Bayesian Machine Learning Approach for Modeling Dynamic Consumer Preferences*

Hengxu Lin† Kohei Onzo‡ Asim Ansari§

May 31, 2026

Abstract

Understanding and predicting consumer preferences in dynamic markets is crucial for marketing decision-making. Existing methods for modeling dynamic consumer heterogeneity are either inflexible in representing individual-level preference trajectories or as in the case of recently proposed Gaussian process-based approaches, induce unimodal cross-sectional heterogeneity distributions. This prevents them from accurately capturing the growth, decline, or divergence of consumer segments over time. We introduce a novel Bayesian nonparametric machine learning framework that leverages a dependent Dirichlet process for modeling heterogeneity to overcome these methodological constraints. Our approach combines the strengths of Dirichlet processes and Gaussian processes to flexibly model a collection of temporally correlated population distributions. This structure simultaneously captures smooth individual-level preference trajectories and evolving multimodal population distributions. Extensive simulations show that our model significantly outperforms a state-of-the-art GP-based benchmark in recovering parameters and predicting choices when the evolving population distributions are multimodal and performs comparably when they are normal. An application to IRI scanner data for multiple categories spanning the Great Recession shows that our model provides more plausible price elasticity estimates, improved predictions, and granular insights into preference shifts during the recession. We finally show that coupon targeting strategies based on our model can generate higher profits.

Keywords:
Dynamic heterogeneity, Bayesian machine learning, Dependent Dirichlet process, Gaussian process, Targeting

*Data availability statement: Our empirical application involves the use of the IRI scanner panel data that are available for purchase and widely prevalent in the marketing academic community.

†Hengxu Lin is a doctoral student at Columbia Business School. Email: helin23@gsb.columbia.edu

‡Kohei Onzo is a doctoral student at Columbia Business School. Email: KOnzo27@gsb.columbia.edu

§Asim Ansari is the William T. Dillard Professor of Marketing at Columbia Business School. Email: maa48@columbia.edu

Understanding and predicting consumer preferences is critically important for marketing success. This is even more so in dynamic markets where consumer preferences and response sensitivities are governed by shifting economic conditions, recent product introductions, consumer learning, and changing individual factors. The ability to accurately capture how customers differ in their preferences and how consumer preferences change over time is crucial for making and fine-tuning segmentation, targeting, and personalization decisions. Therefore, marketing researchers have spent considerable effort in developing approaches for modeling consumer heterogeneity. Much of this research has focused on representing consumer heterogeneity in contexts where preferences are static and time-invariant. Research on dynamic consumer heterogeneity has often relied on constrained formulations with limited flexibility in capturing the complex, often multimodal, nature of real-world consumer populations where distinct segments may grow or diminish in response to external shocks.

Marketing research on static heterogeneity has modeled cross-sectional differences through finite mixtures or continuous distributions involving Gaussian distributions or their mixtures. Given the difficulty in determining the appropriate number of components in a mixture representation, some researchers have used Bayesian nonparametric methods that rely on Dirichlet process (DP) priors to automatically infer the number of latent segments or mixture components. Research on dynamic heterogeneity has modeled the temporal evolution of individual-level preferences using state-space models, hidden Markov models, and, more recently, machine learning methods such as Gaussian processes (GPs). GPs offer a flexible framework for nonparametrically modeling smooth temporal trajectories of individual-level parameters (Dew, Ansari, and Li 2020). However, despite the flexibility they afford, a key limitation remains since the cross-sectional distribution of preferences derived from a GP is necessarily Gaussian at any point in time. This prevents the model from capturing skewed, heavy-tailed, or multimodal cross-sectional population distributions, which are often present in real-world contexts.

In this paper, we address this methodological limitation by introducing a novel probabilistic machine learning framework for modeling dynamic heterogeneity based on a dependent Dirichlet

process (DDP). Dew et al. (2024) provides a comprehensive review of the growing intersection of Bayesian nonparametrics and machine learning methods in the marketing literature. Our framework combines the benefits accruing from Dirichlet processes with those that Gaussian processes offer. The use of a DDP (Quintana et al. 2022; Wade and Inácio 2025) allows us to model a collection of population distributions that are temporally correlated. Moreover, it allows us to model the individual-level parameter trajectories as smooth functions of time as in the case of GPs. Finally, it affords greater flexibility in representing the cross-sectional population distributions. This structure enables the detection of complex distributional shifts, such as the growth and decline of consumer segments or the divergence of existing ones, while simultaneously capturing temporal dynamics.

Our proposed nonparametric machine learning framework offers many advantages over existing methods in the literature. First, it nests existing models as special cases, including static DP based heterogeneity specifications and dynamic heterogeneity specifications based on GPs, and therefore it does well even when simpler heterogeneity structures suffice in a given dataset. Second, it ensures flexibility when complex forms of dynamic heterogeneity are present in the data. Third, it provides granular insights into the temporal evolution at both the market and individual levels, which is crucial for furthering the substantive understanding of market dynamics. Fourth, in contrast to GP specifications, it enables accurate recovery of multimodal and evolving distributions, leading to improved estimates of response sensitivities and key marketing quantities such as price elasticities, and potentially more effective targeting.

We train our machine learning model using a Bayesian approach and employ MCMC methods to obtain draws from the posterior distribution of the parameters and unknown quantities in the model. Our MCMC approach combines data augmentation for utilities, Gibbs sampling for DP components, and elliptical slice sampling for GP kernel parameters. This hybrid MCMC sampling strategy efficiently handles the non-conjugate aspects of the model while avoiding the tuning difficulties associated with traditional Metropolis-Hastings steps.

We conduct extensive simulation studies to highlight the efficacy of our Bayesian machine learning framework. Our simulation uses three data-generating processes (DGPs) in a discrete

choice setting that differ in their representation of the cross-sectional population distributions and their evolution: a time-varying finite mixture distribution (D1), a time-varying normal distribution (D2), and a time-varying mixture of normals (D3). We compare the performance of our model, the Dependent Dirichlet Process Heterogeneity model (DDPH), with that of the Gaussian Process Heterogeneity model (GPH), based on Dew, Ansari, and Li (2020), which is the current state-of-the-art in representing dynamic heterogeneity in marketing. The results demonstrate that DDPH significantly outperforms the GPH benchmark in recovering parameter trajectories and predicting choices when preferences are multimodal (D1 and D3), while performing comparably when the true time-varying cross-sectional population distributions are Gaussian (D2). Visualizations of posterior predictive distributions confirm that DDPH accurately captures complex, time-varying heterogeneity, whereas the GPH model oversmooths the trajectories into unimodal population distributions. Furthermore, our model provides more accurate estimates of response sensitivities and elasticities and better tracks intertemporal distribution shifts, as measured by the Wasserstein distance between population distributions at adjacent time periods. We also investigate the profitability implications of our DDPH and those of the GPH benchmark to compare the optimal coupon policy suggested by each model. Our results indicate that the DDPH yields substantially higher profits under D1 and D3 and comparable profits under D2.

We apply the model to a rich empirical setting using IRI household scanner data from 2006 to 2011, a period spanning the Great Recession. We use six consumer packaged goods categories (e.g., detergent, chips, coffee) and show that our model offers several substantive insights. It provides more plausible price elasticity estimates and reveals interesting temporal dynamics in brand preferences and promotion responsiveness. Model comparisons with the GPH benchmark indicate that the proposed DDPH model improves in-sample fit, prediction for new consumers, and forecasts of future purchases for existing consumers in most categories. Finally, we demonstrate the managerial value of our Bayesian machine learning framework through a counterfactual coupon targeting simulation in our empirical context. We find that the personalized coupon strategies derived from the DDPH model differ substantially from those implied by the GPH and are predicted to generate

greater profits than the GPH benchmark. This illustrates how the model's ability to capture evolving multimodal heterogeneity can translate into more effective marketing interventions.

In summary, this paper contributes to both marketing theory and practice through its novel methodological development. It introduces a flexible nonparametric machine learning framework for modeling dynamic preference distributions, which is novel to the marketing literature. Empirically, it provides insights into how consumer preferences evolve during economic turbulence. Managerially, it offers a robust mechanism for targeting and personalization in dynamic environments. The following section describes the relevant marketing literature and the necessary methodological background. This is followed by a description of our modeling framework and an overview of our estimation approach. We then describe the structure of the simulation studies and discuss the comparative performance of our model relative to the benchmark specification. The subsequent section presents the application to scanner data. We conclude by highlighting the advantages of our framework and propose directions for further research in the last section.

BACKGROUND METHODS AND LITERATURE REVIEW

A considerable amount of work has been done in marketing on the modeling of individual-level differences and unobserved sources of heterogeneity given their importance in segmentation, targeting, and personalization decisions. Past research has studied both static and dynamic forms of heterogeneity. We now briefly review these two streams of research as they directly inform the construction of our modeling framework.

Static Heterogeneity

Models that incorporate static heterogeneity typically assume a two-level structure. In the first stage, an individual-level formulation $\boldsymbol{y}_i \sim f(\boldsymbol{X}_i; \boldsymbol{\beta}_i)$ models the responses $\boldsymbol{y}_i$ for an individual (e.g., customer) i in terms of covariates $\boldsymbol{X}_i$ and individual-specific parameters $\boldsymbol{\beta}_i$. Cross-sectional heterogeneity is then incorporated in the second stage by assuming that the individual-level parameters $\boldsymbol{\beta}_i$ come from a common population distribution $\mathcal{P}$. In static settings, $\boldsymbol{\beta}_i$ is assumed to be

invariant over time. Different approaches have been used to model heterogeneity. Early research assumed that β_i followed a discrete distribution with a pre-specified finite number of point masses (Kamakura and Russell 1989). This, however, comes with the difficulty of figuring out the number of latent classes, which requires repeated model fits.

Continuous mixture models are also popular and researchers often assume that the β_i's come from a Gaussian distribution (Rossi, McCulloch, and Allenby 1996; Arora, Allenby, and Ginter 1998). A Gaussian population distribution allows for fine-grained heterogeneity as individuals are endowed with their own parameter values. However, the Gaussian assumption is still restrictive as it does not allow for multimodal, skewed, or heavy-tailed population distributions. Recognizing this, researchers have employed more expressive population distributions by assuming that the β_i's are drawn from a finite mixture of Gaussians (Allenby, Arora, and Ginter 1998) to allow for multimodal population distributions. While such a specification is very flexible, it requires selecting the appropriate number of Gaussian components, which is difficult to ascertain a priori. To retain flexibility and at the same time seamlessly and automatically infer the shape of the population distribution, some researchers have relied on Bayesian nonparametric approaches. In particular, Dirichlet process (DP) priors have been used for the automatic determination of the appropriate model complexity for a given dataset. We briefly review DP priors as they form an important building block for our proposed framework described in the next section.

Dirichlet Process Heterogeneity DP priors have been widely adopted in marketing applications (Ansari and Mela 2003; Kim, Menzefricke, and Feinberg 2004; Wedel and Zhang 2004; Braun et al. 2006; Li and Ansari 2014; Bruce 2019; Boughanmi and Ansari 2021; Onzo and Ansari 2025). In this nonparametric approach, the individual-level parameters β_i are assumed to come from an unknown population distribution $\mathcal{P}$ such that $\beta_i \sim \mathcal{P}$.[1] Uncertainty over $\mathcal{P}$ is specified using a Dirichlet process prior, i.e., $\mathcal{P} \sim \mathcal{DP}(\alpha, \mathcal{H})$. The DP is defined via its two parameters: the concentration parameter $\alpha > 0$ and the base distribution $\mathcal{H}$, which is typically assumed to be a

[1] This is in contrast to a parametric assumption that the coefficients come from a distribution belonging to a known family of distributions indexed by a finite number of unknown parameters.

(multivariate) Gaussian in the marketing literature, with an unknown mean and covariance matrix.

Draws from the DP are almost surely discrete distributions. For our purposes, the constructive definition of the DP in terms of its stick-breaking representation (Sethuraman 1994) is useful, as the DDP can also be constructed based on this characterization. According to this perspective, $\mathcal{P} \sim \mathcal{DP}(\alpha, \mathcal{H})$ can be almost surely represented as an infinite mixture of point masses,

$$\begin{aligned} \mathcal{P} &= \sum_{k=1}^{\infty} \underbrace{\left\{ V_k \prod_{h<k} (1 - V_h) \right\}}_{\pi_k} \delta_{\tilde{\boldsymbol{\beta}}_k}(\cdot), \\ V_k &\sim \text{Beta}(1, \alpha), \ \tilde{\boldsymbol{\beta}}_k \sim \mathcal{H}, \quad \forall k \in \{1, 2, \dots\}. \end{aligned} \tag{1}$$

The atoms $\{\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \dots, \tilde{\boldsymbol{\beta}}_k, \dots\}$ are realizations from the base distribution $\mathcal{H}$, the probability weights $\{\pi_1, \pi_2, \dots, \pi_k, \dots\}$ are associated with those atoms, and $\delta_{\tilde{\boldsymbol{\beta}}_k}(\cdot)$ is a Dirac measure concentrated at $\tilde{\boldsymbol{\beta}}_k$. The $\tilde{\boldsymbol{\beta}}_k$'s can be intuitively viewed as cluster-specific parameters, as in the case of finite mixtures, although the clusters cannot be interpreted as segments in the traditional marketing sense. The weights π_k are determined by Beta-distributed random variables V_k. The concentration parameter $\alpha > 0$ governs how similar the realized distributions $\mathcal{P}$ are to the base distribution $\mathcal{H}$. As α gets large, $\mathcal{P}$ mimics the base distribution, whereas it tends to place most of its weight on a few atoms when α is small. When the mean and the covariance matrix of the normal base distribution $\mathcal{H}$, as well as α are inferred from the data, the DP offers a seamless mechanism to automatically infer the shape of the distribution $\mathcal{P}$.

While the approaches described above capture static heterogeneity, they do not account for market dynamics, which requires the incorporation of dynamic heterogeneity.

Dynamic Heterogeneity

When dynamic heterogeneity is incorporated, the first-stage individual-level model $\boldsymbol{y}_{it} \sim f(\boldsymbol{X}_{it}; \boldsymbol{\beta}_{it})$ is specified in terms of time-varying individual-level parameters $\boldsymbol{\beta}_{it}$ that capture parameter-driven dynamics. Such dynamics reflect changes in consumer preferences that can

arise from multiple sources. Gordon, Goldfarb, and Li (2013) study how parameters evolve with the market conditions that are powered by the booms and busts of the economic cycle. Other sources could include changes in one's financial conditions stemming from a layoff or a promotion. Jedidi, Mela, and Gupta (1999) model how increased exposure to marketing activities can shift consumer tastes and impact response sensitivities. Consumer learning can also result in preference evolution, as repeated choices and new experiences can update a consumer's valuation of product attributes and preferences for products.

Various methods have been used in the marketing literature to represent the dynamic heterogeneity in the time-varying parameters. One prominent approach is based on state-space specifications. Examples include Lachaab et al. (2006); Sriram, Chintagunta, and Neelamegham (2006); Guhl et al. (2018); Naik et al. (2015), which employ continuous state space models involving a second-stage transition mechanism that captures how the state (i.e., $\boldsymbol{\beta}_{it}$) evolves over time. Other approaches include hidden Markov models that operate on discrete states where the parameters $\boldsymbol{\beta}_{it}$ are state-specific and these discrete states evolve according to a Markov chain (Netzer, Lattin, and Srinivasan 2008). Considering the limited flexibility of previous state-space specifications, recent efforts have used Gaussian processes to flexibly model the time-varying trajectories for the individual-level parameters (Dew and Ansari 2018; Dew, Ansari, and Li 2020; Dew et al. 2024). As GPs are also an integral component of our modeling framework, we briefly review below how they have been used to represent dynamic heterogeneity.

Gaussian Processes When GPs are used for dynamic heterogeneity, the individual-level preference parameters are modeled as continuous functions of calendar time t. Thus, each coefficient in the individual-level model can be represented as a separate function $\beta_i(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$, where $\mathbb{T}$ is a temporal space. A Gaussian process, $\mathcal{GP}(\mu(\cdot), K(\cdot,\cdot))$, acts as a prior over this function and captures the heterogeneity in the individual-level trajectories. The GP specification allows the individual-level trajectories to flexibly deviate from the population mean.

The GP prior is characterized by a mean function $\mu(\cdot) : \mathbb{T} \rightarrow \mathbb{R}$ and a covariance function or

kernel $K(\cdot,\cdot) : \mathbb{T}\times\mathbb{T} \to \mathbb{R}$. The mean function $\mu(\cdot)$ represents the dynamics in the population mean. It models the average, or "typical" trajectory for all individuals in the population, imposing a general structure on the population's behavior over time. Instead of assuming that an individual's parameter trajectory strictly mimics the average temporal pattern, the covariance kernel $K(\cdot,\cdot)$ governs how the individual-specific trajectories flexibly deviate from the mean trajectory. Through its specific functional form and hyperparameters, the kernel encodes temporal dependencies, dictating the smoothness and magnitude of these individual-level variations. Together, the mean function and the kernel allow the model to borrow strength and leverage information across individuals and time.

GP realizations, $\beta_i(\cdot)$, have the property that any finite collection of function values $\beta_i(\boldsymbol{t})$ at time points $\boldsymbol{t} = (t_1, \ldots, t_T)' \in \mathbb{T}^T$ with $T < \infty$ follows a T-dimensional Gaussian distribution, i.e.,

$$\beta_i(\boldsymbol{t}) = (\beta_i(t_1), \ldots, \beta_i(t_T))' \sim \mathcal{N}\Big(\mu(\boldsymbol{t}), K(\boldsymbol{t},\boldsymbol{t}')\Big), \tag{2}$$

where $\mu(\boldsymbol{t}) = \mathbb{E}[\beta_i(\boldsymbol{t})]$ is a T-dimensional mean vector and $K(\boldsymbol{t},\boldsymbol{t}') = \mathrm{Cov}(\beta_i(\boldsymbol{t}), \beta_i(\boldsymbol{t}'))$ is a $T \times T$ covariance matrix. This also implies that for any given time point t, the function value $\beta_i(t)$ is a random variable that follows a Gaussian distribution. To ensure modeling flexibility, the population mean function $\mu(\cdot)$ is typically assumed to come from another Gaussian process, as opposed to being specified parametrically. This results in a hierarchical GP setup for the individual trajectories (Dew, Ansari, and Li 2020). Thus, dynamic heterogeneity for the p-th coefficient $\beta_i^{(p)}(\cdot)$ is characterized as[2]

$$\begin{aligned} \beta_i^{(p)}(\cdot) &\sim \mathcal{GP}\Big(\mu^{(p)}(\cdot), K_{\boldsymbol{\theta}^{(p)}}(\cdot,\cdot)\Big), \\ \mu^{(p)}(\cdot) &\sim \mathcal{GP}\Big(\mathbf{0}, K_{\boldsymbol{\theta}_0}(\cdot,\cdot)\Big), \end{aligned} \tag{3}$$

where $K_{\boldsymbol{\theta}^{(p)}}(\cdot,\cdot)$ is the covariance kernel of the Gaussian process for the p-th coefficient, specified in terms of the hyperparameters in $\boldsymbol{\theta}^{(p)}$. These hyperparameters have additional hyperpriors associated with them. Similarly, the GP associated with the population mean function has a mean

[2] We now have an explicit coefficient index p in the superscript of $\beta_i(\cdot)$.

function set to $\mathbf{0}$ and a covariance kernel $K_{\theta_0}(\cdot,\cdot)$ parametrized by θ_0.

Gaussian processes allow for greater flexibility in handling the evolution of customer-level trajectories when compared to some previously used state-space formulations. However, GPs are still restrictive as the population-level heterogeneity distribution at any time period t is a Gaussian distribution. This implies that one cannot accommodate multimodal, heavy-tailed, or skewed cross-sectional heterogeneity distributions as well as situations where customer segments emerge, grow, or diminish over time. This is a key limitation of the GPH framework for dynamic heterogeneity that we specifically address in this paper.

MODEL

While our Bayesian machine learning framework can be applied in any modeling context where the incorporation of dynamic heterogeneity is important, we illustrate it through a discrete choice model with evolving customer preferences.

Individual-Level Model

We begin by considering a standard utility-based specification. For each consumer i making a discrete choice decision among C alternatives (e.g., brands), the utility for alternative c on the j-th observation at calendar time t is written as:

$$u_{ijtc} = \boldsymbol{x}'_{ijtc}\boldsymbol{\beta}_{it} + \varepsilon_{ijtc}. \tag{4}$$

Here, $\boldsymbol{x}_{ijtc}$ is a P-dimensional vector, which contains observed brand characteristics (e.g., price, promotions) and $C-1$ brand-specific dummy variables for the brand intercepts, $\boldsymbol{\beta}_{it}$ is a P-dimensional vector of preference parameters in period t that captures consumer i's sensitivity to these characteristics and the $C-1$ corresponding brand intercepts that represent the consumer's intrinsic preferences for the brands (relative to the baseline brand whose intercept is normalized to zero). The idiosyncratic component ε_{ijtc} represents factors that influence the consumer's choice but are

unobservable to the researcher. We assume that these are i.i.d. standard normal across observations and alternatives. Consumers select the alternative y_{ijt} that provides the highest utility, i.e., $y_{ijt} = \text{argmax}_c \, u_{ijtc}$. Having specified the individual-level model, we now describe our dynamic heterogeneity specification.

Dynamic Heterogeneity via Dependent Dirichlet Process

We use a dependent Dirichlet process (DDP) (MacEachern 1999, 2000; Quintana et al. 2022; Wade and Inácio 2025) to flexibly capture dynamic heterogeneity over the coefficient vector $\boldsymbol{\beta}_{it}$ in the utility function in Equation 4. A DDP is a class of Bayesian nonparametric methods, which have gained significant attention in the machine learning literature recently (Ascolani et al. 2022; Riva-Palacio, Leisen, and Griffin 2022; Ascolani et al. 2023; Finocchio and Schmidt-Hieber 2023; Feldman and Kowal 2024; Rosa and Rousseau 2025; Um, Sweet, and Adhikari 2025). Under the DDP specification, the cross-sectional population distribution at each time period t marginally follows a Dirichlet process. This differs from the normal distribution assumption that is implied by a GP prior, as in Dew, Ansari, and Li (2020). Thus, the shapes of population distributions can flexibly vary over time. We now describe a DDP in detail before summarizing the complete structure of our model.

