate 90% credible intervals. Results are based on 10,500 MCMC draws with 500 draws discarded as burn-in.

Remark 5.1 (Run-time Comparison)**.** *The table below shows the runtime for the proposed HART and Dirichlet HART logit models for 10,500 MCMC iterations in the dataset described in Section 5.1. The runtime of the conventional linear hierarchical prior computed via the* `bayesm` *package is also reported.*

	Representative Consumer Model		
	Linear	*HART*	*Dirichlet HART*
3 Demos.	*1.82*	*23.68*	*24.83*
3 Demos. + 100 Noise	*508.40*	*18.83*	*21.37*

Notes: The table shows the runtime in minutes for 10,500 MCMC iterations. Columns correspond to the linear hierarchical prior, the HART prior, and the Dirichlet HART prior, respectively. Rows correspond to estimation based on the three respondents' characteristics of AG95 (3 Demos.), and the three respondents' characteristics of AG95 augmented by 100 standard normal irrelevant characteristics (3 Demos. + 100 Noise). All computations were performed sequentially on a single core of a 2019 Intel Core i7-1065G7 processor.

6 Application II: Personalized Mayonnaise Coupons

This section applies the HART logit model to a scanner panel dataset on mayonnaise purchases. I draw two key conclusions: First, incorporating additional consumer characteristics improves HART's out-of-sample predictions but worsens those of conventional approaches. Second, using a double/debiased profit estimator for evaluation of counterfactual coupon policies, I find that personalization with HART improves profits for both new and known consumers when compared to existing pricing and conventional personalization schemes.

I describe the data in Section 6.1, and compare HART logit profit estimates with conventional alternatives in Section 6.2. Section 6.3 describes the construction of counterfactual personalized coupon policies and their nonparametric out-of-sample evaluation.

6.1 Data

The dataset is constructed from the household panel and retailer scanner data provided by NielsenIQ. I focus on the mayonnaise category and consider purchases of 30oz jars of its three primary brands Hellmann's, Kraft, and private label in the 2010-2013 period in seven midwestern states (Illinois, Indiana, Iowa, Michigan, Minnesota, Ohio, Wisconsin). The mayonnaise market in the midwest is substantial, totaling more than 15 million annual sold units of 30oz jars across the three considered brands and the 6,135 stores in the retail scanner dataset.

I split the sample into two subsamples: An "estimation" sample spanning 2010-2011 and a "policy" sample spanning 2012-2013. Households are included in the estimation

sample if they have made at least two purchases of mayonnaise in 2010-2011, resulting in 1,095 households with a total of 8,474 trips. This sample is used for demand estimation in Section 6.2 and serves as the basis for constructing counterfactual personalized coupons in Section 6.3. The policy sample includes the first trip of all households who made at least one mayonnaise purchase in 2012-2013. Of the resulting 1,891 households in the policy sample, 513 households are also in the estimation sample. I refer to the households in both subsamples as *known* consumers and refer to those only in the policy sample as *new* consumers. The policy sample is used for out-of-sample evaluation of personalized coupon policies in Section 6.3.

Each trip is associated with price, feature, and display variables of the purchased mayonnaise product. I match the household panel to the retailer scanner data to obtain prices and promotion variables from non-purchased products.[22] In addition, NielsenIQ provides a set of household characteristics. I focus on two sets of characteristics in the subsequent analysis: First, a set of five demographic variables typically considered in brand choice models (e.g., Gupta and Chintagunta, 1994; Rossi et al., 1996; Ainslie and Rossi, 1998; Horsky et al., 2006; Smith et al., 2023). These are income, family size, and indicators for employment, retirement, and single mothers. Second, a set of 23 extended consumer characteristics that also include the male and female household heads' age, indicators for the completion of high school and college, whether a household head works in a white collar occupation, three marital status indicators, two household composition indicators, three indicators for the type of residence, two indicators for TV setup, and finally indicators for owning a microwave, a garbage disposal, and a dishwasher, respectively. Appendix C provides summary statistics for all subsamples.

6.2 Demand estimates

I consider a total of four brand choice models, each estimated with either the 5 base characteristics or the 23 extended characteristics. In addition to the proposed HART and Dirichlet HART logit models, I estimate a conventional hierarchical logit model that uses a linear specification for the representative consumer (e.g., Smith et al., 2023), as well as a varying coefficient logit model (e.g. Dubé and Misra, 2023; Farrell et al., 2025). Unlike the hierarchical approaches, the varying coefficient logit model allows only for preference heterogeneity as a *deterministic* function of the consumer characteristics—i.e., all preference parameters are fully defined by the representative consumer. I use a second-order polynomial sieve as a varying coefficient function. For ease of discussion, the varying coefficient logit model is referred to as the "VC sieve" model. Note that all considered models are motivated by the same underlying latent utility model outlined in Section

[22]Prices of non-purchased products are imputed using the algorithm developed by Hitsch et al. (2021).

4—the only difference is the specification of heterogeneity across consumers.

The hierarchical Bayesian logit models are estimated with 10,500 MCMC iterations with the first 500 iterations discarded as burn-in. Appendix C provides several MCMC traceplots that indicate convergence with less than 500 iterations. The VC sieve model is estimated using the double/debiased machine learning approach of Farrell et al. (2025).[23] Throughout, the private label intercept is normalized to 0.

Since the key difference between the models is characterizing heterogeneity in preferences, I begin by comparing conditional preference estimates and refer the reader to Appendix C for the unconditional estimated coefficients. Figure 9 plots the empirical distribution of expected price coefficients for each model with both the 5 base characteristics and the 23 extended characteristics.

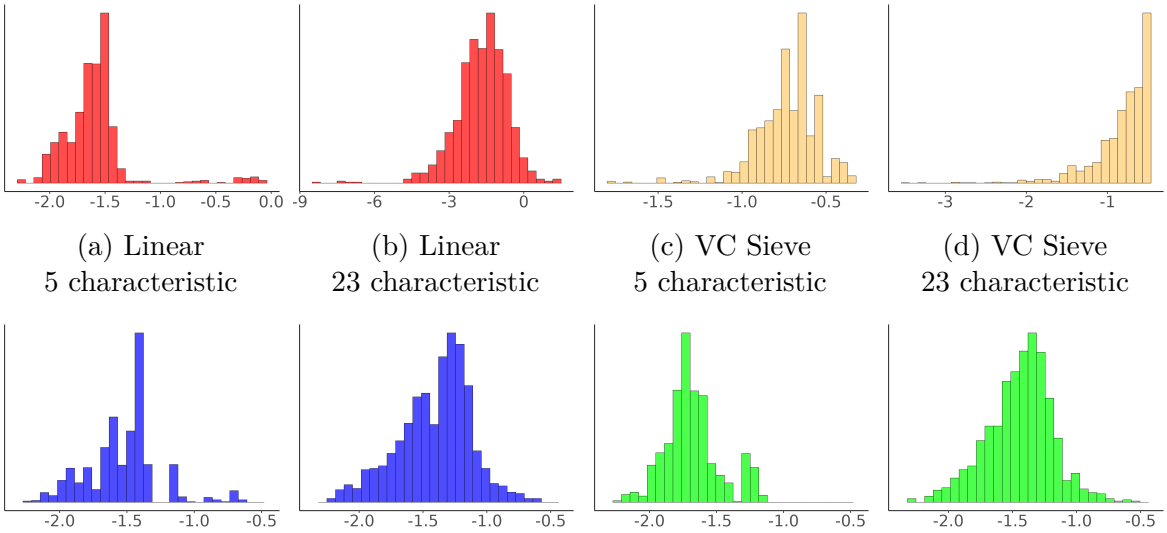
Figure 9 indicates that all models find substantial observed heterogeneity. However, the estimates differ noticeably both across models and included characteristics. First, using 5 characteristics, the hierarchical specifications result in similar distributions of observed heterogeneity with differences primarily in the shrinkage of very low coefficients. While the linear specification in panel (a) shows a mass of representative consumers with price coefficients below -0.5 and close to 0, HART estimates in panel (e) show only a few households between -1 and -0.5. Dirichlet HART, the most regularized model, estimates no households with posterior mean price coefficients above -1. Note also that the coefficients of the VC logit model indicate overall less-negative coefficients than the hierarchical specifications.

Second, when extending the characteristics, all models find richer heterogeneity in preferences as indicated by less-concentrated empirical distributions. However, while the HART and Dirichlet HART estimates in panel (f) and (h) remain within a sensible range of coefficients between -0.5 and slightly below -2, the estimates of the linear specification in panel (b) substantially disperse. For example, several values of the characteristics are associated with expected price coefficients below -6. Even if individual households were to have such extreme preferences, it seems unlikely that whole consumer segments would. Additionally, several other values of the characteristics are associated with price coefficients above 0. These results indicate that despite a relatively moderate number of characteristics, the linear model struggles with extracting meaningful observed heterogeneity from the data.

To analyze which consumer characteristics drive observed heterogeneity in the HART models, I plot the posterior variable selection probabilities in Figure 10. Rather than plotting the raw posterior probabilities for a single split, I compute the posterior over the

[23]The additional nonparametric nuisance functions needed to define the orthogonal scores are estimated with random forest estimators with varying minimum node sizes (1, 10, and 100) and combined via short-stacking as proposed in Ahrens et al. (2025b).

Figure 9: Conditional Price Coefficients



(a) Linear 5 characteristic | (b) Linear 23 characteristic | (c) VC Sieve 5 characteristic | (d) VC Sieve 23 characteristic

(e) HART 5 characteristic | (f) HART 23 characteristic | (g) Dirichlet HART 5 characteristic | (h) Dirichlet HART 23 characteristic

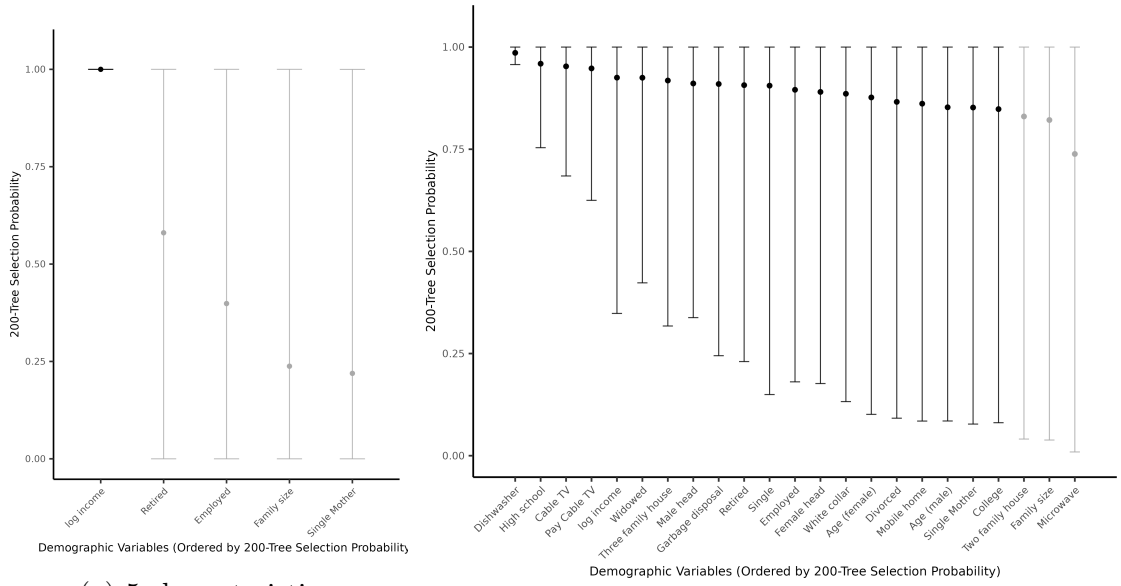
Notes: The figure shows the empirical distribution of estimated conditional price coefficients for models with 5 base characteristics (panels a, c, e, g) and 23 base characteristics (b, d, f, h). Panels (a) and (b) show results for the linear hierarchical prior, panels (c) and (d) show results for the VC Sieve model, panels (e) and (f) show results for the HART prior, and panels (g) and (h) show results for the Dirichlet HART prior. The horizontal axis differs across panels to accommodate the different ranges of the coefficients. The hierarchical Bayesian point estimates are posterior means based on 10,500 MCMC draws with 500 draws discarded as burn-in.

probability that a variable is included in a model with 200 independent tree splits.[24] This amplifies posterior differences in the variable selection probabilities and helps differentiate between variables in a densely dependent tree model. The results highlight substantial posterior selection probabilities for the extended characteristics not typically included in demand analyses (e.g., Rossi et al., 1996). For example, panel (a) shows that the representative consumer loads primarily onto income in the base characteristics, while panel (b) shows that owning a dishwasher has highest posterior selection probability with income's selection probability both lower and less concentrated.

In addition to characterizing observed heterogeneity, the hierarchical Bayesian models also allow for inference on the distribution of unobserved heterogeneity. Table 5 provides posterior estimates of the covariance matrix Σ for the linear and HART logit models with the extended characteristics. First, both models find substantial unobserved heterogeneity. Note that this is despite a richer and more flexible specification of the representative consumer. Second, the distribution of unobserved heterogeneity is more concentrated for the HART model than the linear model. For example, the marginal posterior variance of the price coefficient in the HART model is 2.83 compared to 4.27 in the linear model.

[24] In particular, let $\tau^{(k)}$ be the variable selection probability of the kth characteristic. Then the probability that the kth characteristic is included in a model with 200 splits is $1-(1-\tau^{(k)})^{200}$. Recall that 200 is the default number of trees in the HART and Dirichlet HART models.

Figure 10: Posterior 200-Tree Variable Selection Probabilities (Price)



(a) 5 characteristics

(b) 23 characteristics

Notes: The figure shows the posterior mean selection probabilities for the price sum-of-trees factor across consumer characteristics. Probabilities are computed as the posterior mean of the probability that a variable is included in a model with 200 independent tree splits—i.e., $1-(1-\tau^{(k)})^{200}$ where $\tau^{(k)}$ is the variable selection probability of the kth characteristic for a single split. Panel (a) presents results based on 5 base characteristics. Panel (b) presents results based on 23 extended characteristics. Bars indicate 90% credible intervals. Results are based on 10,500 MCMC draws with 500 draws discarded as burn-in.

Overall, the posterior mean square deviation of a household's preferences, computed as the trace of the posterior covariance matrix, is 28.7%($=\frac{19.1}{26.8}-1$) lower in the HART model than in the linear specification.

It is useful to consider an out-of-sample predictive evaluation to further understand the relative importance of observed and unobserved heterogeneity, as well as to quantify HART's improved modelling with the extended characteristics. Focussing on the prediction of a Hellmann's mayonnaise purchase, which will be the focal brand in the subsequent counterfactual personalized coupon analysis, Table 6 reports the out-of-sample R^2 for each model. Similar to the predictive exercise in Section 5, I compute the out-of-sample R^2 for each model using a 20-fold cross-validation procedure that varies the number of observed trips in the out-of-sample households.[25]

The results reaffirm advantages of HART and Dirichlet HART models over both the linear specification and the VC sieve model. Even with only 5 base characteristics, HART improves the predictive R^2 by 60.9%($=\frac{0.037}{0.023}-1$) compared to the linear specification. This improvement is amplified when extending the characteristics to 23, with HART improving

[25]For example, for predictions based on 1 purchase, the models are first estimated on the full data of 19/20ths of the households *and* the first purchase of the remaining 1/20th. The household-level posterior predictive distributions are then used to create predictions for the remaining trips of the out-of-sample respondents.

Table 5: Posterior Covariance Estimates (23 Characteristics)

	Linear Hierarchical Prior					HART				
Hellmann's	7.39	-0.80	-0.05	-0.03	-0.08	5.33	-0.66	0.04	-0.25	0.48
	(2.91)					(1.75)				
Kraft	-6.52	10.05	0.28	-0.00	0.08	-4.01	7.44	0.41	0.47	-0.63
	(1.03)	(2.53)				(0.77)	(1.69)			
Price	-0.47	1.83	4.27	0.10	-0.21	0.11	1.86	2.83	0.33	-0.23
	(1.34)	(1.19)	(0.92)			(0.64)	(0.56)	(0.53)		
Feature	-0.32	-0.04	0.39	2.38	-0.24	-0.79	1.60	0.69	1.48	-0.43
	(1.25)	(1.90)	(1.16)	(1.11)		(0.78)	(1.00)	(0.54)	(0.65)	
Display	-0.04	0.57	-0.58	-0.69	2.70	1.52	-2.47	-0.58	-0.72	2.02
	(2.37)	(2.38)	(1.32)	(0.97)	(1.22)	(0.70)	(1.05)	(0.65)	(0.36)	(0.74)

Notes: The table shows posterior means of the lower-triangular covariance matrix Σ for linear hierarchical prior and HART logit model with 23 extended characteristics, respectively. Posterior standard deviations are in parentheses. The upper triangular matrices show correlations. Results are based on 10,500 MCMC draws of the HART logit model with 500 draws discarded as burn-in.

Table 6: Out-of-Sample R^2 for Hellmann's Mayonnaise Choice

	0 purchases		1 purchase		5 purchases	
	5 Chars.	23 Chars.	5 Chars.	23 Chars.	5 Chars.	23 Chars.
Linear	0.023	-0.011	0.397	0.385	0.481	0.462
HART	0.037	0.049	0.412	0.412	0.480	0.482
Dirichlet HART	0.039	0.048	0.408	0.411	0.484	0.479
VC Sieve	0.032	0.004	0.053	0.033	0.078	0.069

Notes: The table shows out-of-sample R^2 of consumers' choice of Hellmann's mayonnaise. Results are presented for model predictions based on 5 base characteristics (5 Chars.) and 23 extended characteristics (23 Chars.). R^2 are estimated by a 20-fold cross-validation procedure that varies the number of observed trips in the out-of-sample households (0, 1, and 5 observed purchases). Choice predictions for the linear hierarchical prior, HART prior, and Dirichlet HART prior are computed as posterior means over the choice probabilities. VC Sieve uses plug-in choice probability predictions.

the predictive R^2 by 113%($=\frac{0.049}{0.023}-1$). Both the linear specification as well as the VC logit model struggle with the extended set of characteristic, with the linear specification obtaining a negative out-of-sample R^2 for previously unseen households. This points again to the substantial challenges in incorporating even moderately many characteristics in existing hierarchical approaches.

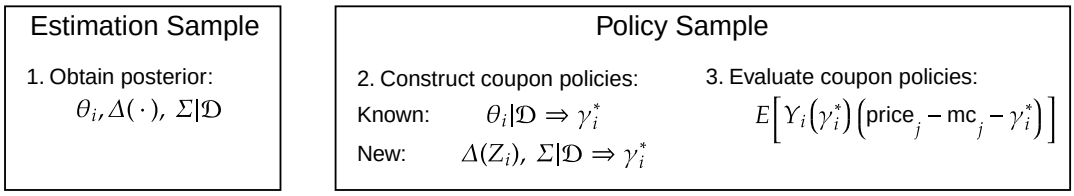
Further, the results indicate the importance of adapting to accumulating purchase history. Even for households with only one historical purchase, the hierarchical approaches obtain substantial improvements in out-of-sample predictive performance. The VC logit model, which does not explicitly adapt to consumers' purchase history, sees no comparable improvement. These results on the importance of purchase history for known consumers mirror those of Rossi et al. (1996) and Smith et al. (2023), and highlight a caveat of machine learning approaches like those considered in Dubé and Misra (2023) and Farrell et al. (2025).

As discussed in Allenby and Rossi (2019), evaluating predictive performance does not suffice to confidently judge the marketing value of competing models. Instead, a concrete decision context is needed. I consider counterfactual personalized coupons in the next section for this purpose.

6.3 Counterfactual personalized coupons

I consider a nonparametric policy evaluation approach to assess the marketing value of the proposed HART and Dirichlet HART approaches. At a high level, this evaluation procedure is conducted in three steps. In step 1, each model is estimated on the 2010-2011 estimation sample. Given the resulting estimates, step 2 constructs personalized coupon policies for all households in the 2012-2013 policy sample. Finally, in step 3, the couponing policies of each model are evaluated by estimating the corresponding expected counterfactual profits. Importantly, no *parametric* assumptions on the demand or supply are needed for estimation of the counterfactual profits. The approach thus presents an even comparison across the considered models of consumer heterogeneity. Figure 11 provides an illustration of the three steps for evaluation of the HART logit model. Throughout, I take the position of the focal manufacturer Hellmann's with the aim to improve their profits from sales of 30oz mayonnaise jars.

Figure 11: Counterfactual Personalized Coupon Evaluation Pipeline



The approach to nonparametric policy evaluation considered here is similar to the evaluation of targeting policies with observational data in Smith et al. (2023). I differ primarily in the use of a double/debiased machine learning (DML) estimator of counterfactual profits rather than an inverse propensity score (IPW) estimator as in Smith et al. (2023). As discussed below, the DML estimator is advantageous as it readily accommodates potentially many controls without imposing parametric functional forms on the propensity score.

The key benefit of the nonparametric policy evaluation approach is its evaluation of coupon policies generated by different demand models on a common metric. This allows for disentangling the *construction* from the *evaluation* of competing coupon policies in a manner not accommodated by within-model evaluation approaches as in, for example, Rossi et al. (1996). However, the nonparametric approach also limits the set of policies

that can be evaluated to those that assign only a few discrete coupon values observed in the policy sample. Researchers thus face a trade-off between the flexibility of the coupon policy construction and their ability to evaluate it. I focus on the nonparametric evaluation of such highly restricted coupon policies but emphasize that the proposed HART and Dirichlet HART models can be used to construct much richer personalization schemes.

Step 1 of the evaluation procedure is completed given the model estimates on the 2010-2011 subsample obtained in Section 6.2. To construct personalized coupon policies for the hierarchical Bayesian models (step 2), I follow the vast literature on Bayes-optimal managerial decisions (e.g., Green, 1963; Rossi et al., 1996) and define the targeting policies γ_i^* as maximizing posterior expected profits:

$$\gamma_i^* \equiv \arg\max_{\gamma \in \Gamma} \mathrm{E}_{\theta_i|\mathcal{D}} \left[\mathrm{P}(Y_i = 1|\theta_i, \mathrm{price}_1 - \gamma)(\mathrm{price}_1 - \mathrm{mc}_1 - \gamma) \right], \tag{11}$$

where Γ denotes the considered coupon values, $\mathrm{P}(Y_i = 1|\theta_i, \mathrm{price}_1 - \gamma)$ denotes the model-implied probability of purchasing Hellmann's at $\mathrm{price}_1 - \gamma$, and $(\mathrm{price}_1 - \mathrm{mc}_1 - \gamma)$ denotes the post-coupon profit associated with a sale of a Hellmann's mayonnaise jar. I set $\mathrm{price}_1 = 4.3$ as the non-discounted price and consider coupons with values $\Gamma = \{0, 0.4, 1.4\}$. Following Smith et al. (2023) who also consider Hellmann's mayonnaise in a similar time period, I set the marginal cost to $\mathrm{mc}_1 = 2.50$. Coupon policies are computed with the prices of Kraft and private label mayonnaise equal to their modal prices of 3.6 and 2.9, and with all feature and display variables equal to zero.

Note that the Bayes-optimal policies defined in (11) fully account for posterior uncertainty in a household's preferences and optimally adapt to their historical choice behavior. For a new consumer who is in the policy sample but not the estimation sample, for example, the optimal personalized coupon is found by integrating profit over the posterior distribution of observed and unobserved heterogeneity:

$$\pi(\theta_i|Z_i, \mathcal{D}) = \mathrm{E}_{\Delta(\cdot),\Sigma|\mathcal{D}} \left[\pi(\theta|\Delta(Z_i), \Sigma) \right].$$

For consumers with longer purchase histories, the posterior distribution of preferences diverges from the representative consumer as discussed in Section 2. Because no posterior distribution is available for the VC sieve model, I instead use the point estimates to construct the corresponding personalized coupon policy. This "plug-in" approach does not account for posterior uncertainty and does not adapt to a household's purchase history.

Step 3 compares the corresponding expected counterfactual profits given the person-

alized coupon policies γ^* that map individuals to discrete coupon values Γ—that is,

$$\Pi_0^{\gamma^*} \equiv \mathrm{E}\left[Y_i(\gamma_i^*)(\mathrm{price}_1 - \mathrm{mc}_1 - \gamma_i^*)\right], \tag{12}$$

where $Y_i(\gamma)$ is a household's counterfactual purchase outcome of Hellmann's mayonnaise at coupon value γ. This nonparametric policy evaluation approach is similar to the evaluation of targeting policies using experimental data in Simester et al. (2020a) and Hitsch et al. (2024) but can (also) be applied in an observational setting as in Smith et al. (2023). In this observational setting, identification of expected counterfactual profits is achieved by assuming that retail prices are as good as randomly assigned *conditional* on a set of control variables. The control variables I consider here are the extended household demographics along with month, year, and state fixed effects.[26]

The observational profit evaluation in this paper differs from that in Smith et al. (2023) by using a double/debiased machine learning (DML, Chernozhukov et al., 2018) estimator rather than an inverse propensity score (IPW) estimator. This DML estimator is obtained by noting that (12) can be rewritten as

$$\Pi_0^{\gamma^*} = \sum_{\gamma \in \Gamma} \mathrm{E}\left[Y_i(\gamma)\mathbb{1}\{\gamma = \gamma_i^*\}(\mathrm{price}_1 - \mathrm{mc}_1 - \gamma)\right]. \tag{13}$$

Note that this is a sum of weighted average potential outcomes with weights given by the product of post-coupon profit $(\mathrm{price}_1 - \mathrm{mc}_1 - \gamma)$ and an indicator equal to one if the coupon value γ and the targeted coupon value γ_i^* are equal. It thus follows that a DML estimator for (12) can be constructed using the sum the orthogonal scores corresponding to weighted average potential outcomes at each value of the coupon.[27]

In contrast to the here-considered DML estimator, the IPW estimator of Smith et al. (2023) is not based on an orthogonal score. As illustrated in, for example, Ahrens et al. (2025a), this implies worse asymptotic properties and prohibits use of flexible machine learning methods for estimation of the propensity scores. The DML estimator is advantageous as it readily accommodates potentially many controls without imposing parametric functional forms on the propensity score. Inference for the DML estimator follows directly from Theorems 3.1 and 3.2 in Chernozhukov et al. (2018) given sufficiently high-quality nuisance estimators. As suggested in Ahrens et al. (2025b), I consider a diverse set of nine different nuisance estimators that are aggregated to minimize out-of-sample MSE. See Appendix D for implementation details and model averaging weights.

[26] Appendix D details and discusses the identification assumptions for counterfactual profit estimation in the mayonnaise scanner data.

[27] See, for example, Appendix A.1.1 in Ahrens et al. (2025a) for the orthogonal score of a (single) weighted average potential outcome. Dudik et al. (2011) construct a parametric doubly robust counterfactual estimator based on the same score.

Table 7 reports the estimated expected counterfactual profits for each model. In addition, I also report the estimated counterfactual profits for uniform coupons of \$0 ("Never Coupon") and \$1.4 ("Always Coupon"), as well as a couponing policy that mimics the realized prices in the policy sample ("Existing Pricing"). Each counterfactual profit should be interpreted as the expected profit of targeting a mayonnaise-purchasing household with the corresponding coupon policy. To put estimates into perspective, recall that in the 6,135 stores in the midwest-US market of the NielsenIQ retail scanner data, there are approximately 15 million annual sales of 30oz mayonnaise jars across the 3 considered brands.