Dependent Dirichlet process A dependent Dirichlet process (DDP) specifies a prior over a collection $\{\mathcal{P}_t\}_{t\in\mathbb{T}}$ of distributions indexed by a covariate (in our context, time t) that is marginally DP distributed for each value of the index. Our DDP setup integrates the two Bayesian nonparametric components that were described in the previous section: the Dirichlet process (DP), which induces clustering over the individual-level parameter trajectories and the Gaussian process (GP), which generates these trajectories.

Stick-breaking representation A DDP extends the stick-breaking representation of the DP to the dynamic context by utilizing an infinite mixture over trajectories. When a time-indexed collection of distributions is DDP distributed, i.e., $\{\mathcal{P}_t\}_{t\in\mathbb{T}} \sim \mathcal{DDP}(\alpha, \mathcal{H})$, the distributions $\mathcal{P}_t$'s can almost

surely be written as

$$\mathcal{P}_t = \sum_{k=1}^{\infty} \underbrace{\left\{ V_k \prod_{h<k} (1 - V_h) \right\}}_{\pi_k} \delta_{\tilde{\boldsymbol{\beta}}_k(t)}(\cdot), \quad \forall t \in \mathbb{T}$$
$$V_k \sim \text{Beta}(1, \alpha), \; \tilde{\boldsymbol{\beta}}_k(\cdot) \sim \mathcal{H}, \quad \forall k \in \{1, 2, \dots\}. \tag{5}$$

Here, the k-th mass point, $\tilde{\boldsymbol{\beta}}_k(\cdot) : \mathbb{T} \to \mathbb{R}^P$, is a vector-valued function and an independent realization of a stochastic process whose law is $\mathcal{H}$. We specify the base measure of the DDP as a (P-dimensional) multi-output Gaussian process, $\mathcal{H} = \mathcal{GP}\Big(\boldsymbol{\mu}(\cdot), K_\Theta(\cdot, \cdot)\Big)$ over the P coefficient trajectories. The mean function $\boldsymbol{\mu}(\cdot) : \mathbb{T} \to \mathbb{R}^P$ is vector-valued, and the kernel $K_\Theta(\cdot, \cdot) : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{P \times P}$ is a matrix-valued function parameterized by Θ. For simplicity, we assume that the P output components of this multi-output GP are independent, which means that the trajectories $\tilde{\beta}_k^{(p)}(\cdot)$ and $\tilde{\beta}_k^{(p')}(\cdot)$ for two distinct coefficients, $p \neq p'$, are independent. This implies that for each coefficient p, we have $\tilde{\beta}_k^{(p)}(\cdot) \sim \mathcal{H}^{(p)} = \mathcal{GP}\Big(\mu^{(p)}(\cdot), K_{\boldsymbol{\theta}^{(p)}}(\cdot, \cdot)\Big)$, where $\mu^{(p)}(\cdot) : \mathbb{T} \to \mathbb{R}$ is a mean function and $K_{\boldsymbol{\theta}^{(p)}}(\cdot, \cdot) : \mathbb{T} \times \mathbb{T} \to \mathbb{R}$ is a kernel function. In other words, the multi-output Gaussian process $\mathcal{H}$ reduces to P independent single-output Gaussian processes $\{\mathcal{H}^{(p)}\}_{p=1}^P$.

Different vector-valued functions $\tilde{\boldsymbol{\beta}}_k(\cdot)$ associated with the different mass points indexed by k can be considered as trajectories belonging to different clusters, and the base Gaussian process $\mathcal{H}$ plays a similar role to the base distribution of a Dirichlet process. The probability weights π_k are determined by Beta distributed random variables V_k, as in the DP. Instead of using a fully general DDP specification, for simplicity, we assume that the weights π_k are fixed across time, resulting in a single-weights DDP (Quintana et al. 2022). The mean function and the covariance kernel of the base GP for coefficient p are specified as follows.

Mean functions We follow Dew, Ansari, and Li (2020) and place an independent GP prior on the mean function $\mu^{(p)}(\cdot) \sim \mathcal{GP}\Big(\mathbf{0}, K_{\boldsymbol{\theta}_0}(\cdot, \cdot)\Big)$ for each coefficient index p, resulting in a hierarchical setup. This allows the mean trajectory to be learned flexibly from the data.

Kernel functions We use a Matérn kernel (Porcu et al. 2024), which is widely used in both statistics and machine learning, to specify the nature of the temporal dependence in the parameter trajectories. The Matérn kernel is given by

$$K_{\boldsymbol{\theta}}(t,t') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\, \frac{\|t-t'\|}{l}\right)^{\nu} B_\nu \left(\sqrt{2\nu}\, \frac{\|t-t'\|}{l}\right), \tag{6}$$

which is parameterized by $\boldsymbol{\theta} = (\kappa, \sigma^2)$, where $\kappa = \sqrt{5}/l$. The parameter ν is positive and controls the smoothness of the function, and the Matérn kernel converges pointwise to the squared exponential (RBF) kernel as $\nu \to \infty$. We define l as a positive length-scale parameter and σ^2 as the variance. The value of the length-scale parameter l dictates how quickly the correlation between two points decreases as the distance between them increases. A small length scale means that a data point has a very local influence and the correlation drops off quickly. In contrast, a large length scale implies that a data point has a far-reaching influence and the correlation between points decays gradually over a larger distance. Finally, $\Gamma(\cdot)$ denotes the gamma function, and $B_\nu(\cdot)$ is a modified Bessel function. As not all parameters in the Matérn kernel are identifiable, researchers often set ν to half-integer values such as $\nu = \frac{1}{2}$, $\frac{3}{2}$, or $\frac{5}{2}$, which yield closed-form kernels with convenient interpretations (Zhang 2004; Dew, Ansari, and Li 2020). In this paper, we set $\nu = \frac{5}{2}$, resulting in

$$K_{\boldsymbol{\theta}}(t,t') = \sigma^2 \left(1 + \kappa\|t-t'\| + \frac{\kappa^2\|t-t'\|^2}{3}\right) \exp(-\kappa\|t-t'\|). \tag{7}$$

The Matérn kernel strikes a balance between flexibility and interpretability, making it particularly well suited for modeling dynamic, heterogeneous consumer preferences.

Individual-level trajectories Individuals are clustered based on their coefficient trajectories. Let $z_i \in \{1, 2, \dots\}$ denote the cluster to which individual i belongs. Then, we have

$$\boldsymbol{\beta}_i(\cdot) = \tilde{\boldsymbol{\beta}}_{z_i}(\cdot), \quad z_i \sim \sum_{k=1}^{\infty} \pi_k \delta_k(\cdot), \tag{8}$$

where the probability weights $\pi_k = V_k \prod_{h<k}(1 - V_h)$ are obtained from the stick-breaking prior. The parameter vector $\boldsymbol{\beta}_{it} = \boldsymbol{\beta}_i(t) = \left(\beta_i^{(1)}(t), \ldots, \beta_i^{(P)}(t)\right)'$ of individual i at any time point t is given by the function value of the individual's trajectory $\boldsymbol{\beta}_i(\cdot)$ evaluated at time point t.

Hyperparameter Specifications

Our model includes several hyperparameters that require prior specifications. For the concentration parameter α, we specify a Gamma prior distribution $\alpha \sim \text{Gamma}(a_0, b_0)$, where a_0 and b_0 are the shape and rate parameters, respectively. The Gamma prior is conjugate to the Beta distribution in the stick-breaking process, facilitating posterior inference.

Let $\boldsymbol{\theta} = (\kappa, \sigma^2)$ denote a generic kernel parameter in a Matérn kernel, where $\kappa = \sqrt{5}/l$ is a scaled inverse length scale. We use log-normal priors: $\kappa \sim \text{LogNorm}(\mu_\kappa, \omega_\kappa)$ and $\sigma^2 \sim \text{LogNorm}(\mu_\sigma, \omega_\sigma)$.[3] The log-normal ensures that both the scaled inverse length scale and signal variance parameters are positive. Moreover, it provides sufficient flexibility to capture parameter uncertainty while maintaining computational tractability. We now describe the full model.

Full Consumer Choice Model (DDPH)

Consumer utility:

$$u_{ijtc} = \boldsymbol{x}'_{ijtc}\boldsymbol{\beta}_{it} + \varepsilon_{ijtc}, \quad \varepsilon_{ijtc} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \tag{9}$$

for each brand c in observation j in calendar time t for consumer i.

Customer trajectories and population dynamics:

$$\begin{aligned}
\boldsymbol{\beta}_i(\cdot) &= \tilde{\boldsymbol{\beta}}_{z_i}(\cdot), \quad z_i \sim \sum_{k=1}^{\infty} \pi_k \delta_k(\cdot), \quad \forall i \in \{1, \ldots, N\} \\
\pi_k &= V_k \prod_{h<k}(1 - V_h), \quad V_k \sim \text{Beta}(1, \alpha), \quad \forall k \in \{1, 2, \ldots\} \\
\tilde{\beta}_k^{(p)}(\cdot) &\sim \mathcal{H}^{(p)} = \mathcal{GP}\left(\mu^{(p)}(\cdot), K_{\boldsymbol{\theta}^{(p)}}(\cdot, \cdot)\right), \quad \forall k \in \{1, 2, \ldots\}, \forall p \in \{1, \ldots, P\} \\
\mu^{(p)}(\cdot) &\sim \mathcal{GP}\left(0, K_{\boldsymbol{\theta}_0}(\cdot, \cdot)\right), \quad \forall p \in \{1, \ldots, P\}.
\end{aligned} \tag{10}$$

[3] The mean and variance of $\log \kappa$ are μ_κ and ω_κ, respectively. Similarly, those of $\log \sigma^2$ are μ_σ and ω_σ.

Hyperparameters and hyperpriors:

$$\begin{aligned}
\boldsymbol{\theta}^{(p)} = \left(\kappa^{(p)}, \sigma^{2(p)}\right),\ \kappa^{(p)} \sim \text{LogNorm}\left(\mu_\kappa, \omega_\kappa\right),\ \sigma^{2(p)} \sim \text{LogNorm}\left(\mu_\sigma, \omega_\sigma\right), \quad \forall p \in \{1, \ldots, P\} \\
\boldsymbol{\theta}_0 = \left(\kappa_0, \sigma_0^2\right),\ \kappa_0 \sim \text{LogNorm}\left(\mu_\kappa, \omega_\kappa\right),\ \sigma_0^2 \sim \text{LogNorm}\left(\mu_\sigma, \omega_\sigma\right), \\
\alpha \sim \text{Gamma}(a_0, b_0),
\end{aligned} \tag{11}$$

where $a_0, b_0, \mu_\kappa, \mu_\sigma, \omega_\kappa, \omega_\sigma$ are hyperparameters of the model.[4] Having summarized the model, we now briefly discuss the benefits of the DDPH specification, especially in comparison with the GPH specification in Dew, Ansari, and Li (2020).

DDPH vs. GPH The DDPH nests a wide class of model specifications, highlighting its flexibility as a framework for capturing heterogeneity. First, the DDPH nests the static heterogeneity DP by imposing a degenerate dependence structure such that $\tilde{\boldsymbol{\beta}}_k(t) \equiv \tilde{\boldsymbol{\beta}}_k$ for all t with an appropriate base prior measure on $\tilde{\boldsymbol{\beta}}_k$. Second, the DDPH nests the GPH as a special case when α is sufficiently large. While the GP provides a highly flexible choice model as shown in Dew, Ansari, and Li (2020), the resulting cross-sectional population distribution is almost surely restricted to be a normal distribution. This limitation prevents GPs from representing the rich, multimodal or skewed distributional shapes often observed or expected in consumer populations. By contrast, the DDP offers a unified and highly flexible approach. It simultaneously allows the individual-level trajectories to be temporally correlated and the population distributions to be highly flexible in shape as well as correlated across time. Figure 1 contrasts the heterogeneity distributions induced by the GPH approach in Dew, Ansari, and Li (2020) (left) with those induced by our proposed approach (right). As can be seen from this figure, our DDPH approach allows the cross-sectional distributions to be shifted flexibly over time. It can therefore capture situations in which new modes or "segments" can appear over time or existing modes can disappear.

[4] We examine the sensitivity of our results to the choice of the hyperparameters in Web Appendix A.

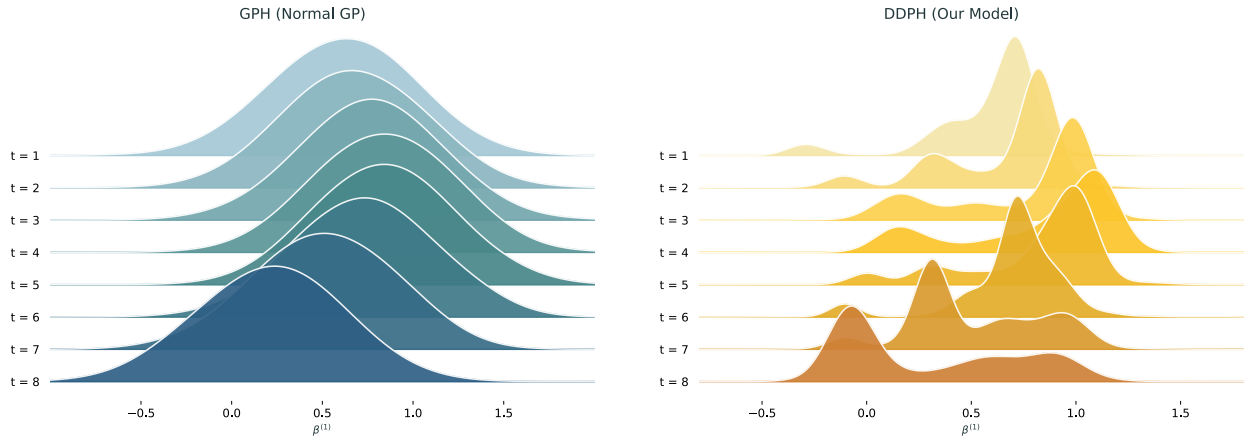

Figure 1: An illustrative example of estimated heterogeneity distributions across time by the GPH model (left) and the DDPH model (right).

Model Training

To train our nonparametric machine learning model, we employ a Bayesian approach that combines data augmentation of the utilities (McCulloch and Rossi 1994; Albert and Chib 1993; McCulloch, Polson, and Rossi 2000), Gibbs sampling as well as elliptical slice sampling (Murray, Adams, and MacKay 2010). The closed-form full conditional distributions are available for most of the model parameters and unknown quantities because of their conditional conjugacy. We therefore use Gibbs sampling steps to draw these parameters. However, the parameters of the kernel function $\boldsymbol{\theta} = (\kappa, \sigma^2)$ present a special challenge due to their positivity constraints and the non-conjugate nature of their full conditionals under the log-normal prior. While Metropolis-Hastings algorithms could be employed for these parameters, we instead use elliptical slice sampling. A key advantage of this choice is that it obviates the need to tune proposal distributions, which is a potentially cumbersome task given the use of multiple GPs in our model since each of which would require separate tuning. As we cannot work with an infinite dimensional object when estimating parameters, we truncate the stick-breaking representation and use a sufficiently large value $M = 300$ for the number of atoms, as suggested in Ishwaran and James (2001).

We run the MCMC chain for 20,000 iterations in our simulation studies and 100,000 to 300,000 iterations in our empirical application. We thin the latter half of each chain to retain a total of

1,000 draws in both the simulation and the application, which are used to construct the posterior distributions of the model parameters and other quantities of interest.

SIMULATION STUDIES

Overview

We conduct extensive Monte Carlo simulations to demonstrate that our Dependent Dirichlet Process Heterogeneity (DDPH) model robustly recovers preference parameters under various data-generating processes and that the less flexible, albeit state-of-the-art Gaussian Process Heterogeneity (GPH) model (Dew, Ansari, and Li 2020) can yield biased estimates of consumer response sensitivities due to its limited ability to capture complex cross-sectional preference distributions. The benchmark model captures continuous preference evolution by treating the customer-level trajectories as realizations from a Gaussian process. However, as mentioned earlier, it constrains the cross-sectional heterogeneity at any time point to be a normal distribution. This restrictive assumption prevents it from capturing complex cross-sectional distributions that evolve over time such as the one shown in Figure 1.

In our simulation studies, we use three data-generating processes that are designed to assess model performance when the preference trajectories and population distributions follow different patterns. To be consistent with our application, we ensure that all observations belonging to a customer for a given discrete time point t (e.g., month) share the same value for a given coefficient (e.g., $\beta_i^{(p)}(t)$ for the p-th coefficient of individual i). This results in a vector of parameter values $\beta_i^{(p)}(\boldsymbol{t})$ for the discrete time points $\boldsymbol{t} = (t_1, \ldots, t_T)'$ for each customer i. We set the total number of time periods to $T = 11$ in each simulated dataset. We use the first 8 time periods for model training and hold out $T_{\text{test}} = 3$ periods for out-of-sample prediction. We also set the number of consumers to $N = 500$, the maximum number of observations per consumer per time period to $J = 7$, the number of brands to $C = 3$, and the number of brand attributes to 2. For each data-generating process, we generate eight datasets with different random seeds and compare the models on these datasets. Additional details regarding the simulated datasets are presented in Web Appendix A. We

now describe the three data-generating processes we use in the simulation studies. We label them based on the nature of the cross-sectional population distribution in a given time period.

Dynamic Finite Mixture Distributions (D1) In this data-generating process, consumers are randomly assigned to one of five latent classes with equal probability. Each latent class is characterized by a dynamic preference trajectory generated from a Gaussian process:

$$\begin{aligned}
\boldsymbol{\beta}_i(\cdot) &= \tilde{\boldsymbol{\beta}}_{z_i}(\cdot), \quad z_i \sim \sum_{k=1}^{5} \pi_k \delta_k(\cdot), \quad \forall i \in \{1, \ldots, N\} \\
\pi_k &= \frac{1}{5}, \quad k \in \{1, 2, \ldots, 5\}, \\
\tilde{\boldsymbol{\beta}}_k^{(p)}(\cdot) &\sim \mathcal{GP}\left(\mu^{(p)}(\cdot), K_{\boldsymbol{\theta}^{(p)}}(\cdot, \cdot)\right), \quad \forall k \in \{1, 2, \ldots, 5\}, \forall p \in \{1, \ldots, P\} \\
\mu^{(p)}(\cdot) &\sim \mathcal{GP}\left(\mathbf{0}, K_{\boldsymbol{\theta}_0}(\cdot, \cdot)\right), \quad \forall p \in \{1, \ldots, P\}.
\end{aligned} \tag{12}$$

The details of the hyperparameter specifications are reported in Web Appendix A. As can be seen in Equation 12, all consumers within a latent class share the same time-evolving coefficients. At any given time point t, the cross-sectional preference heterogeneity distribution is a discrete distribution whose locations are the values of the latent class trajectories evaluated at t. Because the Gaussian process trajectories evolve over time, the cross-sectional heterogeneity distributions vary from one time period to another in their locations. This setup captures discrete class-based heterogeneity in the spirit of Kamakura and Russell (1989) but accounts for parameter dynamics.

Dynamic Normal Distributions: Gaussian Process (D2) In this data-generating process, each consumer's parameter trajectory is independently drawn from a Gaussian process with a common mean function and covariance kernel as in the Gaussian Process Heterogeneity model (GPH) proposed by Dew, Ansari, and Li (2020), i.e.,

$$\begin{aligned}
\boldsymbol{\beta}_i^{(p)}(\cdot) &\sim \mathcal{GP}\left(\mu^{(p)}(\cdot), K_{\boldsymbol{\theta}^{(p)}}(\cdot, \cdot)\right), \quad \forall i \in \{1, \ldots, N\}, \forall p \in \{1, \ldots, P\} \\
\mu^{(p)}(\cdot) &\sim \mathcal{GP}\left(\mathbf{0}, K_{\boldsymbol{\theta}_0}(\cdot, \cdot)\right), \quad \forall p \in \{1, \ldots, P\}.
\end{aligned} \tag{13}$$

Consequently, at any calendar time point t, the cross-sectional distribution of preferences is a normal distribution. The mean and variance of the cross-sectional population distributions as well as the evolution of these population distributions are determined by the mean function $\mu^{(p)}(\cdot)$ and the covariance kernel $K_{\boldsymbol{\theta}^{(p)}}(\cdot,\cdot)$ of the GP in Equation 13.

Dynamic Mixture of Normal Distributions (D3) In this data-generating process, consumers are assigned to one of five latent clusters with fixed probabilities $\boldsymbol{\pi}$. Each cluster is characterized by a Gaussian process, and individual preference trajectories within a cluster are drawn independently from this GP:

$$
\begin{aligned}
\beta_i^{(p)}(\cdot) &\sim \mathcal{GP}\left(\mu_{z_i}^{(p)}(\cdot), K_{\boldsymbol{\theta}_{z_i}^{(p)}}(\cdot,\cdot)\right), \quad \forall i \in \{1,\ldots,N\}, \forall p \in \{1,\ldots,P\} \\
z_i &\sim \sum_{k=1}^{5} \pi_k \delta_k(\cdot), \quad \forall i \in \{1,\ldots,N\} \\
\boldsymbol{\pi} &= (0.05, 0.15, 0.50, 0.20, 0.10) \\
\mu_k^{(p)}(\cdot) &\sim \mathcal{GP}\left(\mathbf{0}, K_{\boldsymbol{\theta}_0}(\cdot,\cdot)\right), \quad \forall k \in \{1,\ldots,5\}, \forall p \in \{1,\ldots,P\}.
\end{aligned}
\tag{14}
$$

At any given time point t, the cross-sectional population heterogeneity distribution is therefore a finite mixture of normal distributions. The mean and variance of normal distributions for each cluster k are derived from the mean function $\mu_k^{(p)}(\cdot)$ and the covariance kernel $K_{\boldsymbol{\theta}_k^{(p)}}(\cdot,\cdot)$ of the cluster-specific GP. Because the GP trajectories evolve over time, the location of each cluster's cross-sectional normal distribution changes dynamically, potentially resulting in multimodal heterogeneity distributions that shift across calendar periods. This setup can be viewed as a dynamic extension of the mixture-of-normals framework used by Allenby, Arora, and Ginter (1998).