Table 7: DML Expected Counterfactual Profit Estimates (in Cents)

	New Consumers		Known Consumers		All Consumers	
	5 Chars.	23 Chars.	5 Chars.	23 Chars.	5 Chars.	23 Chars.
Personalized Coupons						
Linear	30.11	30.05	42.99	33.62	33.60	31.02
	(3.01)	(2.94)	(8.48)	(6.90)	(3.18)	(2.85)
HART	29.36	34.00	44.88	39.35	33.57	35.45
	(3.05)	(4.18)	(8.15)	(8.40)	(3.14)	(3.81)
Dirichlet HART	28.96	31.19	38.12	43.91	33.44	34.64
	(3.05)	(3.54)	(5.83)	(8.23)	(3.19)	(3.41)
VC Sieve	24.83	25.60	37.14	32.87	27.37	27.73
	(5.62)	(4.14)	(5.76)	(5.68)	(4.66)	(3.75)
Ad-hoc Coupons						
Existing Pricing	22.13		34.84		28.41	
	(9.63)		(12.43)		(7.97)	
Never Coupon	29.41		37.90		31.71	
	(3.02)		(5.83)		(2.71)	
Always Coupon	23.31		25.44		23.89	
	(1.60)		(2.75)		(1.39)	

Notes: The table shows double/debiased machine learning counterfactual profit estimates across coupon policies estimated on the 2012-2013 policy sample. Standard errors are in parentheses. Profits are in cents and per mayonnaise-purchasing household. "New Consumers" and "Known Consumers" denote households not included and included in the 2010-2011 estimation sample, respectively. "All Consumers" presents joint profits for the two groups of consumers. The top panel (Personalized Coupons) shows profits corresponding to coupon policies based on the linear hierarchical prior, HART prior, Dirichlet HART prior, and VC Sieve models, respectively. Policies were constructed using the 5 base characteristics and the extended 23 characteristics. The bottom panel (Ad-hoc Coupons) shows profits for non-personalized coupon policies. "Existing Pricing" sets Hellmann's prices to realized prices, "Never Coupon" sets all coupons to \$0, and "Always Coupon" sets all coupons to \$1.4.

The results indicate substantial potential of personalization for both new and known consumers. For example, the coupon policy based on the Dirichlet HART model using the extended characteristics results in 9 cents of additional profits per purchase for both new and known consumers compared to Hellmann's existing pricing. This increases profits by 40% and 26% for new and known consumers, respectively. If Hellmann's were to implement the Dirichlet HART targeting policy across all annual purchases of 30oz mayonnaise jars in the midwest-US market alone, this would result in an additional \$1.35 million in

expected profits.

The results also show that the HART and Dirichlet HART models improve over both the linear hierarchical specification and the VC sieve model. For example, the best performing model for new consumers is HART with the extended characteristics, improving over the best linear specification by 4 cents (13%) and the best VC sieve model by 8.4 cents (33%). For known consumers, the best performing model is HART with base characteristics, improving over the best linear specification by 2 cents (4%) and the best VC sieve model by 8 cents (21%). Overall, the counterfactual profit estimates thus indicate substantial marketing value of the proposed HART and Dirichlet HART models for designing personalized couponing policies.

In addition, note that the HART and Dirichlet HART models are more suitable for taking advantage of a richer set of household characteristics than the linear specification. While HART and Dirichlet HART models largely benefit from the extended characteristics, for example increasing profits by nearly 5 cents for new consumers when using HART, the linear specification either does not benefit or is substantially worsened. For known consumers, a linear model specification with extended characteristics results in a 9 cent reduction in profits compared to the linear specification with base characteristics, and a 1 cent reduction compared to Hellmann's existing pricing. This echoes the results of previous sections on the poor performance of the linear specification with even moderately many characteristics. HART and Dirichlet HART models thus provide a potentially fruitful avenue for future research on the use of richer household characteristics in demand models for targeting to both new and known consumers.

7 Conclusion

This paper introduces hierarchical additive regression trees (HART), a Bayesian machine learning approach for personalization. HART addresses the challenge of making optimal marketing decisions for both new consumers with limited purchase histories and known consumers whose choices have been observed over time. By integrating the flexibility of supervised machine learning with the inferential power of hierarchical Bayesian models, HART flexibly leverages a potentially large number of observed consumer characteristics to form granular initial preference predictions. As a consumer's purchase history accumulates, HART then optimally adapts to their specific preferences, providing a coherent framework for personalization throughout the customer journey. I develop an efficient Metropolis-within-Gibbs sampler that makes fully Bayesian inference practical, allowing managers to make (Bayes) optimal decisions that account for uncertainty at every stage.

I illustrate the value of HART in two discrete choice applications. In a canonical credit

card conjoint study, HART uncovers rich, nonlinear relationships between consumer demographics and preferences that are missed by conventional linear models, leading to potentially more effective targeted product design. In a CPG scanner data application, HART substantially improves out-of-sample prediction for new consumers and generates higher estimated profits in a personalized couponing exercise, particularly when leveraging an extended set of consumer characteristics. These empirical results demonstrate the practical benefits of the proposed approach for marketing decisions. In contrast to existing hierarchical approaches that worsen under even moderately many household characteristics, HART models also provide a potentially fruitful avenue for future research on the use of richer household characteristics in demand models for targeting to both new and known consumers.

This paper advances the development of hierarchical marketing models that flexibly leverage observed characteristics, and several avenues for future research remain. The empirical value of HART could be further explored across different modeling contexts, such as in richer models of consumer search (Morozov, 2023), or in cold start problems where the unit of analysis is not the consumer but, for example, a new product (Neelamegham and Chintagunta, 1999). Wider application to more datasets, including those from large-scale field experiments, would also be valuable for assessing the generalizability of HART's marketing value. Further, rich literatures on hierarchical models, Bayesian supervised machine learning, and efficient sampling provide several opportunities to extend HART. Of particular interest to tackling the cold start problem is the combination of HART with outcome-augmented approaches as in Ainslie and Rossi (1998) and Padilla and Ascarza (2021). Finally, the proposed Metropolis-within-Gibbs sampler could be further accelerated by incorporating recent advances in parallelization and faster sampling algorithms as in Pratola et al. (2014), He et al. (2019), and Bumbaca et al. (2020). Such efficiency improvements could facilitate the application of HART in very large marketing datasets, opening up further opportunities for flexible and fully Bayesian personalization.

References

Ahrens, Achim, Victor Chernozhukov, Christian Hansen, Damian Kozbur, Mark Schaffer and Thomas Wiemann (2025a). *An Introduction to Double/Debiased Machine Learning*.

Ahrens, Achim, Christian B. Hansen, Mark E. Schaffer and Thomas Wiemann (2025b). "Model Averaging and Double Machine Learning". *Journal of Applied Econometrics* 40.3, pp. 249–269.

Ainslie, Andrew and Peter E. Rossi (1998). "Similarities in Choice Behavior Across Product Categories". *Marketing Science* 17.2, pp. 91–106.

Allenby, Greg M., Neeraj Arora and James L. Ginter (1998). "On the Heterogeneity of Demand". *Journal of Marketing Research* 35.3, pp. 384–389.

Allenby, Greg M. and James L. Ginter (1995). "Using Extremes to Design Products and Segment Markets". *Journal of Marketing Research* 32.4, pp. 392–403.

Allenby, Greg M., Nino Hardt and Peter E. Rossi (2019). "Economic Foundations of Conjoint Analysis". *Handbook of the Economics of Marketing*. Edited by Jean-Pierre Dubé and Peter E. Rossi. Volume 1. Handbook of the Economics of Marketing, Volume 1. North-Holland, pp. 151–192.

Allenby, Greg M and Peter J Lenk (1994). "Modeling Household Purchase Behavior with Logistic Normal Regression". *Journal of the American Statistical Association* 89.428, pp. 1218–1231.

Allenby, Greg M. and Peter E. Rossi (1999). "Marketing Models of Consumer Heterogeneity". *Journal of Econometrics* 89.1-2, pp. 57–78.

— (2019). "Inference for Marketing Decisions". *Handbook of the Economics of Marketing*. Volume 1. Elsevier, pp. 69–149.

Ansari, Asim and Carl F. Mela (2003). "E-Customization". *Journal of Marketing Research* 40.2, pp. 131–145.

Ascarza, Eva (2018). "Retention Futility: Targeting High-Risk Customers Might Be Ineffective". *Journal of Marketing Research* 55.1, pp. 80–98.

Athey, Susan and Guido W. Imbens (2019). "Machine Learning Methods That Economists Should Know About". *Annual Review of Economics* 11.Volume 11, 2019, pp. 685–725.

Athey, Susan, Julie Tibshirani and Stefan Wager (2019). "Generalized Random Forests". *The Annals of Statistics* 47.2.

Belloni, Alexandre, Daniel L. Chen, V. Chernozhukov and C. Hansen (2012). "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain". *Econometrica* 80.6, pp. 2369–2429.

Bradlow, Eric T., Manish Gangwar, Praveen Kopalle and Sudhir Voleti (2017). "The Role of Big Data and Predictive Analytics in Retailing". *Journal of Retailing*. The Future of Retailing 93.1, pp. 79–95.

Breiman, Leo (2001). "Random Forests". *Machine Learning* 45.1, pp. 5–32.

Bumbaca, Federico (Rico), Sanjog Misra and Peter E. Rossi (2020). "Scalable Target Marketing: Distributed Markov Chain Monte Carlo for Bayesian Hierarchical Models". *Journal of Marketing Research* 57.6, pp. 999–1018.

Chakraborty, S. (2016). "Bayesian Additive Regression Tree for Seemingly Unrelated Regression with Automatic Tree Selection". *Handbook of Statistics*. Edited by Venkat N.

Gudivada, Vijay V. Raghavan, Venu Govindaraju and C. R. Rao. Volume 35. Cognitive Computing: Theory and Applications. Elsevier, pp. 229–251.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins (2018). "Double/Debiased Machine Learning for Treatment and Causal Parameters". *arXiv:1608.00060 [econ, stat]*.

Chintagunta, Pradeep, Dominique M. Hanssens and John R. Hauser (2016). "Editorial—Marketing Science and Big Data". *Marketing Science* 35.3, pp. 341–342.

Chipman, Hugh A., Edward I. George and Robert E. McCulloch (1998). "Bayesian CART Model Search". *Journal of the American Statistical Association* 93.443, pp. 935–948.

— (2010). "BART: Bayesian Additive Regression Trees". *The Annals of Applied Statistics* 4.1.

Danaher, Peter J. (2023). "Optimal Microtargeting of Advertising". *Journal of Marketing Research* 60.3, pp. 564–584.

Daviet, Remi (2020). "Applications with Limited but Diverse Data: Improving Prediction and Uncertainty Estimation with Bayesian Deep Learning".

Deshpande, Sameer K., Ray Bai, Cecilia Balocchi, Jennifer E. Starling and Jordan Weiss (2024). "VCBART: Bayesian Trees for Varying Coefficients". *Bayesian Analysis* forthcoming.

Dew, Ryan (2025). "Adaptive Preference Measurement with Unstructured Data". *Management Science* 71.5, pp. 3996–4012.

Dew, Ryan and Asim Ansari (2018). "Bayesian Nonparametric Customer Base Analysis with Model-Based Visualizations". *Marketing Science* 37.2, pp. 216–235.

Dew, Ryan et al. (2024). "Probabilistic Machine Learning: New Frontiers for Modeling Consumers and Their Choices". *International Journal of Research in Marketing* forthcoming.

Dubé, Jean-Pierre, Günter J. Hitsch and Peter E. Rossi (2010). "State Dependence and Alternative Explanations for Consumer Inertia". *The RAND Journal of Economics* 41.3, pp. 417–445.

Dubé, Jean-Pierre and Sanjog Misra (2023). "Personalized Pricing and Consumer Welfare". *Journal of Political Economy* 131.1, pp. 131–189.

Dudik, Miroslav, John Langford and Lihong Li (2011). *Doubly Robust Policy Evaluation and Learning*.

Esser, Jonas, Mateus Maia, Andrew C. Parnell, Judith Bosmans, Hanneke van Dongen, Thomas Klausch and Keefe Murphy (2025). *Seemingly Unrelated Bayesian Additive Regression Trees for Cost-Effectiveness Analyses in Healthcare*.

Farrell, Max H., Tengyuan Liang and Sanjog Misra (2021). "Deep Neural Networks for Estimation and Inference". *Econometrica* 89.1, pp. 181–213.

Farrell, Max H., Tengyuan Liang and Sanjog Misra (2025). "Deep Learning for Individual Heterogeneity".

Green, Paul E. (1963). "Bayesian Decision Theory in Pricing Strategy". *Journal of Marketing* 27.1, pp. 5–14.

Green, Peter J (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination". *Biometrika* 82.4, pp. 711–732.

Gupta, Sachin and P K Chintagunta (1994). "On Using Demographic Variables to Determine Segment Membership in Logit Mixture Models". *Journal of Marketing Research* 31.1, pp. 128–136.

Hahn, P. Richard, Jared S. Murray and Carlos M. Carvalho (2020). "Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion)". *Bayesian Analysis* 15.3.

He, Jingyu and P. Richard Hahn (2023). "Stochastic Tree Ensembles for Regularized Nonlinear Regression". *Journal of the American Statistical Association* 118.541, pp. 551–570.

He, Jingyu, Saar Yalov and P. Richard Hahn (2019). "XBART: Accelerated Bayesian Additive Regression Trees". *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 1130–1138.

Hill, Jennifer L. (2011). "Bayesian Nonparametric Modeling for Causal Inference". *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Hill, Jennifer, Antonio Linero and Jared Murray (2020). "Bayesian Additive Regression Trees: A Review and Look Forward". *Annual Review of Statistics and Its Application* 7.1, pp. 251–278.

Hitsch, Günter J., Ali Hortaçsu and Xiliang Lin (2021). "Prices and Promotions in U.S. Retail Markets". *Quantitative Marketing and Economics* 19.3-4, pp. 289–368.

Hitsch, Günter J., Sanjog Misra and Walter W. Zhang (2024). "Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation". *Quantitative Marketing and Economics* 22.2, pp. 115–168.

Horsky, Dan, Sanjog Misra and Paul Nelson (2006). "Observed and Unobserved Preference Heterogeneity in Brand-Choice Models". *Marketing Science* 25.4, pp. 322–335.

Jain, Lalit, Zhaoqi Li, Erfan Loghmani, Blake Mason and Hema Yoganarasimhan (2024). "Effective Adaptive Exploration of Prices and Promotions in Choice-Based Demand Models". *Marketing Science* 43.5, pp. 1002–1030.

Jeong, Seonghyun and Veronika Ročková (2023). "The Art of BART: Minimax Optimality over Nonhomogeneous Smoothness in High Dimension". *Journal of Machine Learning Research* 24, pp. 1–65.

Korganbekova, Malika and Cole Zuber (2023). "Balancing User Privacy and Personalization".