We now compare the performance of our DDPH model to that of the benchmark GPH model in terms of the recovery of individual-level parameter trajectories and the ability to capture shifts in the population distributions over time. Although we fix the number of point masses or clusters in the simulation to generate the data, the DDPH model does not assume a fixed and known number of point masses. Instead, its nonparametric nature allows us to automatically determine

the appropriate complexity present in a given dataset, which is allowed to grow with data size.

Parameter Estimation

We begin by reporting the recovery of the parameters for the different data-generating processes.

Individual Preference Parameters We report results only for the first intercept and the first attribute coefficient to avoid clutter. The estimates of these parameters are representative of the overall results across all utility parameters as the attributes in the simulation are generic.

We measure how much the individual-level parameter estimates deviate from the true parameters $\beta_i^{(p)}(t)$ at each time t. For each of the $S = 8$ simulated datasets indexed by s, let $\hat{\beta}_i^{(p),(s,r)}(t)$ be the r-th draw from the posterior. We compute the posterior mean

$$\bar{\beta}_i^{(p),(s)}(t) = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_i^{(p),(s,r)}(t) \tag{15}$$

across the R retained MCMC iterations and then take its mean squared error relative to the truth,

$$\text{MSE}_i^{(p,t)} = \frac{1}{S}\sum_{s=1}^{S}\left(\bar{\beta}_i^{(p),(s)}(t) - \beta_i^{(p)}(t)\right)^2. \tag{16}$$

We then compute the population average $\text{MSE}^{(p,t)} = \frac{1}{N}\sum_{i=1}^{N}\text{MSE}_i^{(p,t)}$. As $\text{MSE}^{(p,t)}$ can be decomposed in terms of bias and variance, we report this breakdown in Web Appendix A.

Table 1 presents the posterior means of the MSEs for the first intercept $\beta_i^{\text{icept}_1}(\cdot)$ and the first attribute coefficient $\beta_i^{(1)}(\cdot)$ under data-generating processes D1 through D3 for each of the eight calendar time periods pertaining to the training data. Our DDPH model has significantly lower MSE values when compared to those for the GPH, for both data-generating processes. This highlights the advantage of our model in situations where the data-generating process can yield multimodal population distributions. Under D2, we see that the GPH has lower estimation errors because it is the correctly specified model for this data-generating process. However, the difference between the two models is not as substantial as the gap under D1 and D3.

Table 1: Mean square errors of parameter estimates under D1, D2, and D3.

	D1		**D2**		**D3**	
MSE	DDPH	GPH	DDPH	GPH	DDPH	GPH
$\beta^{\text{icept}_1}(1)$	**0.083**	0.359	0.338	**0.295**	**0.113**	0.455
$\beta^{\text{icept}_1}(2)$	**0.057**	0.209	0.321	**0.267**	**0.075**	0.289
$\beta^{\text{icept}_1}(3)$	**0.036**	0.103	0.310	**0.252**	**0.050**	0.125
$\beta^{\text{icept}_1}(4)$	**0.029**	0.039	0.301	**0.241**	**0.040**	0.042
$\beta^{\text{icept}_1}(5)$	**0.029**	0.032	0.307	**0.243**	**0.036**	0.041
$\beta^{\text{icept}_1}(6)$	**0.033**	0.062	0.327	**0.259**	**0.041**	0.076
$\beta^{\text{icept}_1}(7)$	**0.032**	0.069	0.347	**0.277**	**0.041**	0.065
$\beta^{\text{icept}_1}(8)$	**0.038**	0.061	0.374	**0.307**	**0.044**	0.057
$\beta^{(1)}(1)$	**0.007**	0.027	0.049	**0.037**	**0.016**	0.029
$\beta^{(1)}(2)$	**0.008**	0.030	0.040	**0.029**	**0.014**	0.033
$\beta^{(1)}(3)$	**0.006**	0.027	0.035	**0.025**	**0.012**	0.031
$\beta^{(1)}(4)$	**0.006**	0.022	0.034	**0.025**	**0.010**	0.026
$\beta^{(1)}(5)$	**0.008**	0.022	0.036	**0.027**	**0.011**	0.028
$\beta^{(1)}(6)$	**0.010**	0.024	0.039	**0.029**	**0.014**	0.029
$\beta^{(1)}(7)$	**0.013**	0.030	0.040	**0.029**	**0.017**	0.035
$\beta^{(1)}(8)$	**0.017**	0.055	0.048	**0.036**	**0.023**	0.051

Smaller MSE values for each parameter are in bold.

We also simulate from the posterior predictive distributions of the parameter trajectory for the first attribute ($\beta_i^{(1)}(\cdot)$) under each of D1 through D3. This distribution allows us to see how well each model captures the true cross-sectional population distribution of the coefficient at each time point, as it represents the distribution from which the parameter trajectory for an arbitrary consumer is drawn. Figure 2 shows how well the posterior predictive distributions for the two models track the true population distribution for D1. We focus on the first three time periods ($t = 1, 2, 3$). The true parameters (top row) come from a discrete distribution concentrated on five mass points at each time point. We visualize the posterior predictive distributions for the three time periods using discrete histograms. Our model (second row) successfully recovers the locations of the probability masses. However, the GPH benchmark (third row), given its assumption that the cross-sectional distributions are normal, produces a single broad mound of posterior mass centered around the

mean of the true distribution for each time point.

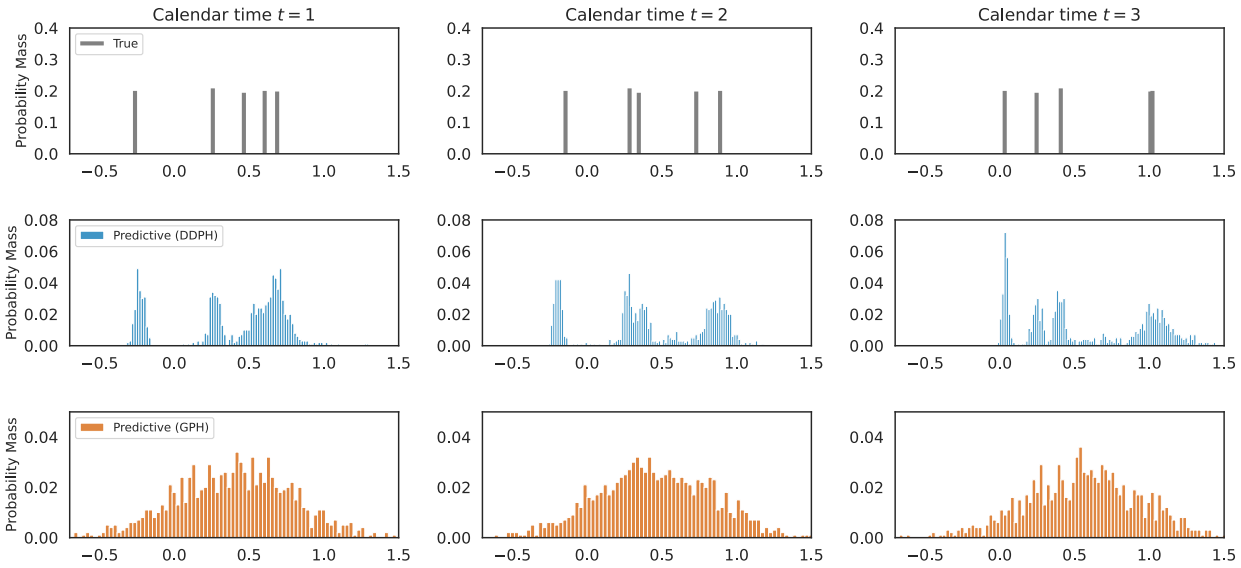

Figure 2: True parameter distribution (top), posterior predictive distribution of the parameter $\beta_i^{(1)}(t)$ from proposed model (middle), and benchmark model (bottom) under D1.

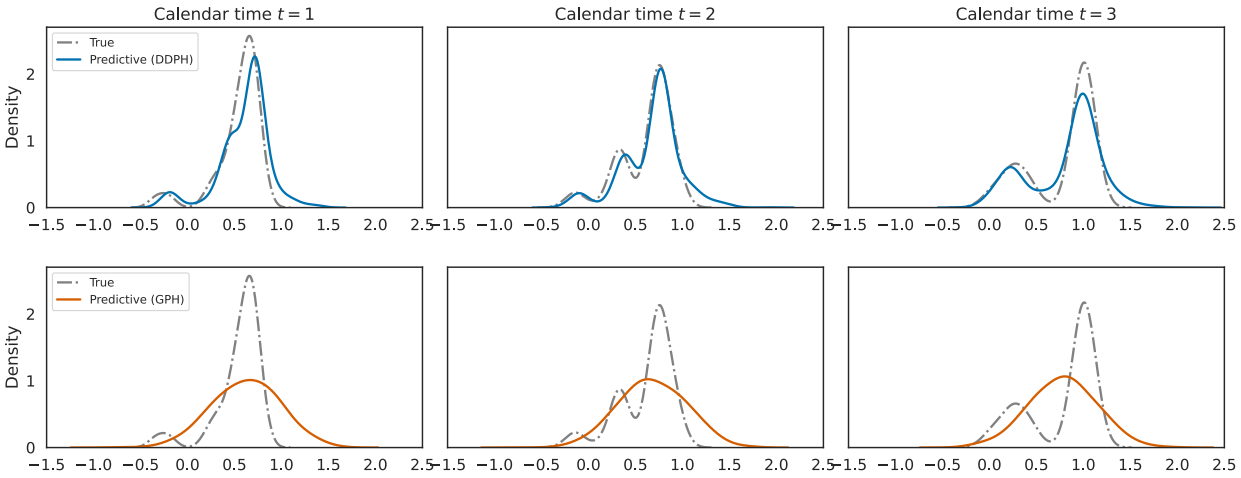

Figure 3: Posterior predictive distribution of the parameter $\beta_i^{(1)}(t)$ for proposed model (top) and benchmark model (bottom) under D3.

Figure 3 shows the cross-sectional population distributions and the posterior predictive distributions for D3, where the true parameters follow a mixture of normals at each time point. We depict the posterior predictive distributions for the two models using kernel density estimates. We see that the true distribution has two distinct modes at $t = 1$, with the right one more prominent. Our

model (top row) accurately reflects the bimodal structure and the relative heights of these peaks. The true preference distribution becomes more complex with three distinct modes as it evolves through $t = 2$, which our DDPH model tracks precisely. By $t = 3$, one of the modes disappears, and we again have a bimodal distribution. Our model successfully captures this distribution shift. It is particularly important to capture such a shift in markets where consumer preferences naturally cluster into distinct groups that evolve differently over time, such as markets with clear premium segments or those experiencing economic shocks such as technological disruption. In contrast, the benchmark GPH model (bottom row of Figure 3) again fails to capture the true preference heterogeneity in any period. It attempts to approximate the multimodal distribution with a normal distribution, resulting in a substantial mischaracterization of how consumer preferences vary in the population at different points in time.

Figure 4 shows the posterior predictive distributions and the true population distributions for three periods for D2. While both models faithfully mimic the normal distributions of preferences and their evolution, we can see that the benchmark model is better at recovering these distributions. These results demonstrate that our approach remains effective even when the underlying preference distributions lack complex patterns such as multimodality.

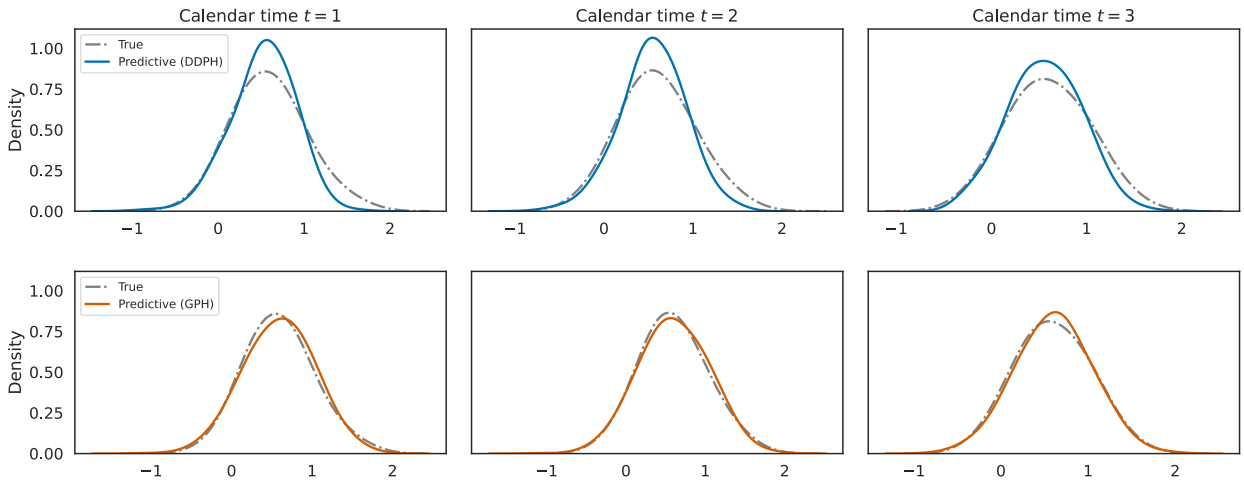

Figure 4: Posterior predictive distribution of the parameter $\beta_i^{(1)}(t)$ for proposed model (top) and benchmark model (bottom) under D2.

Overall, our model's performance is comparable to that of the benchmark when true preference

trajectories are from a Gaussian process, but it outperforms the benchmark when true preferences exhibit complex patterns such as multimodality. The robust performance of our model across different data-generating processes highlights its value for real-world applications, where the true nature of preference heterogeneity is typically unknown a priori. Its practical implications can be significant for marketing decisions. Misspecification of preference distributions can lead to suboptimal strategies, such as developing pricing that targets an "average" consumer, who may lie in between segments. Our model's ability to identify distinct consumer segments, represented by different modes in the preference distribution, can enable more effective targeting and personalization.

Heterogeneity Distribution Shift

The proposed model's ability to flexibly uncover and track shifts in cross-sectional preference distributions over time is a key advantage, as this is crucial for understanding behavioral changes under economic shocks, policy interventions, or other external influences. To see this, we focus on data-generating process D3 and use the Wasserstein distance to quantify intertemporal changes in heterogeneity distribution. We report detailed results for all DGPs in Web Appendix A. The Wasserstein distance between two uni-dimensional distributions P and Q is given by

$$D_W(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| dx, \tag{17}$$

where F_P and F_Q are the cumulative distribution functions of P and Q, respectively.

Figure 5 shows how well the two models track the true distance between temporally adjacent population distributions under the data-generating process D3. We can see from the figure that our model consistently follows the true distribution shifts more accurately across all time transitions, when compared to the benchmark model.

Predictive Performance

In this subsection, we discuss the performance of the two models in predicting the consumer choices. Table 2 presents the prediction accuracy results (hit-rates) under the different data-

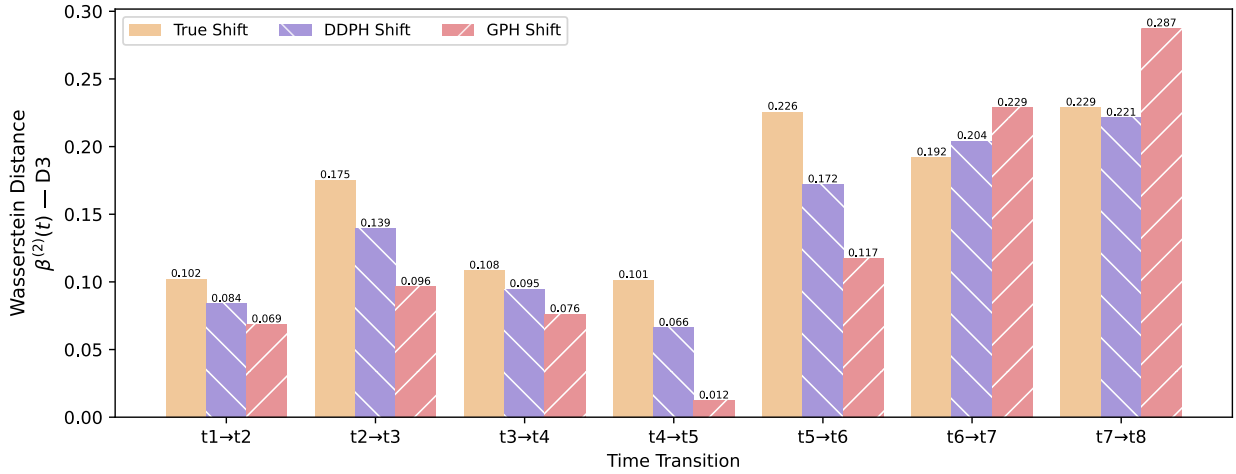

Figure 5: Wasserstein distance between time-adjacent heterogeneity distributions of $\beta^{(2)}(t)$, under D3.

generating processes. We make predictions on the training data in the column titled "Train (all)", which contains the first eight periods, i.e., $t \in \{1, \ldots, 8\}$, and on holdout data in the column titled "Test (all)", which contains the data for the last three periods, i.e., $t \in \{9, 10, 11\}$. We also separately report the accuracy for each of the three holdout periods. To generate these holdout choice predictions, we extrapolate the individual-level trajectories $\hat{\beta}_i^{(p),(r)}(\cdot)$ into the holdout periods for each of the MCMC draws using the covariance kernel and the extrapolated mean function of the underlying GP and then predict the choices by selecting the alternative with the highest choice probability for each observation. Detailed derivation is described in Appendix A. We then compute the prediction accuracy by calculating the proportion of the observations for which the model correctly predicts the chosen alternative. We finally compute the posterior mean of these accuracy measures across the MCMC iterations for each time period t.

For D1 and D3, where multimodal preference distributions evolve dynamically, our model substantially outperforms the benchmark GPH (e.g., 65.5% vs. 53.8% across all three test periods for D1), with the performance gap between the models narrowing as we move further into the future. In D2, where unimodal (normal) preference distributions evolve dynamically, the benchmark GPH model is correctly specified. Thus, the difference in accuracy between the two models is not significant in each holdout period $t \in \{9, 10, 11\}$. This indicates that using our more

Table 2: Prediction accuracy under D1, D2, and D3.

DGP	Model	Train (all)	Test (all)	Test ($t = 9$)	Test ($t = 10$)	Test ($t = 11$)
D1	DDPH	**0.929** (.01)	**0.655** (.03)	**0.823** (.01)	**0.658** (.03)	**0.488** (.04)
	GPH	0.829 (.00)	0.538 (.02)	0.619 (.01)	0.532 (.02)	0.465 (.03)
D2	DDPH	0.818 (.00)	0.621 (.02)	0.694 (.01)	0.622 (.02)	0.548 (.03)
	GPH	**0.823** (.00)	**0.624** (.02)	**0.696** (.01)	**0.626** (.02)	**0.551** (.03)
D3	DDPH	**0.909** (.01)	**0.640** (.02)	**0.794** (.02)	**0.640** (.02)	**0.485** (.02)
	GPH	0.839 (.00)	0.528 (.01)	0.625 (.01)	0.528 (.02)	0.430 (.02)

Standard deviations across datasets are presented in parentheses.

Higher accuracy values are in bold.

flexible approach entails minimal risk even when simpler models might suffice, while offering significant advantage when preferences exhibit complex distributions that may evolve over time. For completeness, macro-averaged recall and specificity are reported in Tables 15 and 16 in Web Appendix A. Across DGPs and forecast horizons, both metrics yield the same qualitative conclusion as the accuracy metric: DDPH outperforms GPH in D1 and D3 and their performance is comparable in D2.

Elasticity of Demand Estimation

In this subsection, we investigate how the consumer response sensitivities can deviate from the truth using different models. The own-attribute elasticity of demand measures the sensitivity of choice probabilities to changes in attributes such as price or promotion. For each MCMC draw r, we compute the elasticity as

$$e_{ijtc}^{(r)} = \frac{\partial \mathbb{P}\left(y_{ijt} = c | \{\boldsymbol{x}_{ijtc'}\}_{c'=1}^{C}, \hat{\boldsymbol{\beta}}_{it}^{(r)}\right)}{\partial x_{ijtc}^{(p)}} \cdot \frac{x_{ijtc}^{(p)}}{\mathbb{P}\left(y_{ijt} = c | \{\boldsymbol{x}_{ijtc'}\}_{c'=1}^{C}, \hat{\boldsymbol{\beta}}_{it}^{(r)}\right)}, \tag{18}$$

where $x_{ijtc}^{(p)}$ is the p-th product attribute variable, $\mathbb{P}\left(y_{ijt} = c | \{\boldsymbol{x}_{ijtc'}\}_{c'=1}^{C}, \hat{\boldsymbol{\beta}}_{it}^{(r)}\right)$ is the probability that consumer i chooses brand c in the j-th observation at time t, conditional on the product attributes $\{\boldsymbol{x}_{ijtc'}\}_{c'=1}^{C}$ and model parameter estimates $\hat{\boldsymbol{\beta}}_{it}^{(r)}$. The choice probabilities and their derivatives in

Equation 18 are obtained using the GHK simulator (Train 2009). To construct elasticity trajectories, we average $e_{ijtc}^{(r)}$ across all observations at each time period t for each brand c. Under D1, Figure 6 plots the resulting posterior mean trajectories and 95% credible intervals (shaded regions) for the proposed model (blue, dash-dotted line), the benchmark GPH model (orange line with circular markers), and the true elasticity trajectory (gray solid line). The corresponding elasticity estimates under D2 and D3, reported in Figures 16 and 17 in Web Appendix A, show that the two models track the truth comparably under D2, whereas under D3 the GPH benchmark systematically biases the elasticity estimates while the DDPH closely follows the true trajectory.

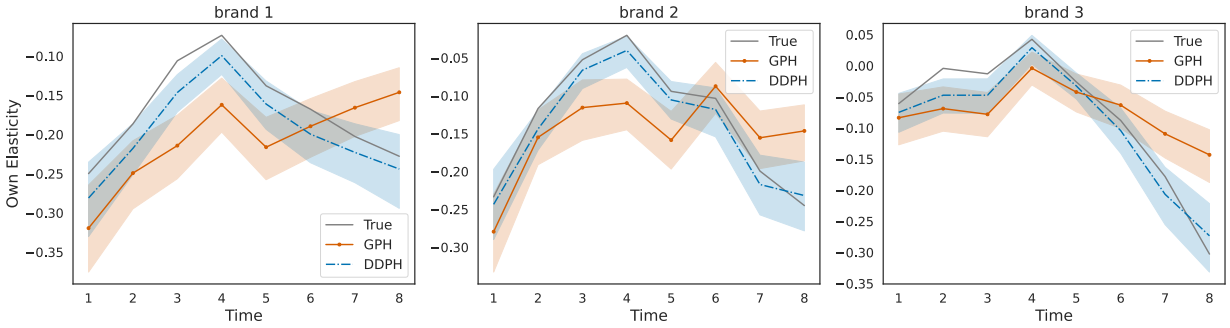

Figure 6: Own-attribute elasticity of demand under D1.

We see from Figure 6 that the DDPH model consistently tracks the true elasticity patterns more accurately than the benchmark GPH model across the brands and time periods under D1. Notably, the DDPH estimates closely follow the evolution of the true elasticity, while the GPH model underestimates the elasticity between $t = 1$ and $t = 6$ and overestimates it between $t = 7$ and $t = 8$ for all three brands shown in the figure. The credible intervals represented by the shaded regions in the figure are important for uncertainty quantification in the elasticity estimates. The intervals from the DDPH model generally encompass the true values, suggesting proper uncertainty quantification, while those from the GPH model exclude the true values in many periods across the brands. The differences in the elasticity estimates between the two models can have substantial implications for managerial decisions. For example, the benchmark model's inaccurate estimation of the attribute elasticity can lead to mis-priced products when focusing on the price attribute, as such pricing decisions would fail to properly capitalize on consumers' evolving price elasticities.

Counterfactual Profit

Having seen significant differences in the parameter estimates and own attribute elasticities between our model and the benchmark across the three data-generating processes, we now investigate how these differences translate into targeting decisions by evaluating the effect of personalized coupons on per-consumer profits.

To ensure that profits are well-defined and economically meaningful, we make the following three modifications to the data-generating processes. First, we draw the price feature from a log-normal distribution so that prices are always positive. Second, we constrain the price coefficient to be negative on average and ensure that demand is downward-sloping. Third, we restrict the simulation to a single feature, namely price, so that the only managerial lever is the discount applied to that price. All other simulation settings remain unchanged from the previous DGPs.

We assume a fixed profit margin of $m = 0.4$ (i.e., 40% of the sales) and consider seven discrete discount levels d for each product manufacturer. These discount levels range from $d = 0.0$ (no discount) up to $d = 0.3$ (30% off) in increments of 0.05. For each consumer i in the j-th observation in period t in the simulated dataset s, the expected profit from brand c is predicted as

$$\text{profit}_{ijtc}^{(s)}\left(d, \mathcal{M}, \{\boldsymbol{x}_{ijtc'}^{(s)}\}_{c'=1}^{C}\right) = (1-d)\cdot m \cdot x_{ijtc}^{\text{price},(s)} \cdot \mathbb{P}\left(y_{ijt}^{(s)} = c | d, \mathcal{M}, \{\boldsymbol{x}_{ijtc'}^{(s)}\}_{c'=1}^{C}\right), \tag{19}$$

where $\mathbb{P}\left(y_{ijt}^{(s)} = c | d, \mathcal{M}, \{\boldsymbol{x}_{ijtc'}^{(s)}\}_{c'=1}^{C}\right)$ is the probability that the consumer chooses brand c when the price of this brand is discounted by d and the rest of the brands are not discounted, based on the model $\mathcal{M}$. We obtain the optimal discount level $d_{i,c}^{*(s)}$ for brand c for each consumer i by maximizing the predicted total profit summed over their holdout period observations in the s-th dataset, $d_{i,c}^{*(s)} = \text{argmax}_d \sum_{j,t} \text{profit}_{ijtc}^{(s)}\left(d, \mathcal{M}, \{\boldsymbol{x}_{ijtc'}^{(s)}\}_{c'=1}^{C}\right)$, under each of the two models. We then compute the realized profit by using $d_{i,c}^{*(s)}$ in conjunction with the true parameter estimates that generated the data (i.e., true DGP). We average the realized profit across consumers within each simulated dataset, and report the mean of these per-dataset averages across the eight simulated datasets in Table 3, with the standard deviation across the datasets shown in parentheses.

Table 3 reports the per-consumer expected profit under D1, D2, and D3. Under D1 and D3, the DGPs with multimodal preference heterogeneity, DDPH yields a higher per-consumer profit than GPH for every product. Moreover, the gap between the two models is sizeable relative to the cross-dataset standard deviations. Under D2, where the GPH benchmark is correctly specified and the cross-sectional preference distribution is normal, both models yield comparable per-consumer profits on average (15.27 vs. 15.32), confirming that our more flexible model carries no meaningful profit penalty when a simpler specification suffices.

Table 3: Mean per-consumer profit under each optimal personalized couponing.

	D1		**D2**		**D3**	
Product	DDPH	GPH	DDPH	GPH	DDPH	GPH
1	**9.48**	8.90	**11.06**	10.98	**9.18**	8.54
	(0.59)	(0.33)	(0.30)	(0.29)	(0.43)	(0.22)
2	**12.00**	11.74	11.02	**11.22**	**12.07**	11.77
	(0.30)	(0.40)	(0.56)	(0.55)	(0.37)	(0.34)
3	**22.48**	21.86	23.75	**23.77**	**21.91**	21.49
	(0.52)	(0.48)	(0.40)	(0.36)	(0.57)	(0.56)
Average	**14.65**	14.17	15.27	**15.32**	**14.39**	13.93
	(0.33)	(0.24)	(0.33)	(0.32)	(0.30)	(0.16)

Combined with the earlier elasticity results, these findings show that DDPH's parameter-recovery advantage translates into meaningful economic gains when consumer preferences exhibit complex patterns such as multimodality. Furthermore, our model does not exhibit a significant impairment when the GPH model is the true model generating the data. In practice, little theory supports assuming normality (or other parametric distributions) for cross-sectional consumer preferences, especially when those distributions may vary over time. In our subsequent empirical application, we find rich and multimodal heterogeneity across different product categories, suggesting that non-normal cross-sectional heterogeneity appears to be prevalent. Combined with the simulation results, this implies that the Bayes decision rule favors our model over the benchmark.

EMPIRICAL APPLICATION

Data

We apply our modeling framework to an empirical application involving U.S. consumer packaged goods (CPG) using IRI household-level scanner panel data that span from January 2006 to December 2011. This period is particularly suitable for studying dynamic consumer heterogeneity, as it includes the 2008 financial crisis during which consumer preferences likely shifted in response to evolving economic uncertainty. Our analysis focuses on six commonly purchased categories: detergent, chips, toilet paper, peanut butter, coffee, and soup. In applying our model to the data, as Kim, Bradlow, and Iyengar (2022, 2026) show, the choice of the data or parameter granularity could be consequential. A trade-off exists between the precision with which a parameter can be estimated due to data sparsity and its bias that results from aggregating over a longer calendar time period. We discretize the parameter trajectories at the monthly level to manage this trade-off. The dataset contains information on two product attributes, price and the feature-display. The latter variable represents whether a product is featured or displayed in the store on a given observation. We hold out the last four months of the data for predictive tasks, as detailed in a later subsection. Table 4 provides descriptive statistics for each category. The columns in the table show the numbers of brands, the number of consumers, the mean and standard deviation (Std.) of the price, the percentage of the observations in which at least one brand was featured or displayed (% Feat.-Disp.), the average number of months in which a consumer made a purchase (Avg. Months/Cons.) and the average number of purchases per consumer (Avg. Purchases/Cons.).

While informative, these averages may obscure nuanced behavioral patterns that unfold over time or vary across individuals. To illustrate these patterns, we provide additional model-free evidence that highlights the presence of both temporal and cross-sectional heterogeneity in consumer behavior. Figure 7 displays the Herfindahl-Hirschman Index (HHI) to represent the brand-level market concentration levels in each quarter. The HHI is defined as the sum of squared market shares, $\sum_{c=1}^{C} s_{ct}^2$, where s_{ct} is the market share for brand c in our dataset in quarter time t. Higher values

Table 4: Summary statistics of the scanner data.

Category	Brands	Consumers	Price Mean (Std.)	% Feat.-Disp.	Avg. Months/Cons.	Avg. Purchases/Cons.
Detergent	6	1,117	1.21 (0.77)	88.31	20	25
Chips	4	1,552	4.16 (0.80)	95.73	28	46
Toilet paper	6	1,512	0.61 (0.17)	82.40	24	34
Peanut butter	5	1,085	1.96 (0.47)	85.31	19	25
Coffee	5	912	5.48 (2.24)	90.62	22	32
Soup	4	1,765	1.56 (0.40)	85.57	28	52

of the HHI indicate greater market concentration, whereas lower values reflect more competitive markets. In Figure 7, the period from 2007 Q4 to 2009 Q2 is highlighted with a square box, which corresponds to the duration of the Great Recession defined by the National Bureau of Economic Research (NBER 2023). We see from the figure that the concentration levels fluctuate for many brands across quarters. For instance, categories such as detergent and chips show visibly different concentration patterns before and during or after the Great Recession. This may suggest that consumers' brand choices and competitive dynamics were disrupted by macroeconomic shocks. Even before the Great Recession, categories like soup display shifting dynamics of brand concentration. These temporal shifts support the need for models that can flexibly capture evolving preferences.

Our data also exhibits significant variation in the number of purchases that consumers make across the time periods. Some consumers are active in nearly all months, while others make sporadic purchases. Given these non-uniform purchase trajectories, a Bayesian framework that adequately pools information across individuals and over time is beneficial for inference.

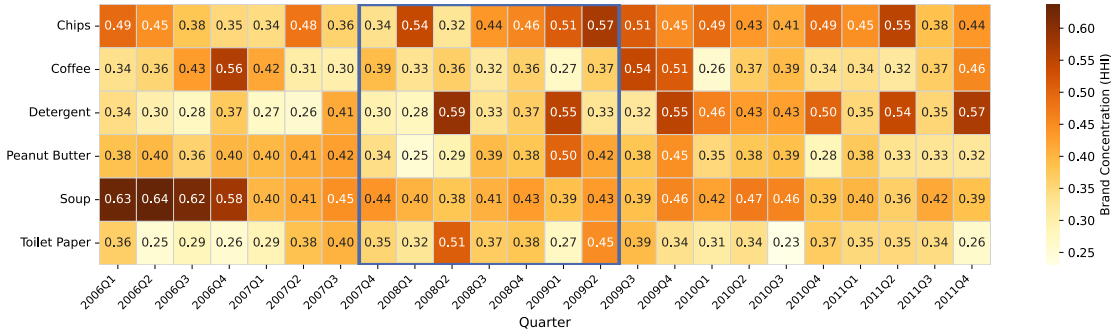

Figure 7: Brand concentration index (HHI) by category and quarter.

Parameter Estimates

We focus on one arbitrary category (chips) in the main manuscript to conserve space and avoid clutter and report the results for the other categories in Web Appendix B. We summarize the posterior distribution of the key parameters by reporting posterior means and standard deviations for the individual-level parameters in Table 5. The parameter estimates are based on the 1,000 thinned post burn-in MCMC draws. Specifically, we compute the mean and variance of the intercepts ($\mathbb{E}_{i,t}[\beta_i^{\text{icept}_c}(t)]$ and $\mathbb{V}_{i,t}[\beta_i^{\text{icept}_c}(t)]$) and the attribute coefficients ($\mathbb{E}_{i,t}[\beta_i^{\text{price}}(t)]$, $\mathbb{E}_{i,t}[\beta_i^{\text{ftdsp}}(t)]$, $\mathbb{V}_{i,t}[\beta_i^{\text{price}}(t)]$, and $\mathbb{V}_{i,t}[\beta_i^{\text{ftdsp}}(t)]$) across time periods and individuals. This yields a summary of population-level tendencies in the intrinsic preference for the brands, price sensitivities, and promotion sensitivities. To characterize the temporal dynamics of the parameter trajectories, we report the scaled inverse length scale (κ) and the signal variance (σ^2) in the Matérn kernel for the first intercept and the price coefficient.

From Table 5, we see noticeable differences in the estimated parameters between the two models. First, the DDPH estimates the average price coefficient across all consumers and time as -0.69, compared to the GPH's estimate of -0.96. Secondly, the variances of the individual-level parameters $\mathbb{V}_{i,t}[\beta_i^{(p)}(t)]$, measured across all individuals and time periods, are consistently larger under the GPH model than under the DDPH model. This difference may arise because the DDPH model, when computing the population variance, pools individuals into clusters with the same preference trajectories, which can mitigate the impact of individuals whose trajectories deviate substantially from the mean. The GPH model, because of its implied normal cross-sectional population distribution, tends to estimate a large variance to cover these far-out individuals. Thirdly, we compare the scaled inverse length scale parameters κ and signal variance parameters σ^2 in the Matérn kernel of the GP component from both models. We highlight the first intercept and the price coefficient as they are representative examples of how both models capture dynamic heterogeneity. Both models estimate small values for κ, indicating that the correlation between parameter values at different time points decays slowly, so the trajectories evolve smoothly. The relatively small signal variance (σ^2) values suggest that the trajectories of both the intercept and

Table 5: Posterior means and standard deviations of the parameters for the chips category.

Parameter Name	Parameter	GPH	DDPH
Individual-level Parameters			
Mean of Intercept 1	$\mathbb{E}_{i,t}[\beta_i^{\text{icept}_1}(t)]$	−0.84 (0.02)	−0.57 (0.02)
Mean of Intercept 2	$\mathbb{E}_{i,t}[\beta_i^{\text{icept}_2}(t)]$	−1.74 (0.04)	−1.11 (0.02)
Mean of Intercept 3	$\mathbb{E}_{i,t}[\beta_i^{\text{icept}_3}(t)]$	−0.47 (0.02)	−0.36 (0.02)
Mean of Price Coef.	$\mathbb{E}_{i,t}[\beta_i^{\text{price}}(t)]$	−0.96 (0.02)	−0.69 (0.01)
Mean of Feature Coef.	$\mathbb{E}_{i,t}[\beta_i^{\text{ftdsp}}(t)]$	0.47 (0.04)	0.36 (0.03)
Variance of Intercept 1	$\mathbb{V}_{i,t}[\beta_i^{\text{icept}_1}(t)]$	0.84 (0.04)	0.31 (0.02)
Variance of Intercept 2	$\mathbb{V}_{i,t}[\beta_i^{\text{icept}_2}(t)]$	1.84 (0.10)	0.56 (0.03)
Variance of Intercept 3	$\mathbb{V}_{i,t}[\beta_i^{\text{icept}_3}(t)]$	0.79 (0.04)	0.31 (0.02)
Variance of Price Coef.	$\mathbb{V}_{i,t}[\beta_i^{\text{price}}(t)]$	0.43 (0.03)	0.08 (0.01)
Variance of Feature Coef.	$\mathbb{V}_{i,t}[\beta_i^{\text{ftdsp}}(t)]$	0.29 (0.03)	0.13 (0.02)
GP Component Parameters			
Inverse Length Scale of Intercept 1	$\kappa^{(\text{icept}_1)}$	0.06 (0.00)	0.05 (0.01)
Inverse Length Scale of Price Coef.	$\kappa^{(\text{price})}$	0.08 (0.01)	0.03 (0.00)
Signal Variance of Intercept 1	$\sigma^{2(\text{icept}_1)}$	0.66 (0.04)	0.33 (0.05)
Signal Variance of Price Coef.	$\sigma^{2(\text{price})}$	0.37 (0.03)	0.03 (0.01)
DP Parameters			
Concentration	α	—	190.52 (11.03)
# of Point Masses	$\lvert\{k : \exists i, z_i = k\}\rvert$	—	119.19 (7.06)

Posterior standard deviations are presented in parentheses.

Individual-level parameters are averaged across time and consumers.

the price coefficient change relatively smoothly. Finally, we report the posterior mean of the DP concentration parameter α and the number of point masses. For the chips category, this number is estimated to be around 119, meaning that each of the 1,552 customers in the sample belongs to one of the 119 unique preference trajectories. We can infer that the DDPH induces clustering among customers, although the mass points cannot be directly interpreted as segments in the traditional marketing sense.[5]

Figure 8 visualizes the estimated evolution of the individual-level parameters in the chips category. To obtain this, we average each of the coefficients across the individuals to compute the

[5] We choose a sufficiently large truncation point at $M = 300$, as we describe in the Model section.

trajectory $\mathbb{E}_i[\hat{\beta}_i^{(p),(r)}(\boldsymbol{t})]$ for each MCMC iteration r, and then we plot the posterior mean of the trajectory with a 95% posterior credible interval for each of the periods. We see from the figure that both models capture similar overall patterns in the preference evolution. For example, they detect a sharp decline in the values of the brand intercepts (brands 1 and 2) during the Great Recession period (in the shaded region in the figure). This pattern suggests an increase in consumers' innate preference for the baseline brand (brand 0) during the period of economic turbulence, which was already the most preferred brand by an average consumer immediately before the recession.

Figure 8 also shows meaningful differences in the mean trajectories between the two models. The GPH trajectories exhibit substantial month-to-month volatility, particularly for the intercepts and price coefficient, whereas the DDPH produces smoother trajectories that also capture systematic shifts during the recession. Across the full training window, the GPH price coefficient is consistently more negative than DDPH's, averaging −0.96 as opposed to −0.69, as shown in Table 5.

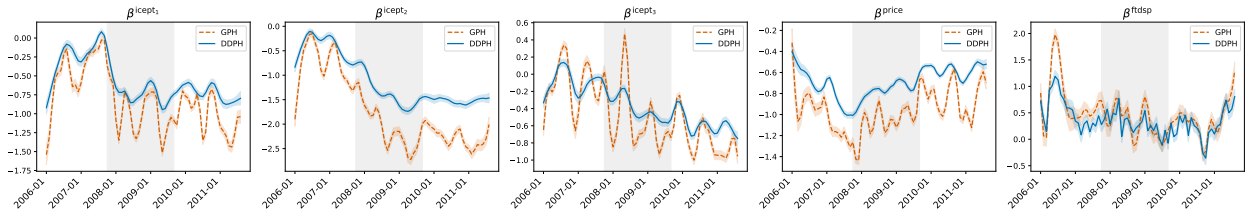

Figure 8: Population mean of the individual-level parameters over time for the chips category.

Figure 9 showcases the posterior predictive distributions of the price coefficient across several months spanning the Great Recession. It shows that the shape and location of these distributions differ significantly between the two models. The GPH model produces broad, unimodal distributions, a consequence of its reliance on the Gaussian assumption that cannot accommodate latent subgroups explicitly. In contrast, the DDPH model captures the bimodality with a sharp peak near −1 and a smaller mode near −0.5 across all the five months shown in the figure. This bimodal structure is preserved throughout the recession window, indicating that the heterogeneity captured by DDPH represents a stable feature of the population in the chips category. Web Appendix B shows the results for other categories and we find meaningful differences between the two models. For example, in Figure 26, the DDPH suggests that the two modes in the price coefficient distribu-

tion shift toward the more negative values during the recession period (from the second to the third panel in the figure) in the detergent category. The GPH, constrained by the cross-sectional normal assumption, suggests the shift in the opposite direction. The ability to capture this multimodal structure underscores the flexibility of the DDPH model in estimating time-varying cross-sectional heterogeneity. Thus, our model can provide more granular insights into how consumer preferences vary across subgroups, which is crucial for pricing and targeting decisions.

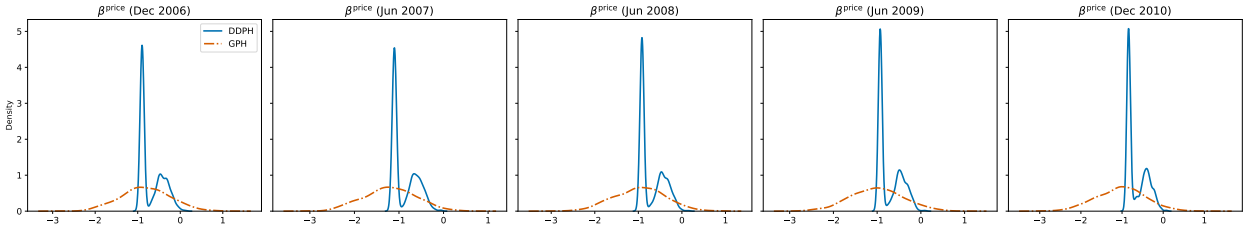

Figure 9: Posterior predictive distribution of the price coefficient for the chips category.

Model Comparison and Fit

We perform a series of posterior predictive checks to assess the adequacy of both models in recovering the market shares of the different brands, both across the data duration and in each time period. For conciseness and brevity, we report the associated results in Figures 31 to 42 in Web Appendix B. These figures indicate that both models are capable of adequately recovering the market shares. This is not surprising as both models involve sophisticated specifications.

To evaluate the predictive performance of the proposed DDPH model relative to the benchmark GPH, we compute the accuracy of choice prediction, defined as the proportion of correct predictions across all consumers, brands, and time periods for each MCMC draw, as is the case in the simulation studies. We report the posterior mean accuracy in Table 6 as well as recall (sensitivity) and specificity in Tables 17 and 18 in Web Appendix B. The results show that the relative performance of DDPH to GPH is consistent across categories and prediction tasks. We evaluate three distinct prediction tasks: (1) in-sample predictive performance, (2) prediction for purchases of new consumers within the training period, and (3) forecasting future purchases of existing consumers. For each of these prediction evaluations, we compare the two models separately across the six

product categories. Table 6 presents the accuracy statistics. We now describe the results.

Table 6: Posterior mean accuracy for three prediction tasks: (1) in-sample fit, (2) new consumer's purchase decision, (3) future purchases for existing consumers.

	In-sample fit		New consumer prediction		Future purchase prediction	
Category	DDPH	GPH	DDPH	GPH	DDPH	GPH
Detergent	0.738	**0.739**	**0.464**	0.464	**0.642**	0.547
Chips	**0.622**	0.592	**0.451**	0.445	**0.609**	0.521
Toilet paper	**0.696**	0.671	**0.393**	0.383	**0.437**	0.338
Peanut butter	**0.729**	0.723	**0.505**	0.483	**0.458**	0.391
Coffee	**0.680**	0.664	**0.325**	0.321	**0.506**	0.475
Soup	**0.565**	0.511	**0.416**	0.404	**0.466**	0.422
Average	**0.672**	0.650	**0.426**	0.417	**0.520**	0.449

Higher values in each row for each task are in bold.