Lenk, Peter J. and Ambar G. Rao (1990). "New Models from Old: Forecasting Product Adoption by Hierarchical Bayes Procedures". *Marketing Science* 9.1, pp. 42–53.

Lenk, Peter and Bryan Orme (2009). "The Value of Informative Priors in Bayesian Inference with Sparse Data". *Journal of Marketing Research* 46.6, pp. 832–845.

Linero, Antonio R. (2018). "Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection". *Journal of the American Statistical Association* 113.522, pp. 626–636.

Linero, Antonio R. and Yun Yang (2018). "Bayesian Regression Tree Ensembles That Adapt to Smoothness and Sparsity". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.5, pp. 1087–1110.

Liu, Xiao (2023). "Dynamic Coupon Targeting Using Batch Deep Reinforcement Learning: An Application to Livestream Shopping". *Marketing Science* 42.4, pp. 637–658.

Logan, Brent R, Rodney Sparapani, Robert E McCulloch and Purushottam W Laud (2019). "Decision Making and Uncertainty Quantification for Individualized Treatments Using Bayesian Additive Regression Trees". *Statistical Methods in Medical Research* 28.4, pp. 1079–1093.

McCulloch, Robert and Peter E Rossi (1994). "An Exact Likelihood Analysis of the Multinomial Probit Model". *Journal of Econometrics* 64.1, pp. 207–240.

McFadden, Daniel (1974). "Conditional Logit Analysis of Qualitative Choice Behavior". *Frontiers in Econometrics*. Edited by P. Zarembka. NY: Academic Press, pp. 105–142.

Morozov, Ilya (2023). "Measuring Benefits from New Products in Markets with Information Frictions". *Management Science* 69.11, pp. 6988–7008.

Morozov, Ilya, Stephan Seiler, Xiaojing Dong and Liwen Hou (2021). "Estimation of Preference Heterogeneity in Markets with Costly Search". *Marketing Science* 40.5, pp. 871–899.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". *Journal of Economic Perspectives* 31.2, pp. 87–106.

Murray, Jared S. (2021). "Log-Linear Bayesian Additive Regression Trees for Multinomial Logistic and Count Regression Models". *Journal of the American Statistical Association* 116.534, pp. 756–769.

Murthi, B. P. S. and Sumit Sarkar (2003). "The Role of the Management Sciences in Research on Personalization". *Management Science* 49.10, pp. 1344–1362.

Neelamegham, Ramya and Pradeep K. Chintagunta (2004). "Modeling and Forecasting the Sales of Technology Products". *Quantitative Marketing and Economics* 2.3, pp. 195–232.

Neelamegham, Ramya and Pradeep Chintagunta (1999). "A Bayesian Model to Forecast New Product Performance in Domestic and International Markets". *Marketing Science* 18.2, pp. 115–136.

Padilla, Nicolas and Eva Ascarza (2021). "Overcoming the Cold Start Problem of Customer Relationship Management Using a Probabilistic Machine Learning Approach". *Journal of Marketing Research* 58.5, pp. 981–1006.

Padilla, Nicolas, Eva Ascarza and Oded Netzer (2024). "The Customer Journey as a Source of Information". *Quantitative Marketing and Economics*.

Park, Trevor and George Casella (2008). "The Bayesian Lasso". *Journal of the American Statistical Association* 103.482, pp. 681–686.

Pratola, Matthew T., Chipman Hugh A., Gattiker James R., Higdon David M., McCulloch Robert and William N. and Rust (2014). "Parallel Bayesian Additive Regression Trees". *Journal of Computational and Graphical Statistics* 23.3, pp. 830–852.

Rafieian, Omid and Hema Yoganarasimhan (2021). "Targeting and Privacy in Mobile Advertising". *Marketing Science* 40.2, pp. 193–218.

— (2023). "AI and Personalization". *Review of Marketing Research* 20. Edited by K. Sudhir and Olivier Toubia, pp. 77–102.

Ročková, Veronika and Stéphanie Van Der Pas (2020). "Posterior Concentration for Bayesian Regression Trees and Forests". *The Annals of Statistics* 48.4.

Rossi, Peter E (2014). *Bayesian Semi-Parametric and Non-Parametric Methods with Applications to Marketing and Micro-Econometrics*. NJ: Princeton University Press.

Rossi, Peter E and Greg M Allenby (2003). "Bayesian Statistics and Marketing". *Marketing Science* 22.3.

Rossi, Peter E., Greg M. Allenby and Robert MacCulloch (2009). *Bayesian Statistics and Marketing*. Reprint. Wiley Series in Probability and Statistics. Chichester: Wiley.

Rossi, Peter E., Robert E. McCulloch and Greg M. Allenby (1996). "The Value of Purchase History Data in Target Marketing". *Marketing Science* 15.4, pp. 321–340.

Rossi, Peter (2023). *Bayesm: Bayesian Inference for Marketing/Micro-Econometrics*. Comprehensive R Archive Network.

Simester, Duncan, Artem Timoshenko and Spyros I. Zoumpoulis (2020a). "Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments". *Management Science* 66.8, pp. 3412–3424.

— (2020b). "Targeting Prospective Customers: Robustness of Machine-Learning Methods to Typical Data Challenges". *Management Science* 66.6, pp. 2495–2522.

Smith, Adam N., Stephan Seiler and Ishant Aggarwal (2023). "Optimal Price Targeting". *Marketing Science* 42.3, pp. 476–499.

Sparapani, Rodney, Charles Spanbauer and Robert McCulloch (2021). "Nonparametric Machine Learning and Efficient Computation with Bayesian Additive Regression Trees: The BART R Package". *Journal of Statistical Software* 97, pp. 1–66.

Um, Seungha, Antonio R. Linero, Debajyoti Sinha and Dipankar Bandyopadhyay (2023). "Bayesian Additive Regression Trees for Multivariate Skewed Responses". *Statistics in Medicine* 42.3, pp. 246–263.

Yin, Mingzhang, Khaled Boughanmi, Anirban Mukherjee and Asim Ansari (2024). *Meta-Learning Customer Preference Dynamics for Fast Customization on Digital Platforms*. SSRN Scholarly Paper. Rochester, NY.

Yoganarasimhan, Hema, Ebrahim Barzegary and Abhishek Pani (2023). "Design and Evaluation of Optimal Free Trials". *Management Science* 69.6, pp. 3220–3240.

Zantedeschi, Daniel, Eleanor McDonnell Feit and Eric T. Bradlow (2017). "Measuring Multichannel Advertising Response". *Management Science* 63.8, pp. 2706–2728.

Appendices

A Default Priors

This appendix describes the default second-stage prior values for the hierarchical Bayesian models. Section A.1 and A.2 describe the default prior values for the HART and Dirichlet HART models. For convenience, Section A.3 provides the default prior values for the linear hierarchical prior from `bayesm` (Rossi, 2023).

A.1 HART

The HART model's priors are defined over the unconditional mean μ, the covariance matrix Σ, and the sum-of-trees parameters. Default prior values are identical to those discussed in Chipman et al. (2010), with the exception of the prior on the terminal leaf coefficients (see Section 2.3).

- **Unconditional Mean (μ):** A normal prior is used for the D-dimensional vector of unconditional means:

  $$\mu \sim \mathcal{N}(\overline{\mu}, A^{-1}).$$

  The default hyperparameter values are $\overline{\mu} = 0$ and $A = 0.01 \cdot I$.

- **Covariance Matrix (Σ):** An inverse-Wishart prior is used for the covariance matrix of unobserved heterogeneity:

  $$\Sigma \sim \mathcal{IW}(\nu, \Psi).$$

  Default hyperparameter values are $\nu = D + 3$ degrees of freedom and a scale matrix $\Psi = \nu \cdot I$.

- **Sum-of-Trees Model:** The prior for the representative unit is defined over the parameters of D independent sum-of-trees models, each being a sum of H trees.

  - **Number of Trees (H):** The default is $H = 200$.

  - **Tree Structure (R_{dh}):** The prior on the tree structure favors shallow trees. The probability of a node at depth q being non-terminal is:

    $$\alpha(1 + q)^{-\beta}.$$

Default values are $\alpha = 0.95$ and $\beta = 2$. When a node is split, a splitting variable is chosen uniformly from the available characteristics. The split point is chosen uniformly over the variable's range. For continuous variables, a grid of 100 quantile-based cutpoints is used by default.

– **Terminal Leaf Coefficients (Λ_{dh}):** The coefficients λ_{dhg} in the terminal leaves are given independent normal priors:

$$\lambda_{dhg} \sim \mathcal{N}(0, \sigma_\lambda^2)$$

The prior variance σ_λ^2 is set via $\sigma_\lambda = \frac{\tau}{\sqrt{H}}$, with a default of $\tau = 1$.

A.2 Dirichlet HART

The Dirichlet HART model augments the HART prior to induce sparsity in the dependence on unit characteristics, following the approach of Linero (2018). This is achieved by modifying the prior on the selection of splitting variables within the tree structure.

Instead of a uniform probability, the vector of selection probabilities $\tau = (\tau^{(1)}, \ldots, \tau^{(K)})$ is given a sparse Dirichlet prior:

$$(\tau^{(1)}, \ldots, \tau^{(K)}) \sim \text{Dirichlet}(\zeta/K, \ldots, \zeta/K)$$

where K is the number of available unit characteristics. The parameter ζ controls the concentration of the distribution and is itself given a hierarchical prior to be learned from the data.[28] In particular, a hyperprior is placed on a transformation of ζ:

$$\frac{\zeta}{\zeta + \rho} \sim \text{Beta}(a, b)$$

This structure is governed by three key hyperparameters:

- a and b are the shape and scale parameters for the Beta prior. A default of $a = 0.5$ and $b = 1$ as proposed in Linero (2018) induces sparsity where few variables are expected to have high selection probabilities. In contrast, setting $a = b = 1$ results in a non-sparse prior similar to the uniform selection of the standard HART model.

- The hyperparameter ρ can be used to induce additional sparsity. Its default value is the total number of characteristics, K. Reducing ρ below K encourages greater sparsity.[29]

[28] Linero (2018) also discusses setting ζ to a pre-specified fixed value or setting it via cross-validation.

[29] See also Sparapani et al. (2021) for a discussion of the Dirichlet prior hyperparameters in an application of BART to nonparametric regression.

All other priors are identical to those in the HART model.

A.3 Linear hierarchical prior

For the conventional linear hierarchical model, the representative unit is defined as $\Delta(Z_i) = \Delta^\top(Z_i - \frac{1}{n}\sum_{i=1}^{n} Z_i) + \mu$, where Δ is a $D \times K$ matrix of coefficients and unit characteristics are de-meaned for ease of interpretation (e.g., Rossi et al., 2009). Following Rossi (2023), the priors for the coefficients Δ are:

$$\text{vec}(\Delta) \sim \mathcal{N}(\overline{\delta}, A_\delta^{-1}.)$$

The default hyperparameter values are a mean of $\overline{\delta} = 0$ and a precision matrix of $A_\delta = 0.01 \cdot I$.

The default priors for the unconditional mean μ and the covariance matrix Σ are the same as for the HART and Dirichlet HART models.

B Application I: Credit Card Conjoint

This appendix provides supplementary details on the first empirical application of Section 5. Section B.1 outlines the conjoint survey. Section B.2 and B.3 provide additional demand and counterfactual estimates, respectively. Section B.4 provides additional results for the high-dimensional placebo exercise.

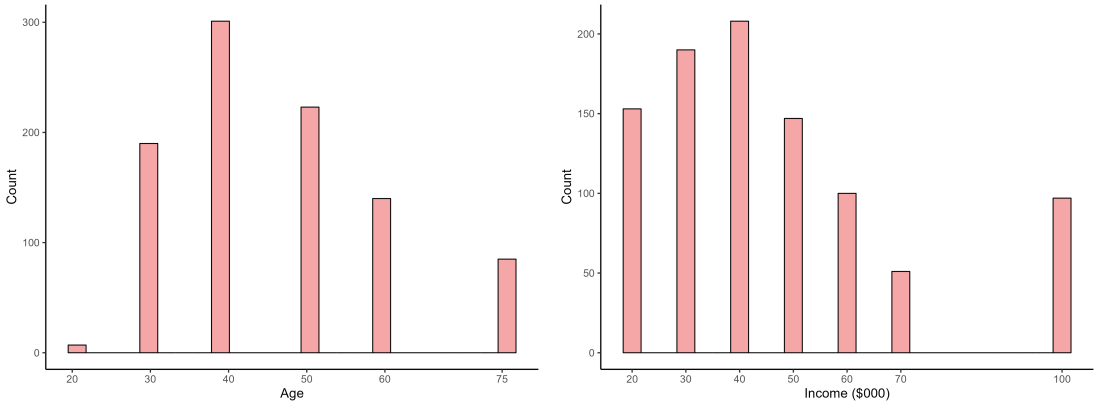
B.1 Conjoint survey of Allenby and Ginter (1995)

The details on the conjoint survey are taken from Allenby and Ginter (1995) and are included here solely for the reader's convenience.

- Interest rate: High (base), medium, low fixed, or medium variable.
- Rewards: Four reward programs (first as base), which consisted of annual fee waivers or interest rebate reductions for specified levels of card usage and/or checking account balance.
- Annual fee: High (base), medium, or low.
- Bank: Bank A (base), Bank B, or Out-of-State.
- Cash rebate: low (base), medium, or high.
- Credit limit: low (base), or high.
- Grace period: short (base), or long.

The data of Allenby and Ginter (1995) includes the three respondent characteristics age, income, and gender. The income and age characteristics are discretized in the original data, the details of which are unknown. Figure 12 shows the empirical distribution of age and income. Table 8 shows the empirical distribution of demographics.

Figure 12: Empirical Distribution of Age and Income



(a) Age (b) Income ($000)

Notes: The figure shows the empirical distribution of age (Panel (a)) and income (Panel (b)) in the conjoint survey data of Allenby and Ginter (1995). The sample totals 946 respondents.