In-Sample fit As we can see from Table 6, the DDPH model achieves higher in-sample accuracy than the GPH in most categories. For example, in the soup category, DDPH attains an average accuracy of 0.565 compared to 0.511 for GPH. Similar improvements are observed in other categories except detergent, where DDPH is almost on par with GPH. This suggests that our proposed model is well calibrated and highlights the importance of estimating dynamic heterogeneity flexibly.

Prediction for new consumers To evaluate predictive performance for new consumers (i.e., a cold-start setting), we first set aside 20% of consumers from the population randomly. These consumers are excluded when fitting the model. We draw the preference trajectories for these unseen consumers during the training period from their posterior predictive distributions. See Web Appendix B for more details on the data-splitting procedure. For each MCMC draw r, we make prediction by $\hat{y}_{ijt}^{(r)}$ for new consumers $i \in \{1, \ldots, N_{\text{test}}\}$ using observed brand attributes $\boldsymbol{x}_{ijtc}$ and consumer-specific preference trajectories $\hat{\boldsymbol{\beta}}_i^{(r)}(\cdot)$. The prediction accuracy is then computed as the proportion of correctly predicted choices for the N_{test} consumers, which is presented in Table 6. The DDPH model shows advantages in predicting new consumer behavior, with the

largest improvements in the peanut butter (accuracy increases by 4.6%) and toilet paper (accuracy increases by 2.6%) categories. While both models exhibit lower overall prediction accuracy than the in-sample fit, our model is consistently superior across all categories.

Prediction of future behavior of existing consumers We predict the choices for the existing consumers (i.e., those in the training data) in the future periods. Specifically, we draw the preference trajectories for the holdout time periods from their posterior predictive distribution. Using these draws, we generate predicted choices as $\hat{y}_{ijt}^{(r)}$ and compute the prediction accuracy for each MCMC iteration r. Table 6 reports the posterior means of the accuracy measures across the iterations. We see from the table that the DDPH model substantially outperforms the GPH model across all categories in the holdout time periods for the existing consumers. This demonstrates the strength of our proposed model in flexibly capturing the evolving patterns of individual-level preferences, which can enhance its ability to project temporal dynamics into the future.

Price Elasticities

To evaluate how each model captures consumer responsiveness to prices, we compute the own-price elasticities $e_{ijtc}^{(r)}$ using Equation 18. For each brand c within a category, we average these estimates across all observations, calendar months, and MCMC draws. Figure 10 plots these aggregated values, contrasting the DDPH model estimates (vertical axis) with the GPH model estimates (horizontal axis). The figure shows that the estimates from our model are more plausible and are consistent with economic expectations. In contrast, the GPH model yields implausibly large negative elasticities for several brands in the detergent and peanut butter categories. This can be a consequence of the broad span of the price coefficient distribution across the time periods in these two categories. We can visually confirm this from Figures 26 and 28 in Web Appendix B.

Next, we examine the evolution of estimated own-price elasticities over time. For both models, we compute the posterior distribution of the mean elasticity at each time period t by averaging $e_{ijtc}^{(r)}$ across all observations and brands within each time period and category. Figure 11 plots the posterior means of these trajectories alongside their 95% credible intervals (shaded regions),

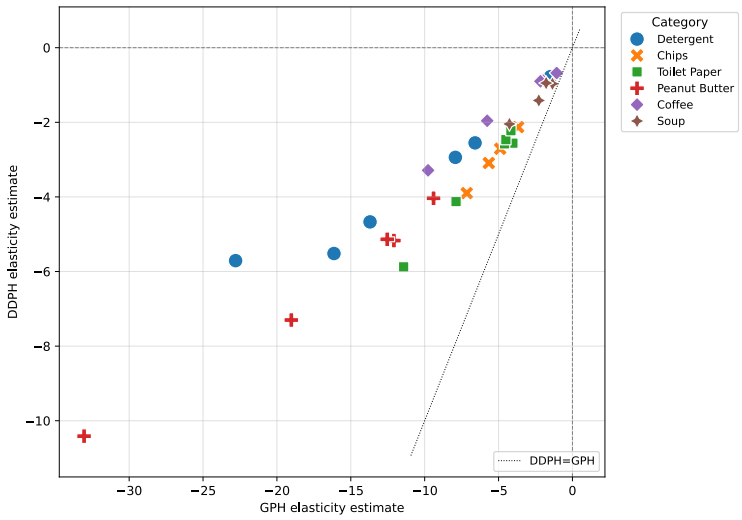

Figure 10: Mean price elasticity for each product and category.

providing a category-level overview of temporal changes in price sensitivity. From the figure, we observe that GPH consistently estimates substantially more negative elasticities than DDPH across all six categories, with the gap most pronounced for detergent, peanut butter, and toilet paper. The GPH trajectories also exhibit much greater volatility over time, while DDPH evolves more smoothly. The clearest temporal shift appears in the soup category, where both models suggest a sharp early-period drop followed by a recovery toward zero by mid-2007.

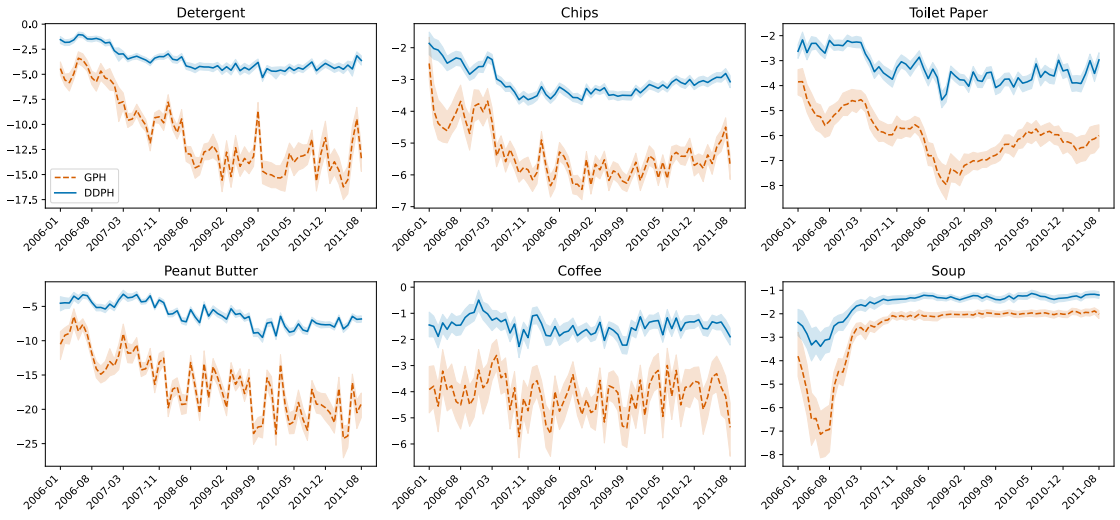

Figure 11: Mean price elasticity over time for each category.

Targeting Implications

To investigate how the model differences translate into managerial value on real scanner data, we now compare the optimal couponing strategies recommended by DDPH and GPH. We first establish our model's credibility by examining how well it predicts observed profits in the status quo (no coupon) in the holdout periods. As shown in Table 7, the profit estimates from the DDPH model are more accurate across all categories. Having established the superiority of our model in tracking profits, we use it to compute the counterfactual profits when evaluating the couponing strategies proposed by each model.[6]

Table 7: Mean absolute difference between predicted and observed profits in holdout periods.

Model	Detergent	Chips	Toilet Paper	Peanut Butter	Coffee	Soup
GPH	0.07	0.27	0.05	0.22	0.49	0.26
DDPH	0.05	0.21	0.04	0.16	0.29	0.09

We use the same counterfactual setup as in the simulation studies (Equation 19) to predict manufacturer profits under personalized couponing, with $m = 0.4$ and seven possible discount levels $d \in \{0, 0.05, 0.10, \ldots, 0.30\}$. The targeting horizon is the first month of the holdout sample: for each consumer, we sum the predicted profits across their observations in that month and then average over consumers. This corresponds to a one-month-ahead forecast of per-customer profits from the end of the training period. Following the credibility check above, we use the parameter estimates from the DDPH model when evaluating the optimal strategies proposed by the two models. We report the resulting per-customer profits along with DDPH's percentage gain over GPH in Table 8, which shows that DDPH outperforms GPH in every category. The percentage gains are largest in coffee (4.6%), peanut butter (4.1%), chips (4.1%), and detergent (4.0%), suggesting that personalized couponing in these categories is particularly impactful when consumer heterogeneity is captured flexibly. These differences in profit gains imply that flexibly capturing dynamic heterogeneity can translate into more effective targeting.

[6]Note that the nearly continuous support of the price distribution for each brand prevents reliable use of an inverse propensity weighted estimator to compute counterfactual profits as in Smith, Seiler, and Aggarwal (2023).

Table 8: Estimated profits under optimal coupon strategies.

Category	GPH	DDPH	DDPH gain over GPH (%)
Chips	0.645	0.671	4.06
Coffee	0.641	0.670	4.59
Detergent	0.082	0.085	4.02
Peanut Butter	0.214	0.222	4.09
Soup	0.282	0.285	0.81
Toilet Paper	0.048	0.049	3.32

We find that both models predominantly favor the no-coupon option, with no-discount recommendations accounting for roughly 45% to 99% of all customers across categories. Figure 12 displays the proportion of consumers assigned to each non-zero discount level within each product category. Focusing on the cases with positive optimal coupon amounts (i.e., $d > 0$), we observe sizable differences between the two models. For example, the GPH tends to give deeper discounts in the chips category (e.g., 20% or higher discounts to more than 10% of the customers), while the DDPH allocates the coupons in a more conservative way. Because of its flexibility, the targeting strategies based on DDPH can better adapt to the specific context of each product category. As shown in Figure 12, our model targets more than 35% of customers in the detergent category but only around 1% in the soup category. In contrast, the strategies derived from the GPH target at least 10% of customers with positive discounts within each category.

The results we have shown in this section highlight the managerial value of the proposed model in capturing dynamic heterogeneity flexibly: not only can it fit individual-level preferences more accurately both in-sample and out-of-sample, but it can also lead to more effective marketing strategies that drive meaningful economic outcomes.

CONCLUSION

We developed a novel Bayesian machine learning framework to model consumer preferences that evolve over time. Dynamic consumer preferences are common in contexts involving macro-level shifts in economic conditions and micro-level changes in individual-level factors. Our model

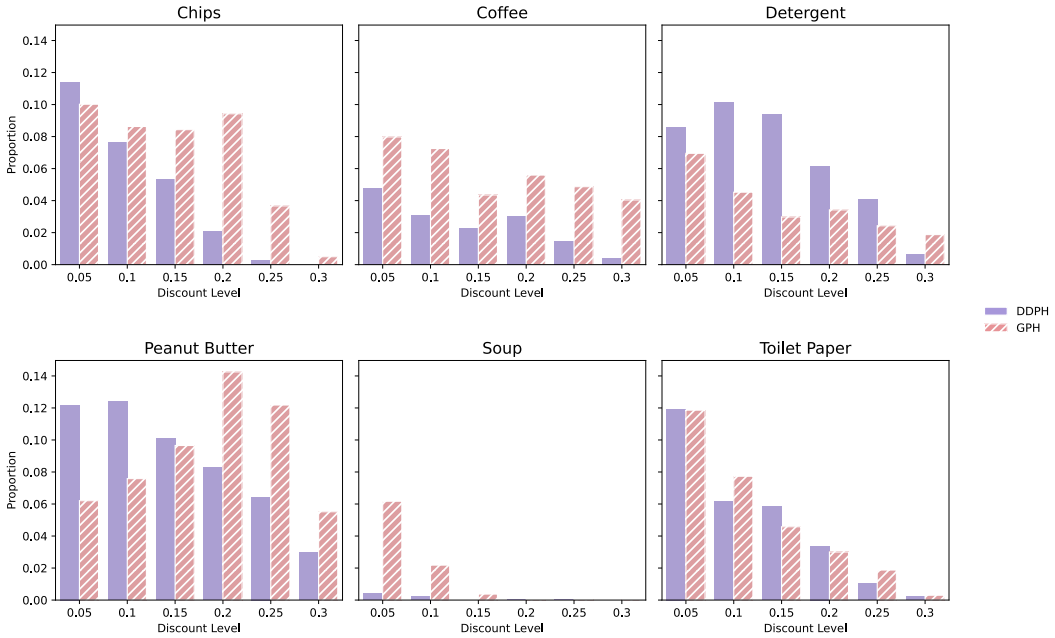

Figure 12: Optimal coupon proportions (only positive discount levels shown).

represents variations in consumer-level parameters such as brand-specific intercepts and price sensitivities using trajectories that span over time and captures the heterogeneity in these trajectories through the novel use of a dependent Dirichlet process in a discrete choice setting. Our nonparametric machine learning framework enables us to infer the appropriate level of model complexity in a data-driven manner. This is particularly useful when markets are characterized by complex cross-sectional heterogeneity involving multimodal or skewed population distributions that change their shape over time because of changes in the individual-level preference trajectories. In such environments, existing approaches to modeling heterogeneity are either inadequate in capturing the complex form of heterogeneity in the data or are too onerous in terms of specification searches. In contrast, our methodology simultaneously allows for flexible and temporally correlated population-level distributions and individual-level trajectories.

We developed a Bayesian estimation approach based on MCMC methods to effectively sample the model parameters and unknown quantities from the posterior distribution. The key feature of our estimation methodology is that it eliminates the need for tuning, which is often computationally burdensome and time-consuming when Metropolis-Hastings steps are involved in inference. We

achieved this through a novel combination of Gibbs sampling, data augmentation, and elliptical slice sampling steps in our MCMC algorithm.

We conducted extensive simulation studies that demonstrate the benefits of our modeling framework when compared to a state-of-the-art dynamic heterogeneity specification recently proposed in the literature. Our simulation results show that our model is able to estimate key model parameters accurately and capture complex heterogeneity patterns that evolve over time, while the GP based benchmark model fails to do so due to its limited flexibility in representing cross-sectional heterogeneity. We also showed that our model performed on par with the GPH when the data are generated from the benchmark, which highlights our model's robustness across settings. We also showed that the flexibility of our model translates into better performance in both in-sample and out-of-sample predictions, as well as in the characterization of attribute elasticities. Finally, we underscored the managerial value of the proposed model by considering personalized couponing strategies derived from the model. Targeting based on our model yielded substantially higher profits than targeting based on the benchmark GPH once the true data-generating process deviated from the GPH assumption. The two models performed comparably when the GPH was the correctly specified model. The results give further credence to our model since consumer dynamic heterogeneity can exhibit complex, potentially multimodal cross-sectional patterns in many real-world scenarios.

We applied our modeling framework and estimation methodology to IRI household scanner data for six common consumer packaged goods categories between 2006 and 2011, which covers a period of the Great Recession. Our model discovered nuanced temporal dynamics in consumer brand preferences and promotion responsiveness. Moreover, it showed superior performance in predicting consumer choices within the training sample and for new consumers or future periods across most categories. We also found that the two models estimated price elasticities differently. The GPH significantly overestimated their magnitude across time periods. Finally, we showed how our model could support managerial decision making through a counterfactual personalized coupon targeting simulation. The results showed that the couponing strategy based on our model

could provide significantly higher profits than that based on the benchmark. One reason for this improvement may be because our model can detect evolving consumer preferences and track the consumer segments within the population more effectively.

We showcased our methodology using a canonical discrete choice setting that is central to many marketing tasks. However, our proposed framework is fairly general and can be applied to other contexts where accounting for both cross-sectional and dynamic heterogeneity is important. For example, researchers can employ our methodology in analyses involving different types of panel data or in other contexts such as social networks (Braun and Bonfrer 2011; McShane, Bradlow, and Berger 2012) and dynamic topic models (Dew, Ansari, and Li 2020). While our framework is very flexible, it could be modified or extended in some directions. For simplicity, we assumed prices to be exogenous. This could be relaxed via the use of control function approaches when appropriate instruments are available. Future research could explore other nonparametric specifications which may be suitable in particular contexts. For example, we assumed that the stick-breaking weights are constant over time, but these can be allowed to vary over time for greater flexibility (Quintana et al. 2022). In our application, we used monthly calendar time periods. In general, this can be a consequential choice. In our situation, we found similar qualitative results for quarterly data. However, the method proposed by Kim, Bradlow, and Iyengar (2022, 2026) could be beneficial in inferring the right level of aggregation in a given situation. While we focused on dynamic heterogeneity, another interesting avenue for future research could extend our approach to incorporate spatial variation, where the pattern of dynamic evolution may vary across different geographic regions.

REFERENCES

Albert, James H and Siddhartha Chib (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88 (422), 669–679.

Allenby, Greg M, Neeraj Arora, and James L Ginter (1998), "On the heterogeneity of demand," *Journal of Marketing Research*, 35 (3), 384–389.

Ansari, Asim and Carl F Mela (2003), "E-customization," *Journal of Marketing Research*, 40 (2), 131–145.

Arora, Neeraj, Greg M Allenby, and James L Ginter (1998), "A hierarchical Bayes model of primary and secondary demand," *Marketing Science*, 17 (1), 29–44.

Ascolani, F, B Franzolini, A Lijoi, and I Prünster (2023), "Nonparametric priors with full-range borrowing of information," *Biometrika*, 111 (3), 945–969.

Ascolani, F, A Lijoi, G Rebaudo, and G Zanella (2022), "Clustering consistency with Dirichlet process mixtures," *Biometrika*, 110 (2), 551–558.

Boughanmi, Khaled and Asim Ansari (2021), "Dynamics of musical success: A machine learning approach for multimedia data fusion," *Journal of Marketing Research*, 58 (6), 1034–1057.

Braun, Michael and André Bonfrer (2011), "Scalable inference of customer similarities from interactions data using Dirichlet processes," *Marketing Science*, 30 (3), 513–531.

Braun, Michael, Peter S Fader, Eric T Bradlow, and Howard Kunreuther (2006), "Modeling the 'pseudodeductible' in insurance claims decisions," *Management Science*, 52 (8), 1258–1272.

Bruce, Norris I (2019), "Bayesian nonparametric dynamic methods: Applications to linear and nonlinear advertising models," *Journal of Marketing Research*, 56 (2), 211–229.

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

Dew, Ryan, Asim Ansari, and Yang Li (2020), "Modeling dynamic heterogeneity using Gaussian processes," *Journal of Marketing Research*, 57 (1), 55–77.

Dew, Ryan, Nicolas Padilla, Lan E. Luo, Shin Oblander, Asim Ansari, Khaled Boughanmi, Michael Braun, Fred Feinberg, Jia Liu, Thomas Otter, Longxiu Tian, Yixin Wang, and Mingzhang Yin (2024), "Probabilistic Machine Learning: New Frontiers for Modeling Consumers and their Choices," *International Journal of Research in Marketing* https://www.sciencedirect.com/science/article/pii/S0167811624000995.

Feldman, Joseph and Daniel R. Kowal (2024), "Nonparametric Copula Models for Multivariate, Mixed, and Missing Data," *Journal of Machine Learning Research*, 25 (164), 1–50.

Finocchio, Gianluca and Johannes Schmidt-Hieber (2023), "Posterior Contraction for Deep Gaussian Process Priors," *Journal of Machine Learning Research*, 24 (66), 1–49.

Gordon, Brett R, Avi Goldfarb, and Yang Li (2013), "Does price elasticity vary with economic growth? A cross-category analysis," *Journal of Marketing Research*, 50 (1), 4–23.

Guhl, Daniel, Bernhard Baumgartner, Thomas Kneib, and Winfried J Steiner (2018), "Estimating time-varying parameters in brand choice models: A semiparametric approach," *International Journal of Research in Marketing*, 35 (3), 394–414.

Ishwaran, Hemant and Lancelot F James (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96 (453), 161–173.

Jedidi, Kamel, Carl F Mela, and Sunil Gupta (1999), "Managing advertising and promotion for long-run profitability," *Marketing Science*, 18 (1), 1–22.

Kamakura, Wagner A and Gary J Russell (1989), "A probabilistic choice model for market segmentation and elasticity structure," *Journal of Marketing Research*, 26 (4), 379–390.

Kim, Jin Gyo, Ulrich Menzefricke, and Fred M Feinberg (2004), "Assessing heterogeneity in discrete choice models using a Dirichlet process prior," *Review of Marketing Science*, 2 (1), 00001022021546561610003.

Kim, Mingyung, Eric T. Bradlow, and Raghuram Iyengar (2022), "Selecting Data Granularity and Model Specification Using the Scaled Power Likelihood with Multiple Weights," *Marketing Science*, 41 (4), 848–866.

Kim, Mingyung, Eric T. Bradlow, and Raghuram Iyengar (2026), "A Bayesian Dual Clustering Approach for Selecting Data and Parameter Granularities," *Marketing Science*, 45 (3), 556–575.

Lachaab, Mohamed, Asim Ansari, Kamel Jedidi, and Abdelwahed Trabelsi (2006), "Modeling preference evolution in discrete choice models: A Bayesian state-space approach," *Quantitative Marketing and Economics*, 4 (1), 57–81.

Li, Yang and Asim Ansari (2014), "A Bayesian semiparametric approach for endogeneity and heterogeneity in choice models," *Management Science*, 60 (5), 1161–1179.

MacEachern, Steven N "Dependent nonparametric processes," "ASA Proceedings of the Section on Bayesian Statistical Science," Vol. 1., pages 50–55, Alexandria, VA (1999).

MacEachern, Steven N (2000), "Dependent Dirichlet processes," *Unpublished manuscript, Department of Statistics, The Ohio State University*, 5.

McCulloch, Robert and Peter E Rossi (1994), "An exact likelihood analysis of the multinomial probit model," *Journal of Econometrics*, 64 (1-2), 207–240.

McCulloch, Robert E, Nicholas G Polson, and Peter E Rossi (2000), "A Bayesian analysis of the multinomial probit model with fully identified parameters," *Journal of Econometrics*, 99 (1), 173–193.

McShane, Blakeley B., Eric T. Bradlow, and Jonah Berger (2012), "Visual Influence and Social Groups," *Journal of Marketing Research*, 49 (6), 854–871.

Murray, Iain, Ryan Adams, and David MacKay "Elliptical slice sampling," "Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics," pages 541–548, JMLR Workshop and Conference Proceedings (2010).

Naik, Prasad A et al. (2015), "Marketing dynamics: A primer on estimation and control," *Foundations and Trends® in Marketing*, 9 (3), 175–266.

NBER "US Business Cycle Expansions and Contractions," (2023) https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions.

Netzer, Oded, James M Lattin, and Vikram Srinivasan (2008), "A hidden Markov model of customer relationship dynamics," *Marketing Science*, 27 (2), 185–204.

Onzo, Kohei and Asim Ansari (2025), "Bayesian nonparametric sequential search," *Journal of Marketing Research*, 62 (2), 362–385.

Porcu, Emilio, Moreno Bevilacqua, Robert Schaback, and Chris J Oates (2024), "The Matérn model: A journey through statistics, numerical analysis and machine learning," *Statistical Science*, 39 (3), 469–492.

Quintana, Fernando A, Peter Müller, Alejandro Jara, and Steven N MacEachern (2022), "The dependent Dirichlet process and related models," *Statistical Science*, 37 (1), 24–41.

Riva-Palacio, Alan, Fabrizio Leisen, and Jim Griffin (2022), "Survival Regression Models With Dependent Bayesian Nonparametric Priors," *Journal of the American Statistical Association*, 117 (539), 1530–1539.

Rosa, Paul and Judith Rousseau (2025), "Nonparametric Regression on Random Geometric Graphs Sampled from Submanifolds," *Journal of Machine Learning Research*, 26 (164), 1–65.

Rossi, Peter E, Robert E McCulloch, and Greg M Allenby (1996), "The value of purchase history data in target marketing," *Marketing Science*, 15 (4), 321–340.

Sethuraman, Jayaram (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, pages 639–650.

Smith, Adam N., Stephan Seiler, and Ishant Aggarwal (2023), "Optimal Price Targeting," *Marketing Science*, 42 (3), 476–499.

Sriram, Srinivasaraghavan, Pradeep K Chintagunta, and Ramya Neelamegham (2006), "Effects of brand preference, product attributes, and marketing mix variables in technology product markets," *Marketing Science*, 25 (5), 440–456.

Train, Kenneth E (2009), *Discrete choice methods with simulation* Cambridge university press.

Um, Seungha, Tracy Sweet, and Samrachana Adhikari (2025), "A Bayesian approach to estimate causal peer influence accounting for latent network homophily," *Journal of the Royal Statistical Society Series A: Statistics in Society*.

Wade, Sara and Vanda Inácio (2025), "Bayesian Dependent Mixture Models: A Predictive Comparison and Survey," *Statistical Science*, 40 (1), 81 – 108.

Wedel, Michel and Jie Zhang (2004), "Analyzing brand competition across subcategories," *Journal of Marketing Research*, 41 (4), 448–456.

Zhang, Hao (2004), "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics," *Journal of the American Statistical Association*, 99 (465), 250–261.

APPENDIX A: DETAILS ON MODEL TRAINING AND INFERENCE

We begin by describing the full conditionals for our proposed model DDPH. They are the distributions of each parameter conditional on all the other parameters and the observed data. We iteratively draw from each of those full conditionals to sample from the posterior distribution of interest.

Model Parameters

DP Parameters

1. The full conditional distribution of the cluster membership z_i is a categorical distribution:

$$p\left(z_i|\{V_k\}_{k=1}^{M-1}, \left\{\tilde{\boldsymbol{\beta}}_k(\boldsymbol{t})\right\}_{k=1}^{M}, \{u_{ijtc}\}_{j,t,c}\right) = \sum_{k=1}^{M} \hat{\pi}_k \delta_k(z_i), \tag{20}$$

where $\delta_k(z_i)$ is a Dirac measure concentrated at $z_i = k$. The cluster weights $\hat{\pi}_k$, summing to 1, are

based on the prior weights $\pi_k = V_k \prod_{h<k}(1 - V_h)$ and the likelihood of u_{ijtc} conditional on $\tilde{\boldsymbol{\beta}}_k(\boldsymbol{t})$:

$$\begin{aligned}\hat{\pi}_k &= p\left(z_i = k | \{V_h\}_{h=1}^{M-1}, \left\{\tilde{\boldsymbol{\beta}}_h(\boldsymbol{t})\right\}_{h=1}^{M}, \{u_{ijtc}\}_{j,t,c}\right) \propto \pi_k \cdot p\left(\{u_{ijtc}\}_{j,t,c} | \tilde{\boldsymbol{\beta}}_k(\boldsymbol{t})\right) \\ &\propto \pi_k \cdot \prod_{j,t,c} \exp\left\{-\frac{1}{2}\left(u_{ijtc} - \boldsymbol{x}'_{ijtc}\tilde{\boldsymbol{\beta}}_{kt}\right)^2\right\}.\end{aligned} \tag{21}$$

2. The full conditional of the stick-breaking length V_k is a Beta distribution:

$$V_k | \{z_i\}_{i=1}^{N}, \alpha \sim \text{Beta}\left(1 + n_k, \alpha + \sum_{h=k+1}^{M} n_h\right), \quad \forall k \in \{1, \ldots, M-1\}, \tag{22}$$

where n_k is the cardinality of cluster membership set $S_k = \{i \in \{1, \ldots, N\} : z_i = k\}$. The stick-breaking weights are then computed as

$$\pi_k = V_k \prod_{h<k}(1 - V_h), \quad \forall k \in \{1, \ldots, M-1\}; \quad \pi_M = 1 - \sum_{k=1}^{M-1} \pi_k. \tag{23}$$

3. The full conditional for the concentration parameter α is a Gamma distribution because the prior is Gamma(a_0, b_0), and the likelihood is determined by the stick-breaking process that $\{V_k\}_{k=1}^{M-1}$ follows. Specifically, it is given by

$$\alpha | \{V_k\}_{k=1}^{M-1} \sim \text{Gamma}\left(a_0 + M - 1, b_0 - \sum_{k=1}^{M-1} \log(1 - V_k)\right). \tag{24}$$

Preference Parameters

4. In model training, we discretize the parameter trajectories at the monthly level. Thus, for a (training) dataset with T time periods (months), we have the function values $\tilde{\beta}_k^{(p)}(\boldsymbol{t})$ and $\boldsymbol{\mu}^{(p)} = \mu^{(p)}(\boldsymbol{t})$ of the cluster-specific parameter trajectory and the mean trajectory as T-dimensional vectors, respectively. Similarly, we have the function values $\mathbf{K}_{\boldsymbol{\theta}^{(p)}} = K_{\boldsymbol{\theta}^{(p)}}(\boldsymbol{t}, \boldsymbol{t}')$ of the kernel, represented as a $T \times T$ covariance matrix. The full conditional for the cluster-level preference parameters thus

follows a Gaussian distribution due to the Gaussian-Gaussian conjugacy:

$$
\begin{aligned}
& p\left(\tilde{\beta}_k^{(p)}(\boldsymbol{t})|\boldsymbol{\mu}^{(p)}, \boldsymbol{\theta}^{(p)}, \left\{\{u_{ijtc}\}_{j,t,c}\right\}_{i\in S_k}, \tilde{\boldsymbol{\beta}}_k^{(-p)}(\boldsymbol{t})\right) \\
& \propto \underbrace{p\left(\tilde{\beta}_k^{(p)}(\boldsymbol{t})|\boldsymbol{\mu}^{(p)}, \boldsymbol{\theta}^{(p)}\right)}_{\mathcal{N}\left(\boldsymbol{\mu}^{(p)}, \mathbf{K}_{\boldsymbol{\theta}^{(p)}}\right)} \cdot \underbrace{p\left(\left\{\{u_{ijtc}\}_{j,t,c}\right\}_{i\in S_k}|\tilde{\boldsymbol{\beta}}_k(\boldsymbol{t})\right)}_{\prod_{i\in S_k}\prod_{j,t,c}\mathcal{N}\left(\boldsymbol{x}'_{ijtc}\tilde{\boldsymbol{\beta}}_k(t),1\right)},
\end{aligned}
\tag{25}
$$

where $\tilde{\boldsymbol{\beta}}_k^{(-p)}(\boldsymbol{t})$ is a matrix collecting preference vectors $\left\{\tilde{\beta}_k^{(p')}(\boldsymbol{t})\right\}_{p'\neq p}$. The full conditional is given by

$$
\tilde{\beta}_k^{(p)}(\boldsymbol{t})|\boldsymbol{\mu}^{(p)}, \boldsymbol{\theta}^{(p)}, \left\{\{u_{ijtc}\}_{j,t,c}\right\}_{i\in S_k}, \tilde{\boldsymbol{\beta}}_k^{(-p)}(\boldsymbol{t}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\text{post},k}^{(p)}, \Sigma_{\text{post},k}^{(p)}\right), \tag{26}
$$

where the mean and covariance are derived as follows. They are different depending on whether the p-th parameter is associated with a brand-specific intercept or an attribute coefficient.

For the intercepts, $p \in \{1, \ldots, C-1\}$, the distribution of the scalar residual utility term $\tilde{u}_{ijt}^{(p)} = u_{ijtp} - \sum_{p'\geq C} x_{ijtp}^{(p')} \cdot \beta_i^{(p')}(t)$ follows $\mathcal{N}\left(\beta_i^{(p)}(t), 1\right)$, conditional on all the other attribute parameters. Then,

$$
\Sigma_{\text{post},k}^{-1(p)} = \mathbf{K}_{\boldsymbol{\theta}^{(p)}}^{-1} + \sum_{i\in S_k, j} \mathbf{I}_T, \quad \boldsymbol{\mu}_{\text{post},k}^{(p)} = \Sigma_{\text{post},k}^{(p)}\left(\mathbf{K}_{\boldsymbol{\theta}^{(p)}}^{-1}\boldsymbol{\mu}^{(p)} + \sum_{i\in S_k, j} \tilde{\boldsymbol{u}}_{ij}^{(p)}\right), \tag{27}
$$

where $\tilde{\boldsymbol{u}}_{ij}^{(p)} = \left(\tilde{u}_{ij1}^{(p)}, \ldots, \tilde{u}_{ijT}^{(p)}\right)'$, and $\mathbf{I}_T$ is a $T \times T$ identity matrix.

For the attribute coefficients, $p \in \{C, \ldots, P\}$, conditional on all other attribute parameters, the distribution of the scalar residual utility term $\tilde{u}_{ijtc}^{(p)} = u_{ijtc} - \sum_{p'\neq p} x_{ijtc}^{(p')} \cdot \beta_i^{(p')}(t)$ follows $\mathcal{N}\left(x_{ijtc}^{(p)} \cdot \beta_i^{(p)}(t), 1\right)$. After stacking the $T \times 1$ vector $\tilde{\boldsymbol{u}}_{ijc}^{(p)} = \left(\tilde{u}_{ij1c}^{(p)}, \ldots, \tilde{u}_{ijTc}^{(p)}\right)'$ and the $T \times 1$ attribute vector $\boldsymbol{x}_{ijc}^{(p)} = \left(x_{ij1c}^{(p)}, \cdots, x_{ijTc}^{(p)}\right)'$, we have the conditional precision and mean as

$$
\begin{aligned}
\Sigma_{\text{post},k}^{-1(p)} &= \mathbf{K}_{\boldsymbol{\theta}^{(p)}}^{-1} + \sum_{i\in S_k, j, c} \text{diag}\left(\boldsymbol{x}_{ijc}^{(p)} \odot \boldsymbol{x}_{ijc}^{(p)}\right) \\
\boldsymbol{\mu}_{\text{post},k}^{(p)} &= \Sigma_{\text{post},k}^{(p)}\left(\mathbf{K}_{\boldsymbol{\theta}^{(p)}}^{-1}\boldsymbol{\mu}^{(p)} + \sum_{i\in S_k, j, c} \boldsymbol{x}_{ijc}^{(p)} \odot \tilde{\boldsymbol{u}}_{ijc}^{(p)}\right),
\end{aligned}
\tag{28}
$$

where $\odot$ denotes the elementwise product and diag $\left(\boldsymbol{x}_{ijc}^{(p)} \odot \boldsymbol{x}_{ijc}^{(p)}\right)$ is a $T \times T$ diagonal matrix.

Utility

5. The full conditional distribution for the utility u_{ijtc} is a truncated normal distribution with the lower bound $\underline{\ell}_{ijtc}$ and upper bound $\overline{\ell}_{ijtc}$ determined by the observed choice y_{ijt} because this augmented utility is assumed to be normally distributed (Albert and Chib 1993; McCulloch and Rossi 1994):

$$p(u_{ijtc}|\boldsymbol{\beta}_{it}, \{u_{ijtc}\}_{c' \neq c}) = \mathcal{TN}\left(\boldsymbol{x}'_{ijtc}\boldsymbol{\beta}_{it}, 1, \underline{\ell}_{ijtc}, \overline{\ell}_{ijtc}\right), \tag{29}$$

where $\underline{\ell}_{ijtc} = -\infty$ and $\overline{\ell}_{ijtc} = \max_{c' \neq c}\{u_{ijtc'}\}$ if $y_{ijt} \neq c$, and $\underline{\ell}_{ijtc} = \max_{c' \neq c}\{u_{ijtc'}\}$ and $\overline{\ell}_{ijtc} = \infty$ if $y_{ijt} = c$.

GP parameters

6. The full conditional for the mean function values $\boldsymbol{\mu}^{(p)}$ of the GP is a multivariate normal distribution because the prior distribution and the likelihood are both multivariate normal distributions:

$$p\left(\boldsymbol{\mu}^{(p)}|\boldsymbol{\theta}_0, \boldsymbol{\theta}^{(p)}, \left\{\tilde{\beta}_k^{(p)}(\boldsymbol{t})\right\}_{k=1}^{M}\right) \propto p\left(\boldsymbol{\mu}^{(p)}|\boldsymbol{\theta}_0\right) \cdot p\left(\left\{\tilde{\beta}_k^{(p)}(\boldsymbol{t})\right\}_{k=1}^{M}|\boldsymbol{\mu}^{(p)}, \boldsymbol{\theta}^{(p)}\right), \tag{30}$$

which results in $\boldsymbol{\mu}^{(p)}|\boldsymbol{\theta}_0, \boldsymbol{\theta}^{(p)}, \left\{\tilde{\beta}_k^{(p)}(\boldsymbol{t})\right\}_{k=1}^{M} \sim \mathcal{N}\left(\boldsymbol{m}_{\text{post}}^{(p)}, \Omega_{\text{post}}^{(p)}\right)$. Here,

$$\Omega_{\text{post}}^{-1(p)} = \mathbf{K}_{\boldsymbol{\theta}_0}^{-1} + M \cdot \mathbf{K}_{\boldsymbol{\theta}^{(p)}}^{-1}, \quad \boldsymbol{m}_{\text{post}}^{(p)} = \Omega_{\text{post}}^{(p)}\mathbf{K}_{\boldsymbol{\theta}^{(p)}}^{-1}\sum_{k=1}^{M}\tilde{\beta}_k^{(p)}(\boldsymbol{t}), \tag{31}$$

where $\mathbf{K}_{\boldsymbol{\theta}_0} = K_{\boldsymbol{\theta}_0}(\boldsymbol{t}, \boldsymbol{t}')$ is a $T \times T$ matrix of the function values of the GP kernel.

7. The full conditional distributions of the kernel parameters $\boldsymbol{\theta}^{(p)} = \left(\kappa^{(p)}, \sigma^{2(p)}\right)'$ and $\boldsymbol{\theta}_0 = \left(\kappa_0, \sigma_0^2\right)'$ are not analytically available. This is because each $\boldsymbol{\theta}^{(p)}$ and $\boldsymbol{\theta}_0$ enters the normal likelihood nonlinearly through its own kernel and the prior is log-normal, resulting in a non-conjugate setup. Thus, we use elliptical slice sampling (Murray, Adams, and MacKay 2010) to draw the parameters $\boldsymbol{\theta}^{(p)}$ and $\boldsymbol{\theta}_0$ from their full conditionals in each MCMC iteration.

Posterior Predictive Distributions

1. Let $\beta_{\text{new}}^{(p)}(\cdot)$ represent a trajectory of a new customer randomly drawn from the population. Then, its posterior predictive distribution is given by $\beta_{\text{new}}^{(p)}(\cdot) \sim \int p(\cdot|\Theta)p(\Theta|\text{data})d\Theta$, where Θ collects all the relevant parameters based on the model specification and the conditioning variable (data) contains all the observed attributes and choices. In GPH, the relevant parameters are $\Theta = \left(\mu^{(p)}(\cdot), \boldsymbol{\theta}^{(p)}\right)$, and the posterior predictive distribution of $\beta_{\text{new}}^{(p)}(\boldsymbol{t})$ is given by $\mathcal{N}\left(\boldsymbol{\mu}^{(p)}, \mathbf{K}_{\boldsymbol{\theta}^{(p)}}\right)$. In DDPH, the relevant parameters are $\Theta = \left(\left\{\tilde{\beta}_k^{(p)}(\cdot), \pi_k\right\}_{k=1}^{M}\right)$. The posterior predictive distribution of $\beta_{\text{new}}^{(p)}(\boldsymbol{t})$ is given by

$$\sum_{k=1}^{M} \pi_k \delta_{\tilde{\beta}_k^{(p)}(\boldsymbol{t})}(\cdot), \tag{32}$$

where $\delta_{\tilde{\beta}_k^{(p)}(\boldsymbol{t})}(\cdot)$ is a Dirac measure concentrated at $\tilde{\beta}_k^{(p)}(\boldsymbol{t})$.

2. Let $\beta_i^{(p)}(\boldsymbol{t}^*)$ be the function values of training-set consumer i's trajectory evaluated at the out-of-sample (future) time periods $\boldsymbol{t}^* \in \mathbb{T}^{T_{\text{test}}}$. Their posterior predictive distribution is given as follows. Note that the extrapolation procedure below is the same across the two models. First, we extrapolate $\mu^{(p)}(\boldsymbol{t}^*) \sim \mathcal{N}(\boldsymbol{\varphi}_0, \Lambda_0)$ based on $\mu^{(p)}(\boldsymbol{t})$ and $\boldsymbol{\theta}_0$, where

$$\begin{aligned} \boldsymbol{\varphi}_0 &= K_{\boldsymbol{\theta}_0}(\boldsymbol{t}^*, \boldsymbol{t}')K_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{t}, \boldsymbol{t}')\mu^{(p)}(\boldsymbol{t}) \\ \Lambda_0 &= K_{\boldsymbol{\theta}_0}(\boldsymbol{t}^*, \boldsymbol{t}^{*'}) - K_{\boldsymbol{\theta}_0}(\boldsymbol{t}^*, \boldsymbol{t}')K_{\boldsymbol{\theta}_0}^{-1}(\boldsymbol{t}, \boldsymbol{t}')K_{\boldsymbol{\theta}_0}(\boldsymbol{t}, \boldsymbol{t}^{*'}). \end{aligned} \tag{33}$$

Here, $\boldsymbol{\varphi}_0$ is a T_{test}-dimensional conditional mean vector, and Λ_0 is a $T_{\text{test}} \times T_{\text{test}}$ conditional covariance matrix. Next, we extrapolate $\beta_i^{(p)}(\boldsymbol{t}^*) \sim \mathcal{N}(\boldsymbol{\varphi}, \Lambda)$ based on $\beta_i^{(p)}(\boldsymbol{t})$, $\mu^{(p)}(\boldsymbol{t})$, and $\boldsymbol{\theta}^{(p)}$ as well as the extrapolated $\mu^{(p)}(\boldsymbol{t}^*)$, where

$$\begin{aligned} \boldsymbol{\varphi} &= \mu^{(p)}(\boldsymbol{t}^*) + K_{\boldsymbol{\theta}^{(p)}}(\boldsymbol{t}^*, \boldsymbol{t}')K_{\boldsymbol{\theta}^{(p)}}^{-1}(\boldsymbol{t}, \boldsymbol{t}')\left(\beta_i^{(p)}(\boldsymbol{t}) - \mu^{(p)}(\boldsymbol{t})\right) \\ \Lambda &= K_{\boldsymbol{\theta}^{(p)}}(\boldsymbol{t}^*, \boldsymbol{t}^{*'}) - K_{\boldsymbol{\theta}^{(p)}}(\boldsymbol{t}^*, \boldsymbol{t}')K_{\boldsymbol{\theta}^{(p)}}^{-1}(\boldsymbol{t}, \boldsymbol{t}')K_{\boldsymbol{\theta}^{(p)}}(\boldsymbol{t}, \boldsymbol{t}^{*'}), \end{aligned} \tag{34}$$

Here, $\boldsymbol{\varphi}$ is a T_{test}-dimensional vector, and Λ is a $T_{\text{test}} \times T_{\text{test}}$ matrix.

WEB APPENDIX A: MORE ON SIMULATION STUDIES

Data-Generating Processes

We generate eight datasets using different random seeds for each data-generating process. In each dataset, we fix the numbers of brands and attributes to $C = 3$ and 2, respectively, so that $P = C - 1 + 2 = 4$ is the total number of parameter in the vector $\boldsymbol{\beta}_{it}$. We set the number of consumers to $N = 500$ and the number of calendar time points to $T = 11$. To mirror our empirical application, in which a consumer's number of purchase occasions varies across time periods, we set the per-consumer, per-period observation count to a value between 0 and $J = 7$. We hold out $T_{\text{test}} = 3$ periods for out-of-sample prediction. The attribute values are drawn independently from a normal distribution,

$$x_{ijtc}^{(p)} \sim \mathcal{N}(0.5, 1) \tag{35}$$

for each $c \in \{1, \ldots, C\}$ and $p \in \{C, C+1, \ldots, P\}$ for each observation. For $p \in \{1, \ldots, C-1\}$, the scalar $x_{ijtc}^{(p)}$ is a brand-specific intercept and takes 1 if $p = c$ and 0 otherwise.

In each DGP, the hyperpriors of the Gaussian process components are specified as

$$\begin{aligned}
\boldsymbol{\theta}^{(p)} &\sim \text{LogNorm}\Big((1, -1)',\ 0.1 \cdot \mathbf{I}_2\Big), \quad \forall p \in \{1, \ldots, P\} \\
\boldsymbol{\theta}_0 &\sim \text{LogNorm}\Big((1, -1)',\ 0.1 \cdot \mathbf{I}_2\Big),
\end{aligned}$$

where $\mathbf{I}_2$ is a 2×2 identity matrix.