Table 8: Distribution of Demographics

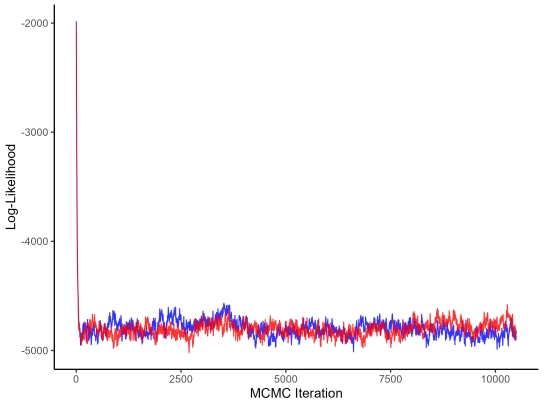
Age	Women							Men						
	Income ($000)							Income ($000)						
	20	30	40	50	60	70	100	20	30	40	50	60	70	100
20	0	0	1	0	0	0	0	3	2	0	0	0	0	1
30	12	14	19	11	6	3	6	13	36	29	17	10	3	11
40	13	22	21	15	16	4	5	13	36	60	35	22	15	24
50	14	11	9	15	9	0	4	8	**25**	27	32	23	14	32
60	22	10	12	3	3	1	4	10	13	19	16	10	10	7
75	**29**	11	3	1	0	1	0	16	10	8	2	1	0	3

Notes: The table shows the joint empirical distribution of demographics in the conjoint survey data of Allenby and Ginter (1995). Numbers indicate the number of respondents in each observed segment. Segments 1 and 2 as used in Section 5 are highlighted in bold. The sample totals 946 respondents.

B.2 Demand estimates

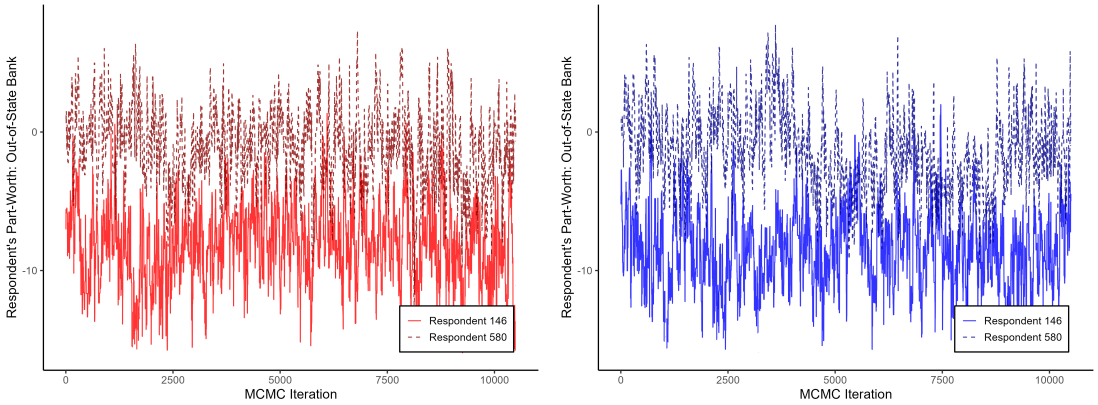
Figure 13 shows the MCMC log likelihood traceplot for the linear hierarchical and HART logit model. Figure 14 shows the MCMC traceplot for the individual part-worths for the out-of-state bank attribute. All traceplots indicate convergence occurs after less than 500 iterations.

Figure 13: MCMC Traceplot of the Log Likelihood



Notes: The figure shows traceplots of the log likelihood for the linear hierarchical logit model (in red) and HART logit model (in blue).

Figure 14: MCMC Traceplot of Individual Part-Worths for Out-of-State Bank



(a) Linear hierarchical prior (b) HART

Notes: The figure shows traceplots of the individual part-worths for the out-of-state bank attribute for the linear hierarchical logit model (Panel (a)) and HART logit model (Panel (b)). Individual part-worths for each model are evaluated for two consumers: respondent 146 (solid lines) and respondent 580 (darker dashed lines), who are members of the consumer segments 1 and 2 considered in Section 5.

Tables 9 and 10 present estimates of the individual part-worths θ_i and expected part-worths $\Delta(\cdot)$ for all credit card attributes. Table 11 presents estimates of the covariance matrix Σ for the linear hierarchical logit model (the HART covariance matrix is shown in Table 2).

Table 9: Individual Part-Worths Estimates

	Respondent 146		Respondent 580	
	Linear	HART	Linear	HART
Interest				
Medium fixed	1.02	1.16	2.47	2.24
	(1.15)	(1.15)	(1.42)	(1.39)
Low fixed	2.73	2.63	5.12	4.36
	(2.16)	(2.29)	(2.48)	(2.36)
Medium variable	0.77	1.21	4.28	3.72
	(2.41)	(2.63)	(2.44)	(2.44)
Rewards				
Rewards program 2	-0.91	-0.83	0.91	0.49
	(0.87)	(0.84)	(1.13)	(1.03)
Rewards program 3	-1.57	-1.24	0.31	0.18
	(1.24)	(1.26)	(1.54)	(1.46)
Rewards program 4	-0.59	0.19	-0.03	0.08
	(1.53)	(1.61)	(1.80)	(1.84)
Annual Fee				
Medium	0.36	-0.04	2.05	2.62
	(1.30)	(1.42)	(1.99)	(1.99)
Low	0.89	0.50	3.59	4.37
	(2.27)	(2.42)	(3.68)	(3.71)
Bank				
Bank B	-1.73	-1.89	-0.74	-0.95
	(1.24)	(1.25)	(1.48)	(1.52)
Out-of-State	-8.24	-8.61	-1.13	-1.50
	(2.85)	(2.83)	(2.78)	(2.85)
Rebate				
Medium	1.09	1.18	-1.15	-1.45
	(1.18)	(1.24)	(1.42)	(1.54)
High	0.63	0.61	-2.78	-3.14
	(2.26)	(2.20)	(2.62)	(2.87)
Credit Limit				
High	-1.68	-1.43	-2.27	-2.05
	(1.17)	(1.18)	(1.68)	(1.54)
Grace Period				
Long	1.16	1.59	0.55	0.19
	(1.82)	(1.92)	(2.12)	(2.36)

Notes: The table shows the posterior means of the individual part-worths θ_i for the linear hierarchical logit model and HART logit model of two respondents. Parentheses indicate posterior standard deviations. Respondent 146 and respondent 580 are members of the consumer segments 1 and 2 as considered in Section 5, respectively.

Table 10: Expected Part-Worths Estimates

	Overall		Age=75; Income=20; Female		Age=50; Income=30; Male	
	Linear (1)	HART (2)	Linear (3)	HART (4)	Linear (5)	HART (6)
Interest						
Medium fixed	2.57	2.56	1.92	1.48	2.29	2.62
	(0.39)	(0.53)	(0.28)	(0.38)	(0.16)	(0.30)
Low fixed	5.01	4.98	4.08	3.17	4.45	5.10
	(0.68)	(0.94)	(0.48)	(0.65)	(0.27)	(0.56)
Medium variable	3.18	3.05	2.30	1.90	2.98	3.24
	(0.78)	(0.94)	(0.57)	(0.70)	(0.29)	(0.56)
Rewards						
Rewards program 2	-0.04	-0.07	-0.05	-0.20	0.04	0.07
	(0.23)	(0.35)	(0.24)	(0.29)	(0.14)	(0.24)
Rewards program 3	-0.56	-0.60	-0.43	-0.68	-0.67	-0.12
	(0.54)	(0.59)	(0.39)	(0.40)	(0.20)	(0.33)
Rewards program 4	-0.53	-0.55	-0.51	-0.34	-0.69	-0.18
	(0.60)	(0.68)	(0.48)	(0.55)	(0.28)	(0.39)
Annual Fee						
Medium	2.19	2.15	2.58	2.02	1.97	2.31
	(0.38)	(0.57)	(0.32)	(0.41)	(0.18)	(0.33)
Low	4.19	4.17	4.78	4.28	3.69	4.21
	(0.70)	(1.03)	(0.52)	(0.74)	(0.30)	(0.58)
Bank						
Bank B	-0.39	-0.46	-0.40	-0.83	-0.47	-0.32
	(0.24)	(0.43)	(0.30)	(0.40)	(0.16)	(0.31)
Out-of-State	-3.78	-3.78	-4.40	-5.19	-3.96	-2.89
	(0.68)	(1.04)	(0.85)	(0.92)	(0.40)	(0.62)
Rebate						
Medium	1.40	1.46	1.34	1.57	1.29	1.14
	(0.24)	(0.41)	(0.30)	(0.37)	(0.16)	(0.28)
High	2.41	2.48	1.71	1.97	1.96	1.71
	(0.62)	(0.85)	(0.50)	(0.62)	(0.29)	(0.50)
Credit Limit						
High	1.15	1.16	1.16	1.42	1.05	1.41
	(0.32)	(0.60)	(0.33)	(0.46)	(0.18)	(0.37)
Grace Period						
Long	3.51	3.52	2.48	2.74	2.96	2.88
	(0.67)	(0.69)	(0.36)	(0.51)	(0.23)	(0.39)

Notes: The table shows the posterior means of the expected part-worths $\Delta(\cdot)$ for the linear hierarchical logit model and HART logit model of two respondents. Parentheses indicate posterior standard deviations. Columns (1)-(2) show overall (or: unconditional) expected part-worths. Columns (3)-(4) and columns (5)-(6) show expected part-worths for two consumer segments, older women with low income and middle-aged men with moderate income, respectively. Odd columns correspond to results for the linear hierarchical prior logit model. Even columns correspond to the HART logit model.

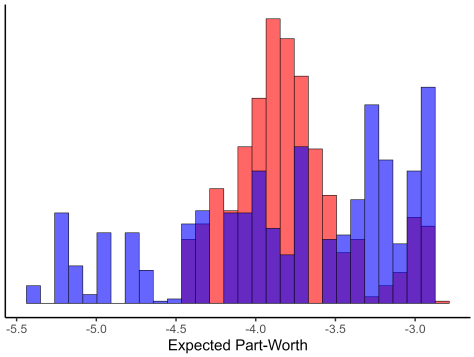
Table 11: Linear Hierarchical Logit Model Covariance Matrix Estimates

Low fixed interest	9.22	0.26	0.20	0.30	0.17	0.24
	(1.84)					
Low annual fee	3.18	14.71	0.56	0.47	0.42	0.61
	(1.71)	(2.05)				
Out-of-state bank	2.40	8.62	16.06	0.36	0.34	0.19
	(1.20)	(1.81)	(2.52)			
High cash rebate	2.79	5.37	4.29	8.70	0.42	0.68
	(1.35)	(1.66)	(1.46)	(1.68)		
High credit limit	1.29	4.11	3.51	3.14	6.47	0.43
	(0.70)	(0.81)	(1.17)	(0.75)	(0.71)	
Long grace period	1.74	5.40	1.75	4.64	2.51	5.28
	(0.94)	(1.45)	(1.13)	(1.25)	(0.65)	(1.03)

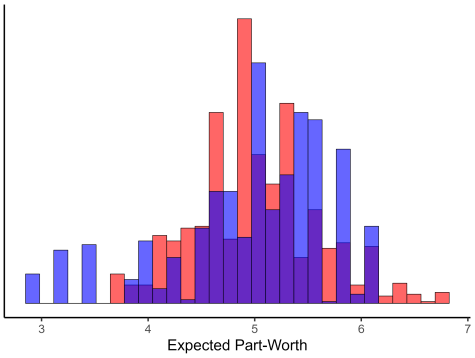
Notes: The table shows posterior means of the lower-triangular covariance matrix Σ for selected attribute levels. Posterior standard deviations are in parentheses. The upper triangular matrix shows correlations.

Figure 15 shows the posterior mean expected part-worths $\Delta(\cdot)$ for the out-of-state bank, low interest rate, and low annual fee attributes. The results indicate that the HART logit model (in blue) estimates richer heterogeneity in the expected part-worths for the out-of-state bank attribute.

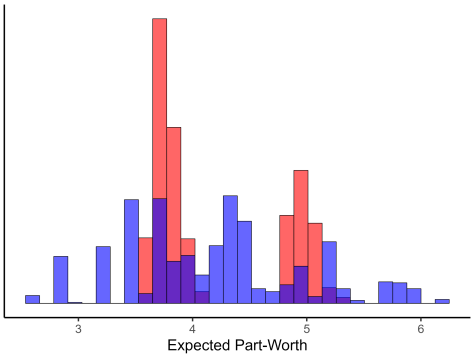
Figure 15: Posterior Mean Expected Part-Worths for all Respondents



(a) Out-of-state bank



(b) Low interest rate



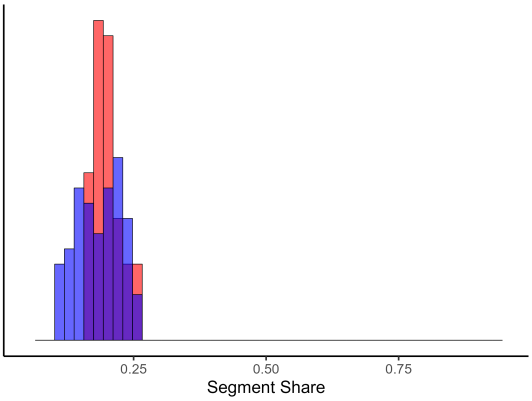
(c) Low annual fee

Notes: The figure shows the empirical distribution of posterior mean expected part-worths $\Delta(\cdot)$ for the out-of-state bank (Panel (a)), low interest rate (Panel (b)), and low annual fee (Panel (c)) attributes across all 946 respondents. Results based on the linear hierarchical logit model are in red, results based on the HART logit model are in blue.

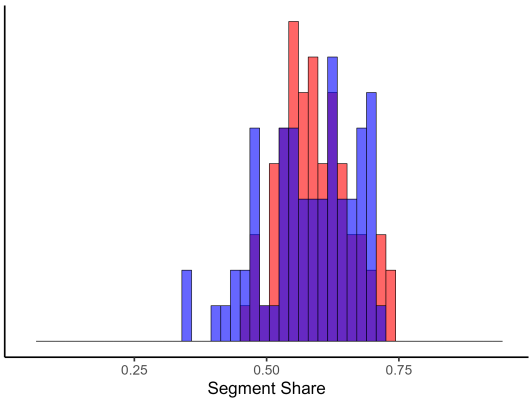
B.3 Counterfactual estimates

Figure 16 shows the empirical distribution of posterior mean counterfactual shares of out-of-state credit card offerings against a baseline in-state credit card across all 70 consumer segments observed in the data. The results indicate that the HART logit model (in blue) estimates richer heterogeneity in the counterfactual shares for the out-of-state bank attribute.