Estimation Results for the Individual-Level Parameters

We further investigate the performance of the models in estimating the individual-level parameters. While we look at the MSE values in the main part of the manuscript, we compute different quantiles, $\tau \in \{0.25, 0.5, 0.75\}$, of the population distributions here. We then compare the posterior

mean of these quantiles for each time period:

$$\frac{1}{R}\sum_{r=1}^{R} q\left(\left\{\beta_i^{(p),(r)}(t)\right\}_{i=1}^{n}, \tau\right). \tag{36}$$

Tables 9, 10, and 11 report the posterior means of the different population quantiles of the individual-level parameters under D1, D2, and D3, respectively. Overall, when the true data-generating process exhibits complex consumer preference patterns such as multimodality (D1 and D3), we find that our model recovers the quantiles more accurately than the benchmark. In D2, where the benchmark model is correctly specified, its overall performance is slightly better than our model. However, both models estimate the quantiles of the individual-level parameters well when the underlying cross-sectional heterogeneity is characterized by a normal distribution for each time period.

Table 9: Quantile estimates under D1.

	25th			Median			75th		
	DDPH	True	GPH	DDPH	True	GPH	DDPH	True	GPH
$\beta^{\text{icept}_1}(1)$	**-1.22** (.09)	-1.25	-0.90 (.05)	**-0.70** (.14)	-0.66	-0.62 (.05)	-0.28 (.14)	-0.35	**-0.33** (.05)
$\beta^{\text{icept}_1}(2)$	-1.14 (.07)	-1.03	**-1.05** (.03)	**-0.71** (.10)	-0.60	-0.77 (.03)	**-0.39** (.07)	-0.43	-0.48 (.04)
$\beta^{\text{icept}_1}(3)$	**-1.17** (.07)	-1.12	-1.22 (.02)	**-0.94** (.06)	-0.92	-0.94 (.03)	**-0.54** (.10)	-0.53	-0.66 (.04)
$\beta^{\text{icept}_1}(4)$	**-1.28** (.04)	-1.22	-1.35 (.03)	**-1.06** (.05)	-0.98	-1.08 (.03)	**-0.86** (.07)	-0.96	-0.80 (.04)
$\beta^{\text{icept}_1}(5)$	**-1.26** (.04)	-1.09	-1.40 (.04)	**-1.11** (.04)	-1.02	-1.12 (.04)	**-0.95** (.08)	-1.00	-0.84 (.04)
$\beta^{\text{icept}_1}(6)$	**-1.24** (.04)	-1.16	-1.32 (.05)	**-1.07** (.05)	-1.14	-1.05 (.04)	-0.84 (.10)	-0.75	**-0.77** (.04)
$\beta^{\text{icept}_1}(7)$	**-1.14** (.06)	-1.15	-1.17 (.05)	-0.92 (.07)	-0.90	**-0.89** (.04)	**-0.72** (.09)	-0.69	-0.62 (.04)
$\beta^{\text{icept}_1}(8)$	-0.97 (.09)	-0.99	**-0.98** (.04)	-0.74 (.05)	-0.64	**-0.70** (.03)	**-0.53** (.07)	-0.61	-0.42 (.03)
$\beta^{(1)}(1)$	**0.27** (.03)	0.28	0.22 (.02)	**0.45** (.02)	0.45	0.39 (.02)	**0.55** (.02)	0.47	0.57 (.02)
$\beta^{(1)}(2)$	**0.24** (.02)	0.24	0.17 (.01)	**0.41** (.01)	0.43	0.34 (.01)	**0.46** (.01)	0.43	0.51 (.01)
$\beta^{(1)}(3)$	**0.23** (.01)	0.26	0.13 (.01)	**0.29** (.01)	0.29	0.29 (.01)	**0.33** (.01)	0.30	0.45 (.01)
$\beta^{(1)}(4)$	**0.11** (.03)	0.11	0.10 (.01)	0.24 (.01)	0.26	**0.25** (.01)	**0.27** (.01)	0.27	0.41 (.01)
$\beta^{(1)}(5)$	**0.15** (.02)	0.20	0.06 (.01)	**0.26** (.02)	0.26	0.23 (.01)	**0.33** (.02)	0.32	0.39 (.01)
$\beta^{(1)}(6)$	**0.01** (.02)	0.04	-0.00 (.01)	**0.32** (.01)	0.33	0.20 (.01)	**0.37** (.01)	0.34	0.38 (.01)
$\beta^{(1)}(7)$	**-0.17** (.03)	-0.14	-0.09 (.01)	**0.13** (.02)	0.13	0.13 (.01)	**0.47** (.02)	0.43	0.35 (.01)
$\beta^{(1)}(8)$	**-0.29** (.04)	-0.28	-0.15 (.01)	**-0.17** (.02)	-0.18	0.05 (.01)	**0.47** (.02)	0.44	0.29 (.01)

Standard deviations across datasets are presented in parentheses.
Mean values closer to the true values are in bold.

Table 10: Quantile estimates under D2.

	25th			Median			75th		
	DDPH	True	GPH	DDPH	True	GPH	DDPH	True	GPH
$\beta^{\text{icept}_1}(1)$	**-1.53** (.07)	-1.62	-1.71 (.08)	-1.18 (.05)	-1.24	**-1.26** (.07)	-0.82 (.07)	-0.79	**-0.82** (.06)
$\beta^{\text{icept}_1}(2)$	**-1.53** (.05)	-1.63	-1.72 (.06)	-1.18 (.04)	-1.25	**-1.28** (.04)	-0.83 (.07)	-0.81	**-0.83** (.04)
$\beta^{\text{icept}_1}(3)$	-1.53 (.03)	-1.63	**-1.72** (.03)	-1.18 (.04)	-1.23	**-1.27** (.03)	**-0.84** (.05)	-0.83	-0.83 (.04)
$\beta^{\text{icept}_1}(4)$	**-1.53** (.02)	-1.61	-1.72 (.02)	-1.19 (.04)	-1.25	**-1.28** (.03)	**-0.85** (.05)	-0.87	-0.83 (.05)
$\beta^{\text{icept}_1}(5)$	**-1.55** (.03)	-1.61	-1.73 (.03)	-1.19 (.04)	-1.28	**-1.29** (.04)	**-0.85** (.06)	-0.88	-0.84 (.06)
$\beta^{\text{icept}_1}(6)$	-1.56 (.05)	-1.72	**-1.75** (.03)	-1.20 (.05)	-1.28	**-1.30** (.04)	-0.84 (.05)	-0.88	**-0.85** (.05)
$\beta^{\text{icept}_1}(7)$	-1.56 (.06)	-1.73	**-1.74** (.04)	-1.20 (.06)	-1.28	**-1.29** (.03)	-0.84 (.04)	-0.89	**-0.85** (.04)
$\beta^{\text{icept}_1}(8)$	-1.52 (.06)	-1.71	**-1.70** (.05)	-1.17 (.06)	-1.27	**-1.26** (.04)	-0.81 (.06)	-0.90	**-0.81** (.04)
$\beta^{(1)}(1)$	-0.04 (.01)	-0.04	**-0.05** (.02)	0.13 (.02)	0.16	**0.15** (.02)	0.32 (.02)	0.35	**0.35** (.02)
$\beta^{(1)}(2)$	-0.04 (.01)	-0.05	**-0.05** (.01)	0.13 (.02)	0.15	**0.15** (.02)	0.31 (.03)	0.34	**0.34** (.02)
$\beta^{(1)}(3)$	**-0.04** (.01)	-0.03	-0.05 (.01)	0.14 (.01)	0.15	**0.15** (.01)	0.30 (.02)	0.33	**0.34** (.02)
$\beta^{(1)}(4)$	**-0.03** (.02)	-0.03	-0.03 (.01)	0.16 (.01)	0.16	**0.16** (.01)	0.31 (.01)	0.35	**0.36** (.01)
$\beta^{(1)}(5)$	-0.02 (.02)	-0.03	**-0.02** (.02)	0.17 (.01)	0.18	**0.18** (.01)	0.34 (.02)	0.36	**0.37** (.01)
$\beta^{(1)}(6)$	-0.01 (.01)	-0.02	**-0.01** (.02)	0.17 (.02)	0.18	**0.19** (.01)	0.35 (.02)	0.37	**0.38** (.01)
$\beta^{(1)}(7)$	-0.01 (.01)	-0.03	**-0.01** (.01)	**0.16** (.01)	0.17	0.18 (.01)	0.34 (.01)	0.36	**0.38** (.01)
$\beta^{(1)}(8)$	**-0.02** (.03)	-0.02	-0.03 (.02)	**0.15** (.03)	0.15	0.17 (.02)	0.33 (.03)	0.35	**0.36** (.02)

Standard deviations across datasets are presented in parentheses.
Mean values closer to the true values are in bold.

Table 11: Quantile estimates under D3.

	25th			Median			75th		
	DDPH	True	GPH	DDPH	True	GPH	DDPH	True	GPH
$\beta^{\text{icept}_1}(1)$	**-1.65** (.17)	-1.56	-1.12 (.05)	**-1.00** (.06)	-1.28	-0.80 (.04)	**-0.28** (.11)	-0.31	-0.48 (.06)
$\beta^{\text{icept}_1}(2)$	**-1.53** (.10)	-1.47	-1.26 (.05)	**-1.07** (.07)	-1.05	-0.94 (.03)	**-0.48** (.11)	-0.45	-0.62 (.04)
$\beta^{\text{icept}_1}(3)$	-1.37 (.11)	-1.38	**-1.38** (.05)	**-1.18** (.05)	-1.19	-1.06 (.03)	**-0.74** (.07)	-0.60	-0.75 (.03)
$\beta^{\text{icept}_1}(4)$	**-1.31** (.08)	-1.26	-1.44 (.06)	**-1.23** (.09)	-1.18	-1.12 (.03)	**-0.94** (.05)	-0.99	-0.81 (.02)
$\beta^{\text{icept}_1}(5)$	**-1.22** (.05)	-1.09	-1.40 (.06)	**-1.03** (.07)	-1.00	-1.09 (.04)	**-0.92** (.05)	-0.94	-0.78 (.04)
$\beta^{\text{icept}_1}(6)$	**-1.12** (.07)	-1.13	-1.28 (.07)	**-0.83** (.09)	-0.79	-0.97 (.06)	**-0.74** (.09)	-0.72	-0.66 (.06)
$\beta^{\text{icept}_1}(7)$	**-0.93** (.10)	-0.90	-1.10 (.08)	**-0.68** (.09)	-0.73	-0.79 (.06)	**-0.61** (.08)	-0.64	-0.48 (.07)
$\beta^{\text{icept}_1}(8)$	**-0.76** (.11)	-0.70	-0.91 (.07)	-0.58 (.05)	-0.62	**-0.60** (.06)	**-0.46** (.06)	-0.55	-0.29 (.06)
$\beta^{(1)}(1)$	**-0.04** (.02)	-0.03	0.02 (.02)	0.23 (.03)	0.24	**0.24** (.02)	0.47 (.03)	0.46	**0.46** (.02)
$\beta^{(1)}(2)$	**-0.15** (.02)	-0.12	-0.06 (.01)	**0.17** (.03)	0.19	0.16 (.01)	**0.42** (.02)	0.43	0.38 (.02)
$\beta^{(1)}(3)$	**-0.09** (.01)	-0.09	-0.05 (.01)	0.14 (.03)	0.18	**0.14** (.01)	**0.31** (.02)	0.31	0.34 (.02)
$\beta^{(1)}(4)$	**0.10** (.01)	0.07	0.01 (.01)	**0.12** (.01)	0.14	0.17 (.01)	**0.23** (.01)	0.22	0.35 (.02)
$\beta^{(1)}(5)$	**0.13** (.02)	0.18	0.05 (.01)	**0.31** (.01)	0.28	0.24 (.01)	**0.34** (.02)	0.35	0.41 (.02)
$\beta^{(1)}(6)$	0.05 (.03)	0.05	**0.05** (.01)	**0.37** (.03)	0.35	0.29 (.01)	**0.43** (.02)	0.42	0.48 (.02)
$\beta^{(1)}(7)$	**-0.08** (.04)	-0.11	0.02 (.01)	**0.41** (.01)	0.39	0.31 (.01)	**0.42** (.02)	0.46	0.52 (.02)
$\beta^{(1)}(8)$	**-0.17** (.04)	-0.26	0.02 (.02)	**0.43** (.02)	0.40	0.30 (.02)	0.44 (.03)	0.50	**0.53** (.02)

Standard deviations across datasets are presented in parentheses.
Mean values closer to the true values are in bold.

Decomposing the MSE

The MSE values reported in Table 1 in the main text admit the standard bias-variance decomposition. Let $\bar{\beta}_i^{(p),(s)}(t) = R^{-1}\sum_r \hat{\beta}_i^{(p),(s,r)}(t)$ be the posterior mean of consumer i's p-th parameter in time period t on simulated dataset s. The cross-dataset MSE is decomposed as follows:

$$\text{MSE}_i^{(p,t)} = \underbrace{\left(\bar{\beta}_i^{(p)}(t) - \beta_i^{(p)}(t)\right)^2}_{\text{Bias}^2} + \underbrace{\frac{1}{S}\sum_{s=1}^{S}\left(\bar{\beta}_i^{(p),(s)}(t) - \bar{\beta}_i^{(p)}(t)\right)^2}_{\text{Variance}}, \tag{37}$$

where $\bar{\beta}_i^{(p)}(t) = S^{-1}\sum_s \bar{\beta}_i^{(p),(s)}(t)$ is the average posterior mean across the eight datasets. We averaged the squared bias and variance components as well as the MSE over the consumers and report them in Tables 12 to 14 for D1, D2, and D3, respectively. Under D1 and D3, where the population distributions are multimodal, DDPH's lower MSE is driven almost entirely by lower

bias, at the cost of a modest increase in variance, reflecting its flexibility in tracking complex distributions. GPH attains marginally lower bias and variance than DDPH under D2, where the GPH benchmark is correctly specified.

Table 12: Bias-variance decomposition of population mean MSE under D1.

	MSE (DDPH)	MSE (GPH)	Bias^2 (DDPH)	Bias^2 (GPH)	Var (DDPH)	Var (GPH)
$\beta^{\text{icept}_1}(1)$	0.083	0.359	0.018	0.344	0.065	0.015
$\beta^{\text{icept}_1}(2)$	0.057	0.209	0.011	0.192	0.047	0.016
$\beta^{\text{icept}_1}(3)$	0.036	0.103	0.006	0.087	0.030	0.016
$\beta^{\text{icept}_1}(4)$	0.029	0.039	0.004	0.025	0.025	0.015
$\beta^{\text{icept}_1}(5)$	0.029	0.032	0.007	0.016	0.022	0.016
$\beta^{\text{icept}_1}(6)$	0.033	0.062	0.008	0.046	0.026	0.016
$\beta^{\text{icept}_1}(7)$	0.032	0.069	0.010	0.053	0.022	0.016
$\beta^{\text{icept}_1}(8)$	0.038	0.061	0.013	0.048	0.026	0.013
$\beta^{(1)}(1)$	0.007	0.027	0.002	0.014	0.006	0.013
$\beta^{(1)}(2)$	0.008	0.030	0.002	0.016	0.006	0.014
$\beta^{(1)}(3)$	0.006	0.027	0.001	0.013	0.005	0.014
$\beta^{(1)}(4)$	0.006	0.022	0.001	0.008	0.005	0.013
$\beta^{(1)}(5)$	0.008	0.022	0.002	0.009	0.006	0.013
$\beta^{(1)}(6)$	0.010	0.024	0.002	0.012	0.008	0.012
$\beta^{(1)}(7)$	0.013	0.030	0.003	0.019	0.010	0.011
$\beta^{(1)}(8)$	0.017	0.055	0.004	0.044	0.013	0.010

Table 13: Bias-variance decomposition of population mean MSE under D2.

	MSE (DDPH)	MSE (GPH)	Bias^2 (DDPH)	Bias^2 (GPH)	Var (DDPH)	Var (GPH)
$\beta^{\text{icept}_1}(1)$	0.338	0.295	0.264	0.230	0.074	0.065
$\beta^{\text{icept}_1}(2)$	0.321	0.267	0.241	0.195	0.080	0.072
$\beta^{\text{icept}_1}(3)$	0.310	0.252	0.231	0.179	0.078	0.072
$\beta^{\text{icept}_1}(4)$	0.301	0.241	0.226	0.169	0.075	0.072
$\beta^{\text{icept}_1}(5)$	0.307	0.243	0.232	0.172	0.075	0.070
$\beta^{\text{icept}_1}(6)$	0.327	0.259	0.249	0.189	0.078	0.070
$\beta^{\text{icept}_1}(7)$	0.347	0.277	0.268	0.209	0.078	0.068
$\beta^{\text{icept}_1}(8)$	0.374	0.307	0.304	0.249	0.071	0.058
$\beta^{(1)}(1)$	0.049	0.037	0.030	0.024	0.019	0.013
$\beta^{(1)}(2)$	0.040	0.029	0.023	0.016	0.017	0.013
$\beta^{(1)}(3)$	0.035	0.025	0.018	0.012	0.017	0.013
$\beta^{(1)}(4)$	0.034	0.025	0.017	0.012	0.017	0.013
$\beta^{(1)}(5)$	0.036	0.027	0.019	0.014	0.018	0.013
$\beta^{(1)}(6)$	0.039	0.029	0.020	0.015	0.019	0.013
$\beta^{(1)}(7)$	0.040	0.029	0.021	0.015	0.019	0.014
$\beta^{(1)}(8)$	0.048	0.036	0.029	0.022	0.019	0.014

Table 14: Bias-variance decomposition of population mean MSE under D3.

	MSE (DDPH)	MSE (GPH)	Bias^2 (DDPH)	Bias^2 (GPH)	Var (DDPH)	Var (GPH)
$\beta^{\text{icept}_1}(1)$	0.113	0.455	0.035	0.436	0.077	0.019
$\beta^{\text{icept}_1}(2)$	0.075	0.289	0.024	0.269	0.051	0.020
$\beta^{\text{icept}_1}(3)$	0.050	0.125	0.019	0.107	0.030	0.018
$\beta^{\text{icept}_1}(4)$	0.040	0.042	0.012	0.024	0.028	0.018
$\beta^{\text{icept}_1}(5)$	0.036	0.041	0.011	0.022	0.025	0.019
$\beta^{\text{icept}_1}(6)$	0.041	0.076	0.013	0.055	0.028	0.021
$\beta^{\text{icept}_1}(7)$	0.041	0.065	0.012	0.041	0.028	0.024
$\beta^{\text{icept}_1}(8)$	0.044	0.057	0.017	0.035	0.027	0.022
$\beta^{(1)}(1)$	0.016	0.029	0.007	0.012	0.009	0.017
$\beta^{(1)}(2)$	0.014	0.033	0.006	0.016	0.008	0.017
$\beta^{(1)}(3)$	0.012	0.031	0.006	0.015	0.006	0.016
$\beta^{(1)}(4)$	0.010	0.026	0.005	0.010	0.005	0.016
$\beta^{(1)}(5)$	0.011	0.028	0.005	0.012	0.006	0.016
$\beta^{(1)}(6)$	0.014	0.029	0.006	0.014	0.009	0.015
$\beta^{(1)}(7)$	0.017	0.035	0.007	0.021	0.011	0.015
$\beta^{(1)}(8)$	0.023	0.051	0.010	0.037	0.013	0.014

Heterogeneity Distribution Shift

In the main text, we examine how well our model and the benchmark are able to track shifts in the cross-sectional preference distributions over time under D3, using parameter of one attribute as an illustrative instance. We now report the results for all parameters under all the data-generating processes, D1, D2, and D3, in more detail.

Figure 14 shows the results for D2, where the benchmark GPH model is correctly specified. While the DDPH model captures the shifts in some time periods less precisely, its performance is comparable to or even better than that of the benchmark in other periods.

Figure 13 and Figure 15 report the results for D1 and D3, respectively. We see from the figures that our DDPH model recovers the shifts more precisely than the benchmark GPH model. This difference in performance may stem from the benchmark's assumption that the cross-sectional distribution is Gaussian in any time period, limiting its ability to characterize intertemporal shifts in distributions accurately when the true consumer preference patterns exhibit complex evolution patterns.

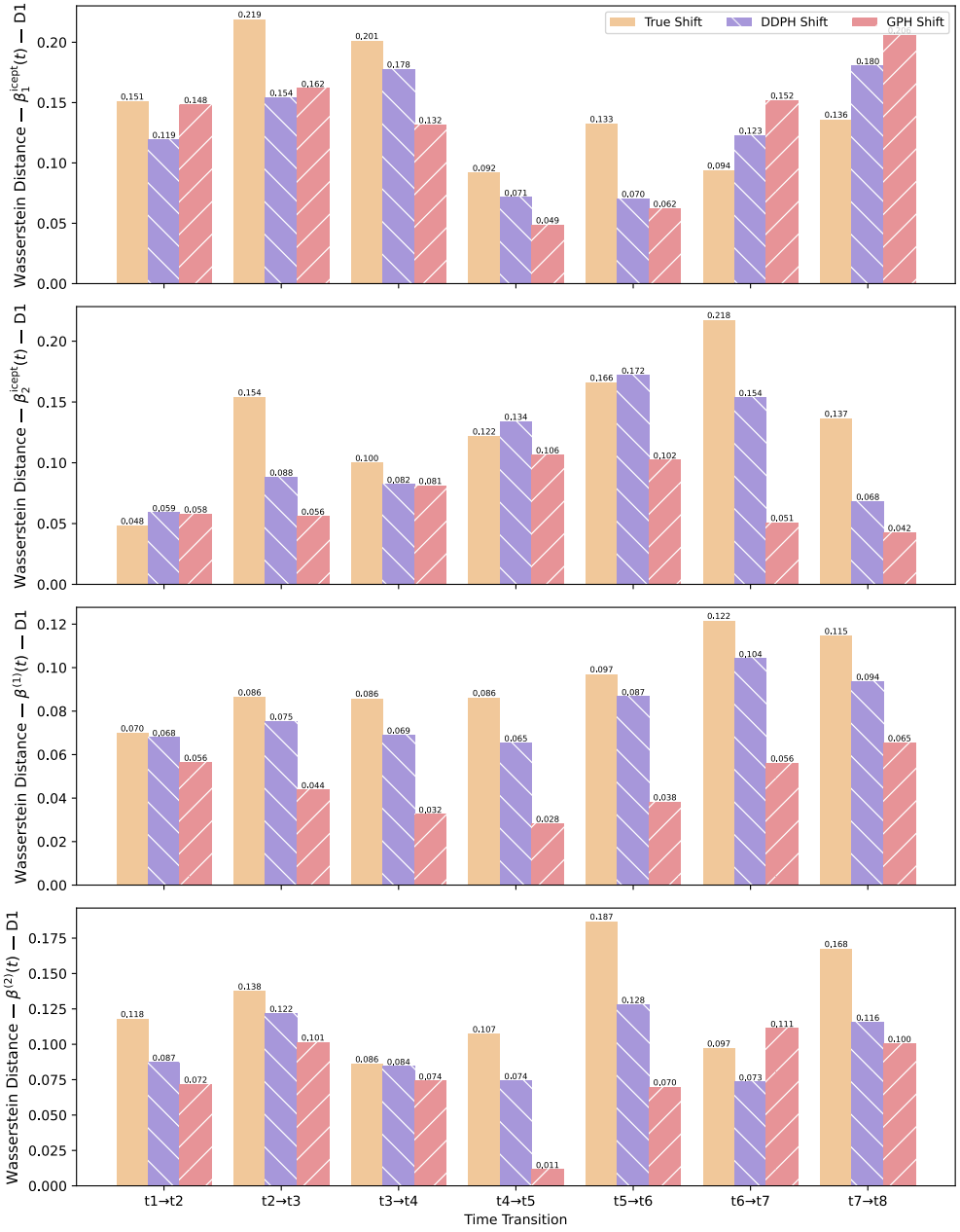

Figure 13: Intertemporal heterogeneity distribution divergence of all preference parameters (intercepts and coefficients), measured by Wasserstein distance D1.