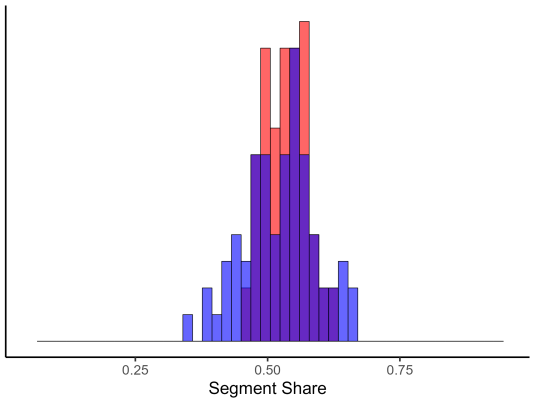
Figure 16: Posterior Mean of Counterfactual Shares for all 70 Segments



(a) Out-of-state



(b) Out-of-state & low interest rate

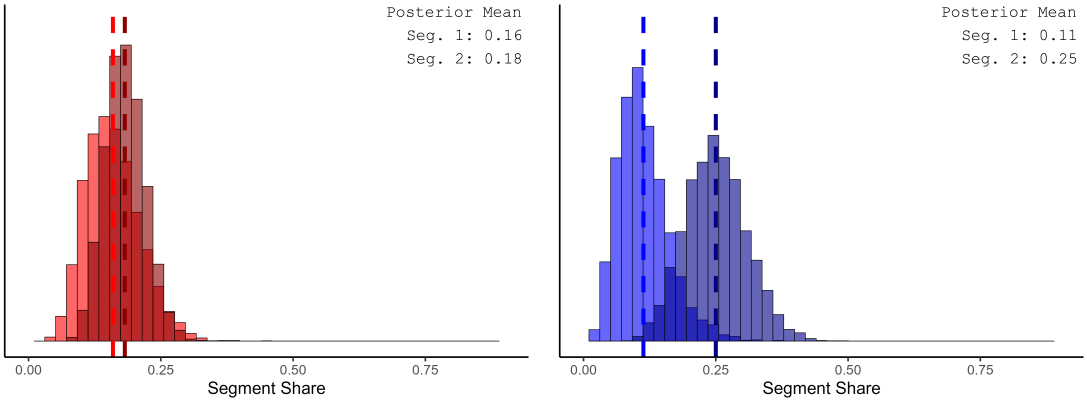


(c) Out-of-state & low annual fee

Notes: The figure shows the empirical distribution of posterior mean counterfactual shares of out-of-state credit card offerings against a baseline in-state credit card across all 70 consumer segments observed in the data. Panel (a) shows the counterfactual shares of an out-of-state credit card, Panel (b) shows the counterfactual shares of an out-of-state credit card with a low interest rate, and Panel (c) shows the counterfactual shares of an out-of-state credit card with a low annual fee. Results based on the linear hierarchical logit model are in red, results based on the HART logit model are in blue.

Figures 17 and 18 show the posterior distribution of counterfactual shares of an out-of-state credit card and an out-of-state credit card with a low annual fee for the two considered consumer segments. As in Section 5.3, the counterfactual results based on the linear approach (Panel (a)) indicate little gain to targeted advertising or other personalized approaches. In contrast, the HART logit posteriors (Panel (b)) outline the bank's opportunity to differentiate between the two segments.

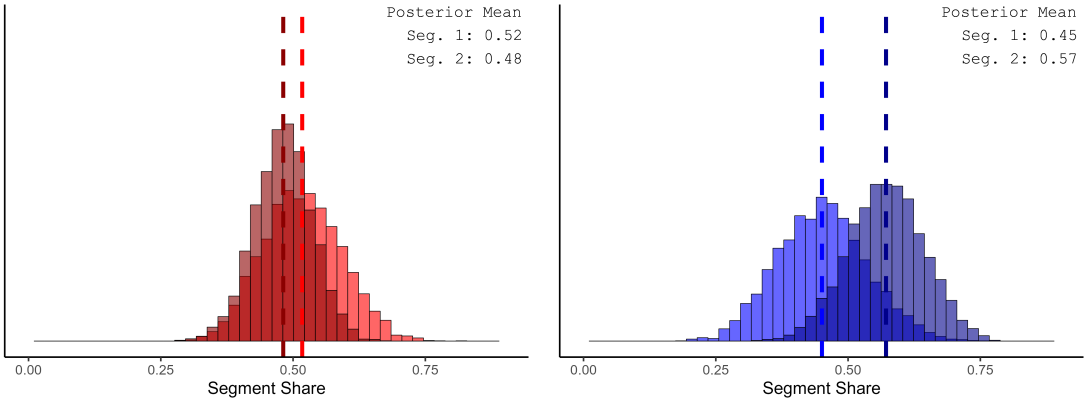
Figure 17: Counterfactual Shares for an Out-of-State Credit Card



(a) Linear hierarchical prior (b) HART

Notes: The figure shows the posterior distribution of expected counterfactual shares of an out-of-state credit card against a baseline in-state credit card. Panel (a) and (b) show results for the linear hierarchical prior and the HART prior, respectively. Expected counterfactual shares for each model are evaluated for two consumer segments: older women with low income (solid) and middle-aged men with moderate income (dark shaded).

Figure 18: Counterfactual Shares for an Out-of-State Credit Card with a Low Annual Fee



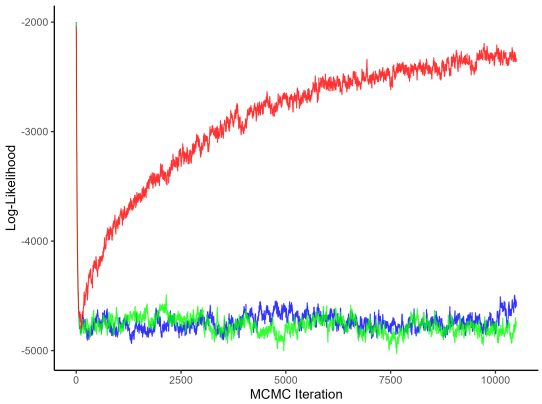
(a) Linear hierarchical prior (b) HART

Notes: The figure shows the posterior distribution of expected counterfactual shares of an out-of-state credit card with low annual fee against a baseline in-state credit card.

B.4 Robustness to high-dimensional characteristics

Figure 19 shows the MCMC log likelihood traceplots for the conventional linear, HART, and Dirichlet HART logit models when respondent characteristics are extended with 100 standard normal (irrelevant) characteristics. The results indicate convergence of the HART and Dirichlet HART logit models within less than 500 iterations. In contrast, the linear hierarchical logit model does not appear to converge within the 10,000 iterations.

Figure 19: MCMC Log Likelihood Traceplot with 100 Irrelevant Characteristics



Notes: The figure shows traceplots of the log likelihood for the linear hierarchical logit model (in red), the HART logit model (in blue), and the Dirichlet HART logit model (in green).

C Application II: Personalized Mayonnaise Coupons

This appendix provides supplementary details on the second empirical application of Section 6. Section C.1 provides summary statistics for the NielsenIQ household panel scanner data. Section C.2 provides MCMC diagnostics for the hierarchical logit models. Section C.3 provides additional demand estimates. Section C.4 details DML nuisance estimation for inference about counterfactual profits.

C.1 Summary Statistics

Table 12 summarizes the household characteristics of three considered subsamples of the NielsenIQ household panel scanner data. There are no substantial differences in the household characteristics of the three subsamples.

Table 12: Summary Statistics

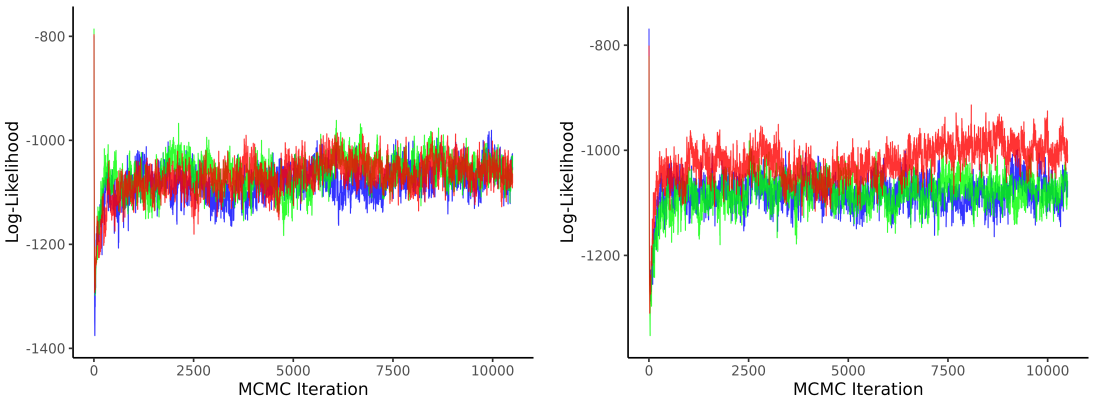
	Estimation Sample		Policy Sample New Consumers		Policy Sample Known Consumers	
	Mean	SD	Mean	SD	Mean	SD
Base Characteristics						
Log income	10.64	0.68	10.62	0.69	10.68	0.66
Family size	0.82	0.50	0.82	0.51	0.77	0.48
Employed	0.83	0.38	0.88	0.32	0.82	0.39
Retired	0.33	0.47	0.21	0.41	0.34	0.47
Single mother	0.04	0.19	0.04	0.20	0.03	0.18
Extended Characteristics						
Age (female)	53.47	16.82	47.02	18.16	53.75	17.02
Age (male)	46.16	24.71	40.51	24.23	47.05	24.48
High school	0.69	0.46	0.67	0.47	0.70	0.46
College	0.37	0.48	0.37	0.48	0.35	0.48
White collar	0.80	0.40	0.79	0.40	0.78	0.41
Widowed	0.07	0.25	0.05	0.22	0.07	0.25
Divorced	0.13	0.33	0.14	0.34	0.13	0.33
Single	0.09	0.29	0.13	0.34	0.09	0.28
Female head	0.05	0.22	0.07	0.26	0.06	0.23
Male head	0.19	0.39	0.22	0.41	0.19	0.39
Two family house	0.02	0.15	0.02	0.14	0.02	0.14
Three family house	0.07	0.26	0.10	0.30	0.06	0.25
Mobile home	0.03	0.18	0.03	0.18	0.04	0.20
Cable TV	0.40	0.49	0.39	0.49	0.39	0.49
Pay cable TV	0.24	0.43	0.28	0.45	0.23	0.42
Microwave	0.99	0.08	0.99	0.12	0.99	0.08
Garbage disposal	0.09	0.29	0.06	0.24	0.10	0.29
Dishwasher	0.72	0.45	0.71	0.46	0.71	0.45

Notes: The table shows summary statistics for the NielsenIQ household panel scanner data. "Estimation Sample" refers to the 1,095 households with at least two mayonnaise purchases in 2010-2011. "Policy Sample" refers to the 1,891 households with at least one mayonnaise purchase in 2012-2013. "Known Consumers" refers to the 513 households that are in both the estimation and policy samples. "New Consumers" refers to the 1,378 households that are in the policy sample but not the estimation sample.

C.2 MCMC Diagnostics

Figure 20 shows the MCMC log likelihood traceplots for the conventional linear, HART, and Dirichlet HART logit models, using both the base 5 characteristics and the extended 23 characteristics. Figure 21 shows corresponding traceplots of the individual price coefficients. For ease of exposition, each plot considers two consumers with visually distinct posterior distribution. Figure 22 shows corresponding traceplots of the expected price coefficients, evaluated at the consumer characteristics of the two consumers considered in Figure 21. All traceplots indicate convergence of all models within less than 500 iterations.

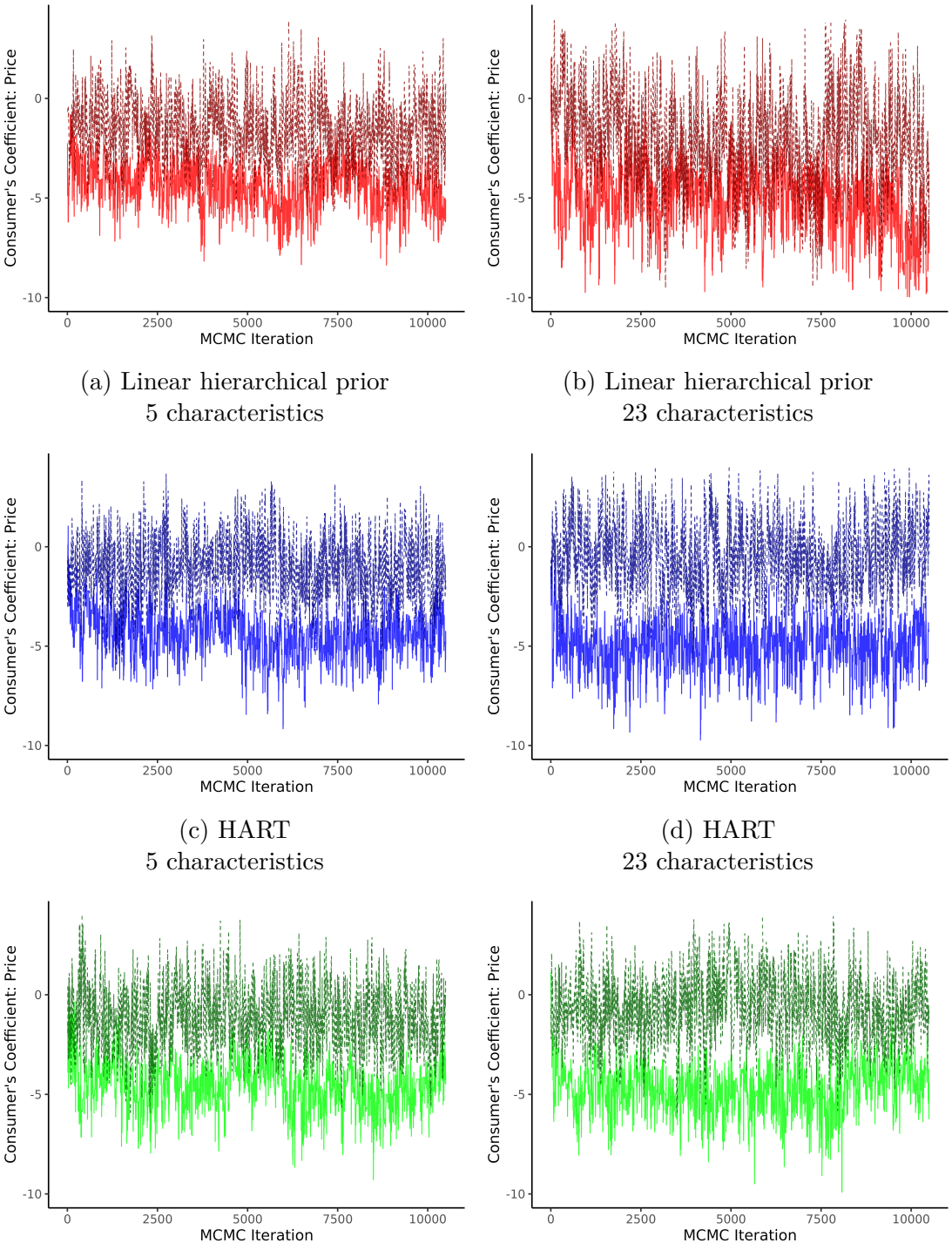
Figure 20: MCMC Traceplot of the Log Likelihood



(a) Linear (b) HART

Notes: The figure shows traceplots of the log likelihood for the linear hierarchical logit model (in red), the HART logit model (in blue), and the Dirichlet HART logit model (in green).

Figure 21: MCMC Traceplot of Individual Price Coefficients



(a) Linear hierarchical prior
5 characteristics

(b) Linear hierarchical prior
23 characteristics

(c) HART
5 characteristics

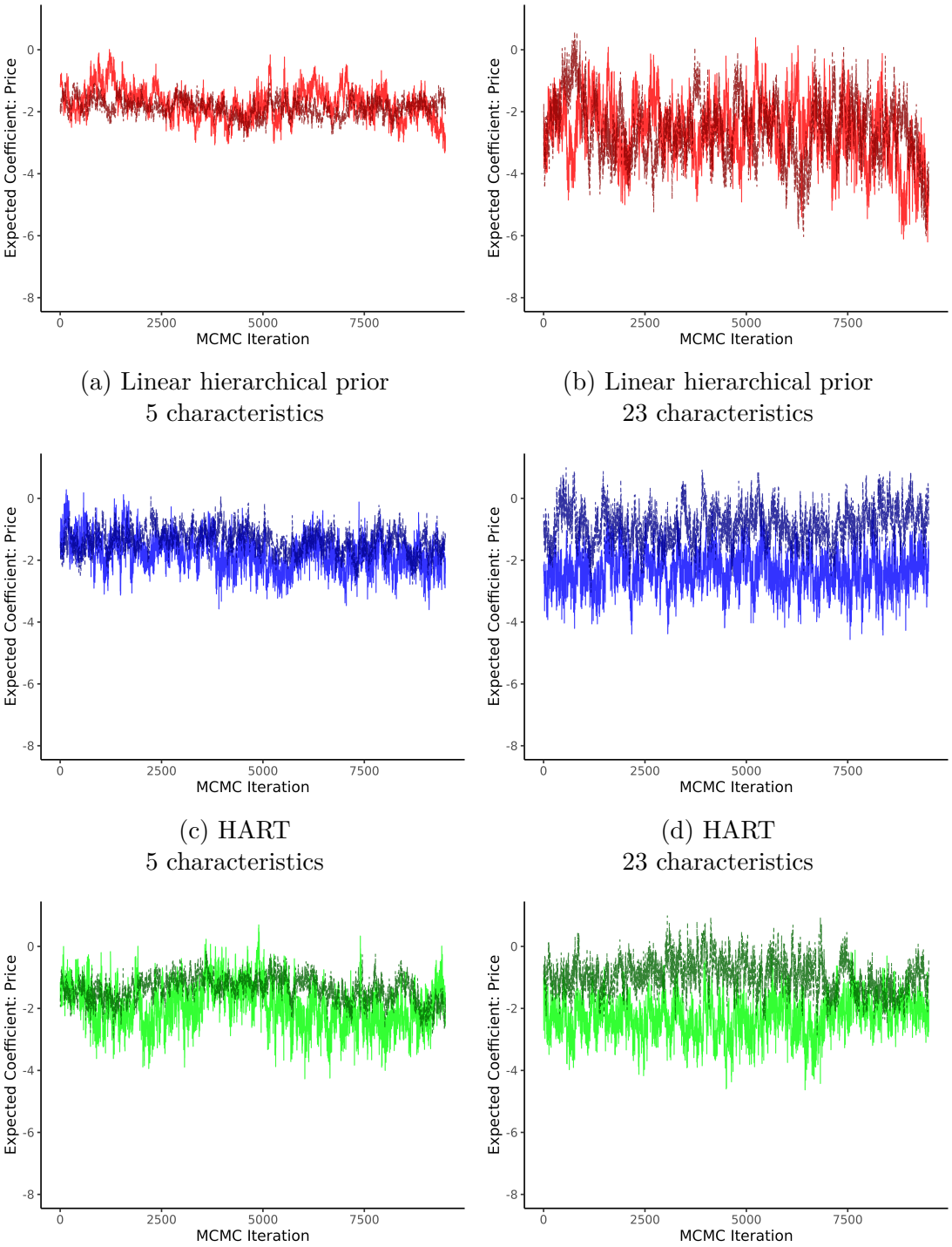
(d) HART
23 characteristics

(e) Dirichlet HART
5 characteristics

(f) Dirichlet HART
23 characteristics

Notes: The figure shows traceplots of the individual price coefficients for the linear hierarchical logit model (Panel (a) and (b)), the HART logit model (Panel (c) and (d)), and the Dirichlet HART logit model (Panel (e) and (f)). Traceplots on the left-hand side (Panels (a), (c), and (e)) show results for the base 5 characteristics. Traceplots on the right-hand side (Panels (b), (d), and (f)) show results for the extended 23 characteristics. Individual price coefficients for each model are evaluated for two consumers in the 2010-2011 estimation sample with visually distinct posterior distribution.

Figure 22: MCMC Traceplot of Expected Price Coefficients



(a) Linear hierarchical prior
5 characteristics

(b) Linear hierarchical prior
23 characteristics

(c) HART
5 characteristics

(d) HART
23 characteristics

(e) Dirichlet HART
5 characteristics

(f) Dirichlet HART
23 characteristics

Notes: The figure shows traceplots of the expected price coefficients for the linear hierarchical logit model (Panel (a) and (b)), the HART logit model (Panel (c) and (d)), and the Dirichlet HART logit model (Panel (e) and (f)). Traceplots on the left-hand side (Panels (a), (c), and (e)) show results for the base 5 characteristics. Traceplots on the right-hand side (Panels (b), (d), and (f)) show results for the extended 23 characteristics. Expected price coefficients for each model are evaluated at the characteristics of the two consumers considered in Figure 21.

C.3 Demand Estimates

Table 13 reports the unconditional mean of the estimated coefficients for all models. The results indicate similar overall demand patterns, including a stark preference in the midwest market for Kraft mayonnaise compared to both Hellmann's and private label. Notably, when extending the base characteristics to the 23 extended characteristics, the posterior uncertainty in the coefficients of the linear specification increases substantially. In contrast, posterior uncertainty about the coefficients of the HART and Dirichlet HART logit model remains stable.

Table 13: Expected Coefficient Estimates

	5 Characteristics				23 Characteristics			
	Linear	HART	Dirichlet HART	VC Sieve	Linear	HART	Dirichlet HART	VC Sieve
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hellmann's	6.20	5.60	6.43	2.79	5.37	4.71	5.19	3.75
	(1.67)	(1.05)	(1.01)	(0.32)	(2.80)	(0.84)	(0.98)	(0.31)
Kraft	8.92	8.28	9.20	3.72	8.65	7.42	7.89	4.86
	(1.41)	(1.03)	(0.91)	(0.33)	(2.86)	(0.88)	(1.03)	(0.30)
Price	-1.62	-1.54	-1.68	-0.76	-1.70	-1.40	-1.44	-0.98
	(0.49)	(0.48)	(0.43)	(0.10)	(1.34)	(0.56)	(0.52)	(0.13)
Feature	0.99	0.56	0.51	0.25	0.62	0.67	0.55	0.06
	(1.77)	(0.64)	(0.54)	(0.19)	(2.94)	(0.54)	(0.71)	(0.22)
Display	0.23	-0.02	-0.42	0.17	0.42	-0.24	-0.33	-0.03
	(0.72)	(0.54)	(0.51)	(0.18)	(2.21)	(0.64)	(0.61)	(0.20)

Notes: The table shows coefficient estimates for brand, price, and marketing mix effects. The left and right panels correspond to results using the base 5 and the extended 23 consumer characteristics, respectively. Columns (1)-(3) and (5)-(6) show posterior means of the hierarchical Bayesian models with posterior standard deviations in parentheses. Columns (4) and (8) show double/debiased machine learning point estimates and standard errors based on orthogonal scores of Farrell et al. (2025). "Hellmann's" and "Kraft" denote the corresponding brand intercepts. The private label brand intercept (not shown here) is normalized to zero throughout. "Price", "Feature", and "Display" denote the own-effect coefficients of the marketing mix variables price, and whether the corresponding brand was featured or in-store display.

Table 14 presents the posterior estimates of the covariance matrix Σ for the linear and HART logit models with the base characteristics (the covariance estimates with extended characteristics are in Table 5 in the main text). Overall, the posterior mean square deviation of a household's preferences, computed as the trace of the posterior covariance matrix, is -9%($=\frac{22.46}{24.66} - 1$) lower in the HART model than in the linear specification.

Table 14: Posterior Covariance Estimates (5 Characteristics)

	Linear					BART				
Hellmann's	6.17	-0.21	0.41	-0.19	-0.04	5.43	-0.28	0.31	-0.19	0.13
	(1.07)					(1.72)				
Kraft	-1.49	12.28	0.44	0.60	-0.14	-1.94	11.21	0.39	0.49	-0.35
	(2.05)	(3.94)				(1.41)	(2.96)			
Price	1.59	2.37	2.48	0.32	-0.05	1.09	1.96	2.38	0.34	-0.21
	(1.03)	(0.78)	(0.59)			(0.95)	(0.84)	(0.55)		
Feature	-0.61	3.34	0.83	2.37	-0.22	-0.70	2.29	0.80	2.10	-0.31
	(1.23)	(1.70)	(0.67)	(1.12)		(0.97)	(1.59)	(0.59)	(0.73)	
Display	-0.13	-0.67	-0.08	-0.38	1.36	0.37	-1.42	-0.39	-0.54	1.34
	(0.71)	(1.29)	(0.45)	(0.34)	(0.64)	(0.65)	(1.25)	(0.48)	(0.44)	(0.65)

Notes: The table shows posterior means of the lower-triangular covariance matrix Σ for linear hierarchical prior and HART logit model with 5 characteristics, respectively. Posterior standard deviations are in parentheses. The upper triangular matrices show correlations.

C.4 DML Estimates

There are six nuisance functions arising in the estimation of counterfactual profits in Section 6.3: the conditional expectation functions of profit given controls at each coupon level $(0, 0.4, 1.4)$, and the conditional probability of observing each coupon level in the sample. Following suggestions in Ahrens et al. (2025b), I estimate these nuisance functions by averaging over several machine learners via "short-stacking". This helps alleviate the problem of choosing a single machine learner.

I consider nine machine learners: ordinary least squares (OLS), lasso and ridge regression with cross-validated penalty parameters, three random forest estimators with minimum node size of 100 (high regularization), 10 (medium regularization), and 1 (low regularization), and three gradient tree boosting estimators with 500 boosting rounds and max tree depths of 1 (high regularization), 3 (medium regularization), and 5 (low regularization).[30]

Table 15 reports the short-stacking weights corresponding to each of the nine learners and target nuisance function along with out-of-sample R-squared of the short-stacked estimator. The results show that the two machine learners that are assigned the most model average weight are cross-validated lasso (CV Lasso) and the random forest with a minimum node size of 100 (Random Forest high regularization).

[30]Lasso and ridge regression are implemented using the `glmnet` package, random forests are implemented using the `ranger` package, and gradient tree boosting is implemented using the `xgboost` package. Hyperparameters are set to the respective package defaults unless otherwise specified.

Table 15: Short-Stacking Weights

	E[Y|price, X]			P(price|X)		
Price level:	\$2.9	\$3.9	\$4.3	\$2.9	\$3.9	\$4.3
Linear						
OLS	0.00	0.03	0.00	0.02	0.00	0.00
CV-Lasso	0.00	0.94	0.81	0.00	0.00	0.00
CV-Ridge	0.01	0.00	0.00	0.00	0.00	0.00
Random Forest						
High regularization	0.83	0.00	0.00	1.02	1.07	0.99
Medium regularization	0.00	0.00	0.00	0.00	0.00	0.00
Low regularization	0.00	0.00	0.07	0.00	0.00	0.00
Gradient Tree Boosting						
High regularization	0.00	0.00	0.00	0.00	0.00	0.00
Medium regularization	0.16	0.00	0.03	0.00	0.00	0.05
Low regularization	0.00	0.00	0.06	0.00	0.00	0.00
R^2	0.03	0.01	0.04	0.08	0.08	0.16

Notes: The table shows the short-stacking weights corresponding to each of the nine learners and target nuisance function along with out-of-sample R-squared of the short-stacked estimator. Short-stacking weights are restricted to be non-negative during estimation.

D Profit Estimation in Observational Data

This section describes and characterizes the counterfactual profit estimator applied in Section 6.3. Section D.1 describes the target parameter, Section D.2 states sufficient identification assumptions, and Section D.3 constructs the double/debiased machine learning (DML) estimator.

The discussion of counterfactual profit estimation here is partially based on ongoing work with Andrew Bai and Sanjog Misra.

D.1 Setup

Consider a sample of $i \in \{1, \ldots, n\}$ consumers. For each consumer, the manager observes $W_i \equiv (Y_i, p_i, X_i, Z_i)$, where Y_i is the purchase outcome of the focal product (e.g., Hellmann's) equal to one if they purchased and zero otherwise, p_i is the price of the focal product, X_i is a vector of market environment variables such as the time and place of their purchase, and Z_i is a vector of time-invariant consumer characteristics. Throughout, I consider the conventional cross-sectional setting with independent consumers. Inference generalizes naturally to clustered data.[31]

The manager is interested in estimating the expected profit value Π_0^γ associated with a given personalized coupon policy γ that maps consumer characteristics Z_i to a finite set of potential coupon levels—i.e., $\gamma : \text{support}(Z_i) \to \Gamma \equiv \{\gamma_1, \ldots, \gamma_K\}$. In particular,

$$\Pi_0^\gamma \equiv \mathrm{E}\left[Y_i(\gamma_i)(\text{price}_1 - \gamma_i - mc_1)\right], \tag{14}$$

where $Y_i(\cdot)$ denotes potential purchasing outcomes for the focal product at different coupon levels, $\gamma_i \equiv \gamma(Z_i)$ is the assigned coupon level based on consumer i's characteristics Z_i, price_1 denotes the constant base price of the focal product, and mc_1 denotes the corresponding constant marginal cost.

Expected counterfactual profits as in (14) are frequently considered in the personalization and targeting literature. For example, Rossi et al. (1996) consider maximizing counterfactual profits via targeted couponing. Smith et al. (2023) and Section 6.3 in this paper take a similar approach to targeting coupons for Hellmann's mayonnaise. The difference between (14) and the profits maximized in the abovementioned papers is that the counterfactuals $Y_i(\cdot)$ are not explicitly specified as a *parametric* demand model. For example, I leverage the choice probabilities of the proposed HART and Dirichlet HART as a parametric model for the choice counterfactuals but no such model is imposed on (14).

[31] For an illustration of DML with one-way clustered data, see, e.g., Ahrens et al. (2025a).

Not specifying a parametric demand model for defining counterfactual profits allows evaluation of policies constructed under different demand models without giving any policy an *a priori* advantage. Indeed, Equation (14) is agnostic about how policies γ are constructed. This motivates recent applications of nonparametric counterfactual policy evaluation in marketing (e.g., Dudik et al., 2011; Simester et al., 2020a; Rafieian and Yoganarasimhan, 2023; Smith et al., 2023; Hitsch et al., 2024).

A fundamental challenge in counterfactual profit evaluation is that the manager does not observe the potential outcomes $Y_i(\cdot)$ and instead sees only the realized outcomes Y_i. Π_0^γ is thus not identified without further assumptions that relate the realized outcomes to the potential outcomes.

D.2 Identification under selection on observables

Identification of counterfactual profits is simplified by rewriting (14) as

$$\Pi_0^\gamma = \sum_{\tilde{\gamma}\in\Gamma} \mathrm{E}\left[Y_i(\tilde{\gamma})\mathbb{1}\{\tilde{\gamma}=\gamma_i\}\right](\mathrm{price}_1 - \mathrm{mc}_1 - \tilde{\gamma}). \tag{15}$$

This shows that identification of Π_0^γ can be achieved through identification of the weighted average potential outcomes (wAPO) $\mathrm{E}\left[Y_i(\tilde{\gamma})\mathbb{1}\{\tilde{\gamma}=\gamma_i\}\right]$ at each coupon value $\tilde{\gamma}\in\Gamma$, where the weights $\mathbb{1}\{\tilde{\gamma}=\gamma_i\}$ indicate whether coupon value $\tilde{\gamma}$ matches the policy assignment γ_i for consumer i. In particular, scaling these wAPOs by the marginal profit constants $(\mathrm{price}_1 - \mathrm{mc}_1 - \tilde{\gamma})$ yields Π_0^γ.

A growing literature in marketing leverages (15) in combination with *randomization-by-action* experimental designs (e.g., Ascarza, 2018; Simester et al., 2020a; Yoganarasimhan et al., 2023; Hitsch et al., 2024). These approaches use random assignment of treatment actions $\tilde{\gamma}$ to identify the wAPOs. Here, in the absence of experimental data, I instead follow the approach by Smith et al. (2023) and consider identification under a selection on observables assumption.

In particular, Assumptions D.1-D.2 are sufficient for identification of $\mathrm{E}\left[Y_i(\tilde{\gamma})\right]$ for all coupon values $\tilde{\gamma}\in\Gamma$.

Assumption D.1 states that potential outcomes are independent of prices *encountered* by consumers, conditional on the encountered market environment and consumer characteristics. Importantly, even if firms are not personalizing their prices, encountered prices may still be correlated with potential outcomes due to self-selection. I.e., different consumers exhibit different shopping behavior and may thus be exposed to different prices.

Assumption D.1 (Conditional Independence of Prices; CIP)**.** $Y_i(\cdot) \perp\!\!\!\perp p_i \,|\, Z_i, X_i, \forall i \in [n]$.

Smith et al. (2023) consider an analogous selection on observables assumption. In

their application, however, only the identity of the retail chain is controlled for. In Section 6.3, I consider a more general approach where the controls include extended consumer characteristics, as well as month, year, and state fixed effects.

Assumption D.2 requires common support of prices across consumer and market characteristics. This restricts the nonparametric policy evaluation approach to counterfactual profits within the support of observed prices. Consequently, I only consider the most frequently observed coupon values in Section 6.3.

Assumption D.2 (Common Support; CS)**.** $\mathrm{P}(p_i = \mathit{price}_1 - \tilde{\gamma} \mid Z_i, X_i) > 0$ *w.p. 1,* $\forall \tilde{\gamma} \in \Gamma, i \in [n]$.

Given Assumptions D.1-D.2, standard identification arguments show identification of expected counterfactual profits (14). For completeness, this result is stated and proven in Proposition D.3.

Proposition D.3. *Suppose Assumptions D.1-D.2 hold. Then, under regularity assumptions, we have for any policy* $\gamma : \mathrm{support}(Z_i) \to \Gamma$ *that*

$$\Pi_0^\gamma = \sum_{\tilde{\gamma} \in \Gamma} \mathrm{E}\left[\mathrm{E}\left[Y_i | p_i = \mathit{price}_1 - \tilde{\gamma}, X_i, Z_i\right] \mathbb{1}\{\tilde{\gamma} = \gamma_i\}\right] (\mathit{price}_1 - \mathit{mc}_1 - \tilde{\gamma}), \tag{16}$$

Proof. Take an arbitrary $\tilde{\gamma} \in \Gamma$. We have,

$$\begin{aligned} \mathrm{E}\left[\mathrm{E}\left[Y_i | p_i = \mathrm{price}_1 - \tilde{\gamma}, X_i, Z_i\right] \mathbb{1}\{\tilde{\gamma} = \gamma_i\}\right] =& \mathrm{E}\left[\mathrm{E}\left[Y_i(\tilde{\gamma}) | p_i = \mathrm{price}_1 - \tilde{\gamma}, X_i, Z_i\right] \mathbb{1}\{\tilde{\gamma} = \gamma_i\}\right] \\ \overset{[1]}{=}& \mathrm{E}\left[\mathrm{E}\left[Y_i(\tilde{\gamma}) \mathbb{1}\{\tilde{\gamma} = \gamma_i\} | X_i, Z_i\right]\right] \\ \overset{[2]}{=}& \mathrm{E}\left[Y_i(\tilde{\gamma}) \mathbb{1}\{\tilde{\gamma} = \gamma_i\}\right], \end{aligned} \tag{17}$$

where [1] follows from Assumption D.1 and [2] follows from the law of iterated expectations. Assumption D.2 implies that the left-hand side of (17) is well-defined. Since the choice of $\tilde{\gamma}$ was arbitrary, all wAPOs $\mathrm{E}\left[Y_i(\tilde{\gamma}) \mathbb{1}\{\tilde{\gamma} = \gamma_i\}\right]$ are identified. Multiplying with the known marginal profit constants $(\mathrm{price}_1 - \mathrm{mc}_1 - \tilde{\gamma})$ and summing over all $\tilde{\gamma} \in \Gamma$ then implies identification of Π_0^γ. □

D.3 Estimation and inference

While Proposition D.3 provides a clear identification strategy, estimation presents statistical challenges. The conditional expectation functions $\mathrm{E}\left[Y_i | p_i = \mathrm{price}_1 - \tilde{\gamma}, X_i, Z_i\right]$ are high-dimensional objects unless strong functional form assumptions are imposed on the data generating process. Simply replacing these expectations with sample analogues in (16) leads to biased estimates and invalidates standard inference procedures (e.g., Ahrens

et al., 2025a). To address these issues, I construct a double/debiased machine learning (DML; Chernozhukov et al., 2018) estimator, which enables valid inference despite the presence of high-dimensional nuisance parameters.

There are two main components to every DML estimator: 1) a Neyman orthogonal score, and 2) cross-fitting. Neyman orthogonal scores for wAPOs are readily available in the literature. See, for example Appendix A.1.1 of Ahrens et al. (2025a). Adapting the notation to the present setting, the Neyman orthogonal score function for the wAPO $\omega_0^{\tilde{\gamma}} \equiv \mathrm{E}\left[Y_i(\tilde{\gamma})\mathbb{1}\{\tilde{\gamma} = \gamma_i\}\right]$ is given by

$$\begin{aligned}\psi^{\tilde{\gamma}}(W_i; \omega^{\tilde{\gamma}}, \eta) + \omega^{\tilde{\gamma}} =& \frac{\mathbb{1}\{p_i = \text{price}_1 - \tilde{\gamma}\}Y_i}{m(\tilde{\gamma}, X_i, Z_i)}\mathbb{1}\{\tilde{\gamma} = \gamma_i\} \\ &- \frac{g(\tilde{\gamma}, X_i, Z_i)\mathbb{1}\{\tilde{\gamma} = \gamma_i\}}{m(\tilde{\gamma}, X_i, Z_i)}\left(\mathbb{1}\{p_i = \text{price}_1 - \tilde{\gamma}\} - m(\tilde{\gamma}, X_i, Z_i)\right),\end{aligned} \tag{18}$$

where the nuisance parameter $\eta \equiv (m, g)$ takes true values η_0 at

$$\begin{aligned}m_0(\tilde{\gamma}, X_i, Z_i) &\equiv \mathrm{P}\left(p_i = \text{price}_1 - \tilde{\gamma} | X_i, Z_i\right), \\ g_0(\tilde{\gamma}, X_i, Z_i) &\equiv \mathrm{E}[Y_i | p_i = \text{price}_1 - \tilde{\gamma}, X_i, Z_i].\end{aligned}$$

The score in (18) combines the inverse propensity weight score with the regression score. For this reason, it is also referred to as the doubly robust score or augmented inverse propensity weighted score (e.g., Dudik et al., 2011; Rafieian and Yoganarasimhan, 2023).

To obtain the Neyman orthogonal score for counterfactual profits, note that Equation (15) implies that $\Pi_0^{\tilde{\gamma}}$ is a linear function of $\omega_0^{\tilde{\gamma}}$. It thus follows that a Neyman orthogonal score for the counterfactual profit can be constructed as

$$\psi^{\gamma}(W_i; \Pi^{\gamma}, \eta) = \sum_{\tilde{\gamma} \in \Gamma} \psi^{\tilde{\gamma}}(W_i; \omega^{\tilde{\gamma}}, \eta)\left(\text{price}_1 - \text{mc}_1 - \tilde{\gamma}\right), \tag{19}$$

where the nuisance parameters η are the same as in (18).

A DML estimator $\hat{\Pi}^{\gamma}$ for Π_0^{γ} can be constructed as the solution to the cross-fitted sample analog of (19):

$$\hat{\Pi}^{\gamma}: \quad \frac{1}{n}\sum_{l=1}^{L}\sum_{i \in I_l} \psi^{\gamma}(W_i; \hat{\Pi}^{\gamma}, \hat{\eta}_{-l}) = 0, \tag{20}$$

where $\{I_l\}_{l=1}^{L}$ is a random partition of the sample of consumers $\{1, \ldots, n\}$ into L sub-samples of approximately equal size, and $\hat{\eta}_{-l}$ denotes a cross-fitted nuisance parameter estimator computed using only consumers *not* in sub-sample l.

Chernozhukov et al. (2018) derive conditions for valid inference about Π_0^γ with the DML estimator $\hat{\Pi}^\gamma$. The central requirement, beyond standard sampling and regularity conditions, is that the nuisance parameter estimator $\hat{\eta}$ satisfies moderate convergence rate conditions. A crude sufficient convergence rate for η is that nuisance functions have $n^{-1/4}$ ℓ_2 convergence rates; see, e.g., discussion on p. C25 of Chernozhukov et al. (2018). A large and active literature characterizes settings in which variants of modern machine learning methods achieve these sufficient convergence rates, thus motivating their use for nuisance estimation. This includes, for example, versions of lasso, random forests, and neural networks (see, e.g., Belloni et al., 2012; Athey et al., 2019; Farrell et al., 2021). Because the choice of nuisance estimator is both difficult and potentially consequential in practice, Ahrens et al. (2025b) propose to average over several machine learners via "short-stacking" (see Appendix C.4 for the short-stacking estimator in the coupon application).

Assuming that estimators of the nuisance parameters converge at appropriate rates, Theorems 3.1 and 3.2 in Chernozhukov et al. (2018) imply that the sampling distribution of $\hat{\Pi}^\gamma$ is well approximated by a normal distribution as the number of consumers n grows large—i.e.,

$$\sqrt{n}\hat{\Sigma}_\gamma^{-1/2}(\hat{\Pi}^\gamma - \Pi_0^\gamma) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\hat{\Sigma}_\gamma \equiv \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l} \psi^\gamma(W_i; \hat{\Pi}^\gamma, \hat{\eta}_{-l})^2.$$

Appropriate DML standard errors for $\hat{\Pi}^\gamma$ are then given by the square root of the diagonal values of $\hat{\Sigma}_\gamma/n$. I report these standard errors for counterfactual profit evaluation in Table 7 in Section 6.3.