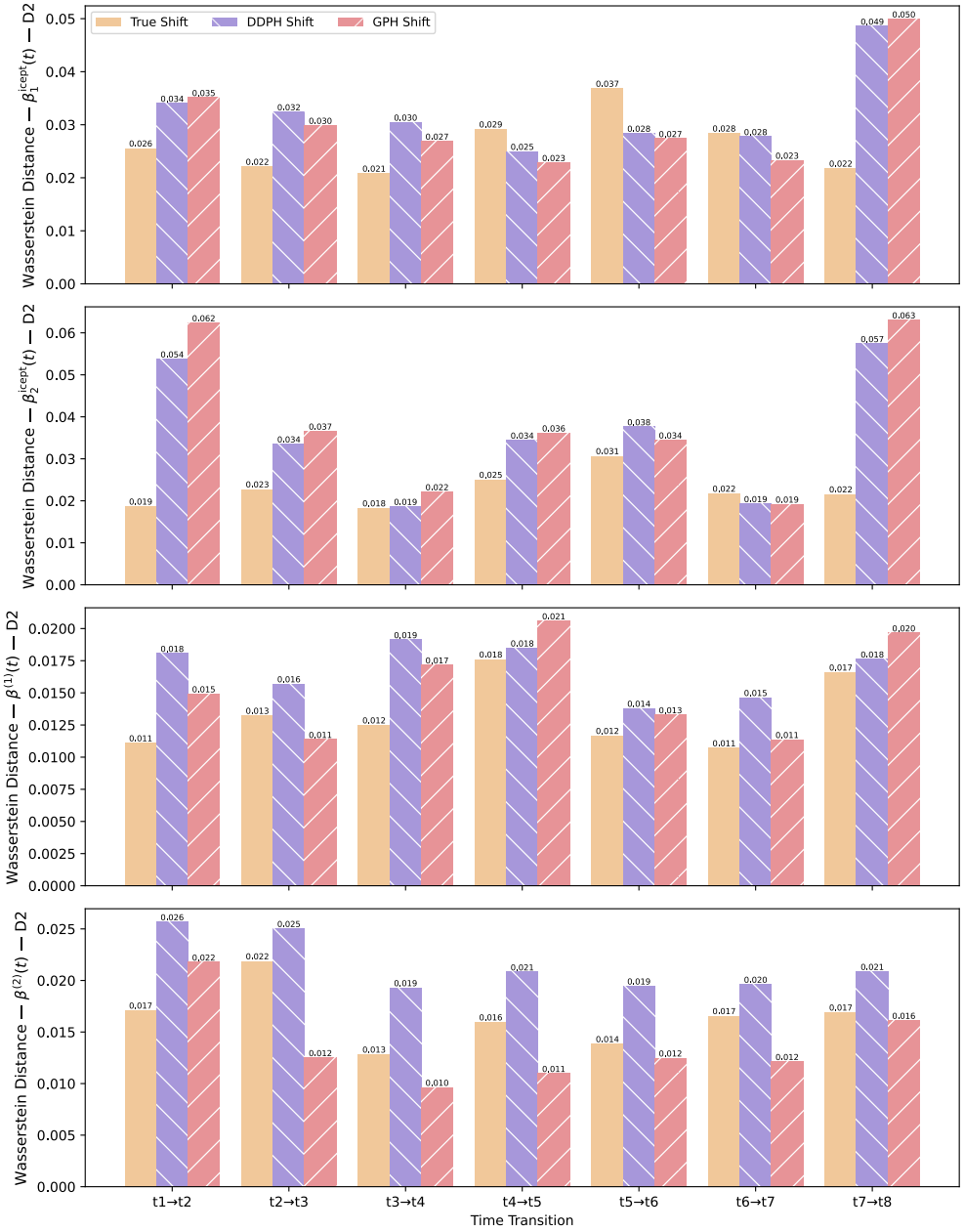

Figure 14: Intertemporal heterogeneity distribution divergence of all preference parameters (intercepts and coefficients), measured by Wasserstein distance D2.

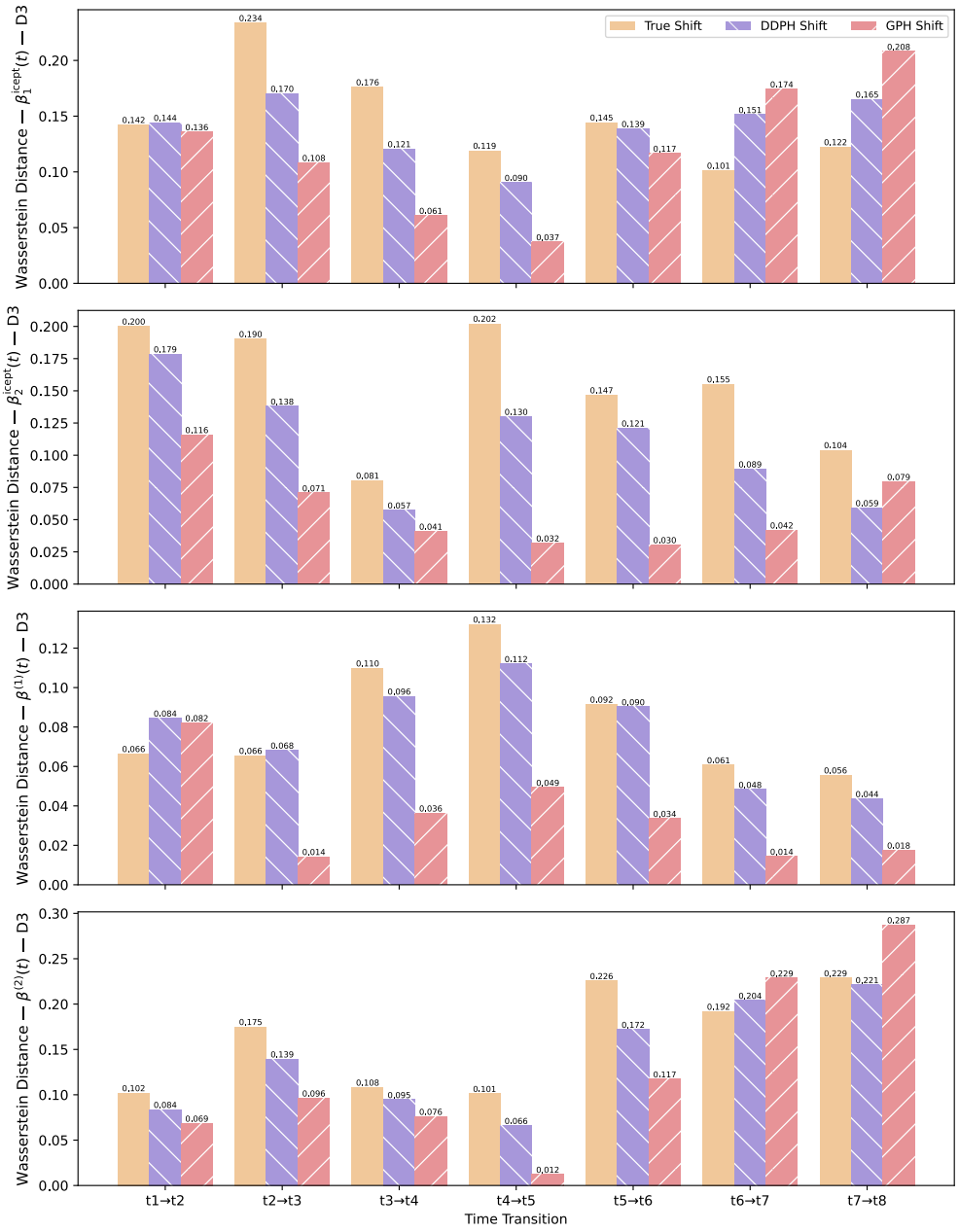

Figure 15: Intertemporal heterogeneity distribution divergence of all preference parameters (intercepts and coefficients), measured by Wasserstein distance D3.

Predictive performance

We supplement the overall prediction accuracy reported in the Predictive Performance subsection of the main text with two further metrics commonly used to assess classification performance: macro-averaged recall (the average across classes of sensitivity, where the per-class sensitivity

is the proportion of observations in product class c that the model correctly predicts as c) and macro-averaged specificity (the average across classes of specificity, where the per-class specificity is the proportion of observations of choices not in product class c that the model correctly predicts as not belonging to class c). The computation of these two metrics follows the procedure that we use to compute the accuracy metric reported in the main text. Tables 15 and 16 present these metrics under D1, D2, and D3. The relative performance of DDPH and GPH is consistent with that for accuracy across all DGPs and forecast horizons. DDPH outperforms GPH under D1 and D3, whereas GPH performs slightly better under D2. Moreover, the performance difference between the two models diminishes as the forecast horizon increases.

Table 15: Macro-averaged recall under D1, D2, and D3.

DGP	Model	Train (all)	Test (all)	Test ($t = 9$)	Test ($t = 10$)	Test ($t = 11$)
D1	DDPH	**0.926** (.01)	**0.642** (.03)	**0.810** (.01)	**0.640** (.04)	**0.481** (.05)
	GPH	0.824 (.00)	0.523 (.02)	0.596 (.01)	0.514 (.03)	0.466 (.04)
D2	DDPH	0.806 (.00)	0.611 (.03)	0.681 (.01)	0.611 (.03)	0.545 (.04)
	GPH	**0.813** (.00)	**0.615** (.03)	**0.685** (.01)	**0.616** (.03)	**0.546** (.04)
D3	DDPH	**0.908** (.01)	**0.644** (.02)	**0.789** (.02)	**0.638** (.02)	**0.500** (.04)
	GPH	0.837 (.00)	0.523 (.02)	0.614 (.01)	0.520 (.02)	0.432 (.03)

Standard deviations across datasets are presented in parentheses.

Higher recall values are in bold.

Table 16: Macro-averaged specificity under D1, D2, and D3.

DGP	Model	Train (all)	Test (all)	Test ($t = 9$)	Test ($t = 10$)	Test ($t = 11$)
D1	DDPH	**0.964** (.00)	**0.820** (.02)	**0.907** (.01)	**0.821** (.02)	**0.735** (.03)
	GPH	0.913 (.00)	0.758 (.01)	0.798 (.00)	0.755 (.01)	0.725 (.02)
D2	DDPH	0.904 (.00)	0.803 (.01)	0.839 (.01)	0.804 (.01)	0.769 (.02)
	GPH	**0.907** (.00)	**0.805** (.01)	**0.841** (.00)	**0.806** (.01)	**0.770** (.02)
D3	DDPH	**0.954** (.00)	**0.818** (.01)	**0.895** (.01)	**0.816** (.01)	**0.743** (.02)
	GPH	0.919 (.00)	0.757 (.01)	0.806 (.01)	0.756 (.01)	0.711 (.01)

Standard deviations across datasets are presented in parentheses.

Higher specificity values are in bold.

Own-Attribute Elasticity under D2 and D3

The main text presents the own-attribute elasticity for D1 (Figure 6). Here we report the own-attribute elasticities for the remaining two data-generating processes (D2 and D3). Figures 16 and 17 show the posterior mean estimates of the DDPH model (blue, dash-dotted line) and the GPH benchmark (red line with circular markers), together with the true elasticity (gray solid line), for the three brands across the eight training periods. Shaded regions denote 95% credible intervals of the posterior distributions. Under D2 (the correctly-specified case for GPH), both models track the true elasticity comparably across brands overall. Under D3, DDPH closely follows the true elasticity throughout the training window, whereas GPH systematically fails to do so.

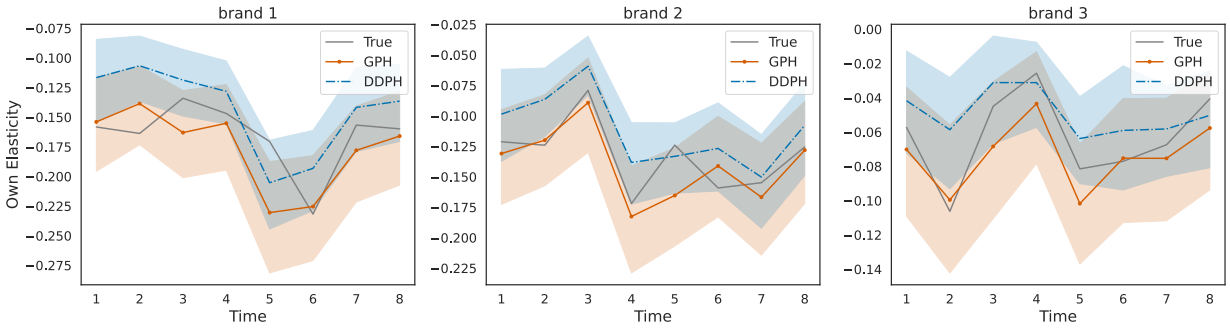

Figure 16: Own-attribute elasticity of demand under D2.

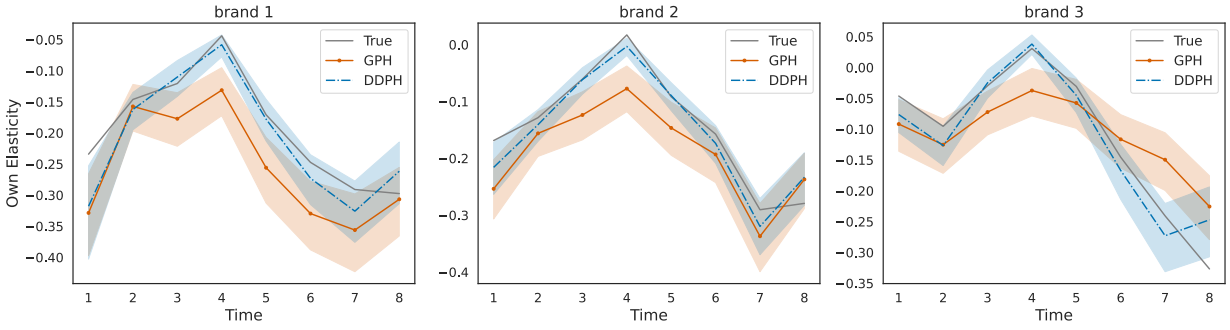

Figure 17: Own-attribute elasticity of demand under D3.

Sensitivity of the Results to the Choice of Hyperparameters

We conduct two hyperparameter sensitivity checks. In the first, we vary the model hyperparameters while holding the DGP hyperparameters fixed. In the second, we vary the DGP hyperparameters while holding the model hyperparameters fixed. We find that the simulation results are robust to both. The following details each hyperparameter sensitivity check.

We investigate the sensitivity of the simulation results to the choice of hyperparameters of the models: (μ_κ, μ_σ) and $(\omega_\kappa^2, \omega_\sigma^2)$ that govern the prior of the kernel parameters $\boldsymbol{\theta}^{(p)}, \boldsymbol{\theta}_0$, and (a_0, b_0) that govern the prior of the DP concentration parameter α. We use the data from D3 and modify one aspect of the DDPH model hyperparameters at a time relative to the default specification we used in the main text ($\mu_\kappa = 0$, $\mu_\sigma = 0$, $\omega_\kappa^2 = 1$, $\omega_\sigma^2 = 1$, $a_0 = 2$, $b_0 = 1$): (i) a shifted kernel prior mean $(\mu_\kappa, \mu_\sigma) = (1, -1)$, (ii) a widened kernel prior variance $(\omega_\kappa^2, \omega_\sigma^2) = (5, 5)$, and (iii) a tighter Gamma prior on the DP concentration $(a_0, b_0) = (2, 4)$. For each modification, we re-fit DDPH and compare the results from this model to those from the DDPH with the default hyperparameter specification. Figure 18 reports the posterior predictive distribution of $\beta_i^{(1)}(t)$ at $t = 1, 2, 3$ from different hyperparameter specifications of DDPH. We see from the figure that the posterior predictive distributions from DDPH track the true multimodal distributions across time periods accurately regardless of the choice of the model hyperparameters.

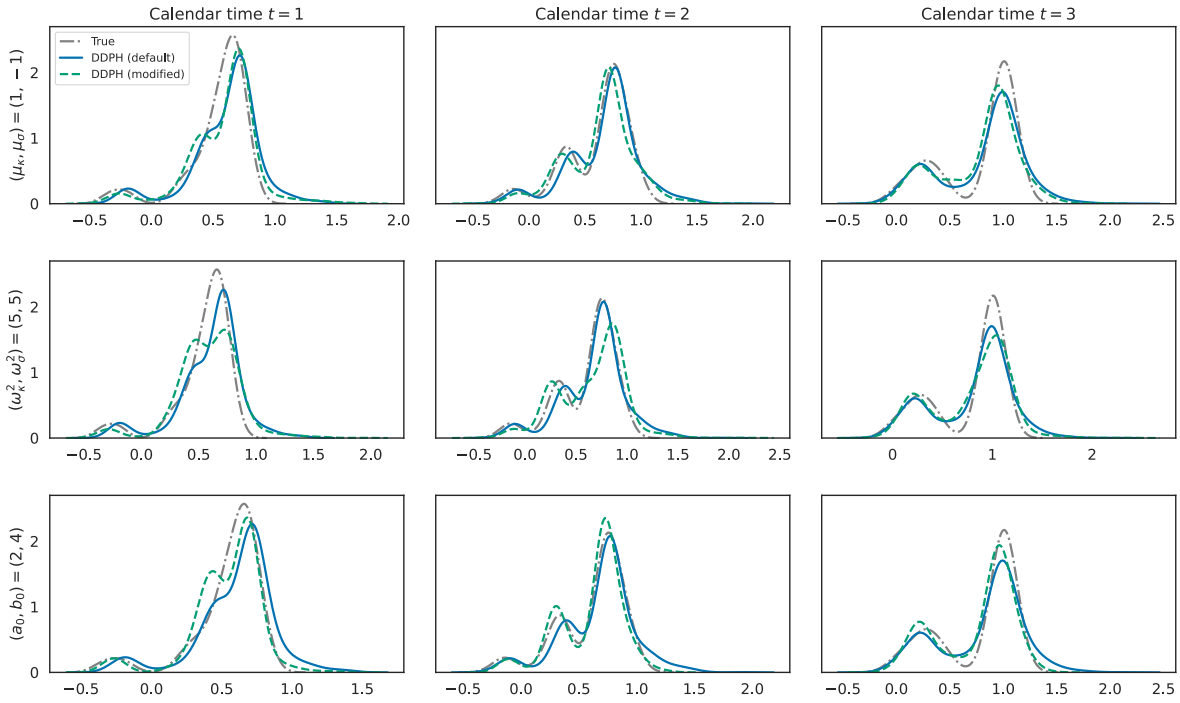

Figure 18: Posterior predictive distributions of the first coefficient $\beta_i^{(1)}(t)$ at time periods $t = 1, 2, 3$ (columns) under three model hyperparameter specifications of DDPH. Top row: $(\mu_\kappa, \mu_\sigma) = (1, -1)$, middle row: $(\omega_\kappa^2, \omega_\sigma^2) = (5, 5)$, and bottom row: $(a_0, b_0) = (2, 4)$. Each subplot overlays the true distribution (gray, dash-dotted) with the DDPH posterior predictive under the default hyperparameter values (blue, solid) and under the modified hyperparameter values (green, dashed).

We also assess the sensitivity of our results to the hyperparameters of the data-generating process. We used the following specification when we simulated the data in the simulation studies in the main text: $(\mu_\kappa, \mu_\sigma) = (0, -2)$, $(\omega_\kappa^2, \omega_\sigma^2) = (0.1, 0.1)$, and $(a_0, b_0) = (2, 1)$. Here, we modify one aspect of the DGP hyperparameters of D3 at a time, (i) a shifted prior mean of the kernel hyperparameters, $(\mu_\kappa, \mu_\sigma) = (5, -5)$, (ii) a widened prior variance of the kernel hyperparameters, $(\omega_\kappa^2, \omega_\sigma^2) = (2, 2)$, and (iii) a tighter Gamma prior on the DP concentration, $(a_0, b_0) = (2, 4)$. In each of the three cases, we re-fit DDPH and GPH on the dataset based on the modified DGP hyperparameters and plot the posterior predictive distributions of $\beta_i^{(1)}(t)$ at $t = 1, 2, 3$ overlaid on the true cross-sectional distributions in Figure 19.

The DDPH posterior predictive continues to track the true, potentially multimodal distributions closely, while the GPH benchmark continues to produce broad unimodal posterior predictive

distributions. The overall results are similar to those in Figure 3 presented in the main text, confirming that our model's recovery of the heterogeneity distributions is robust to changes in the DGP hyperparameter specifications.

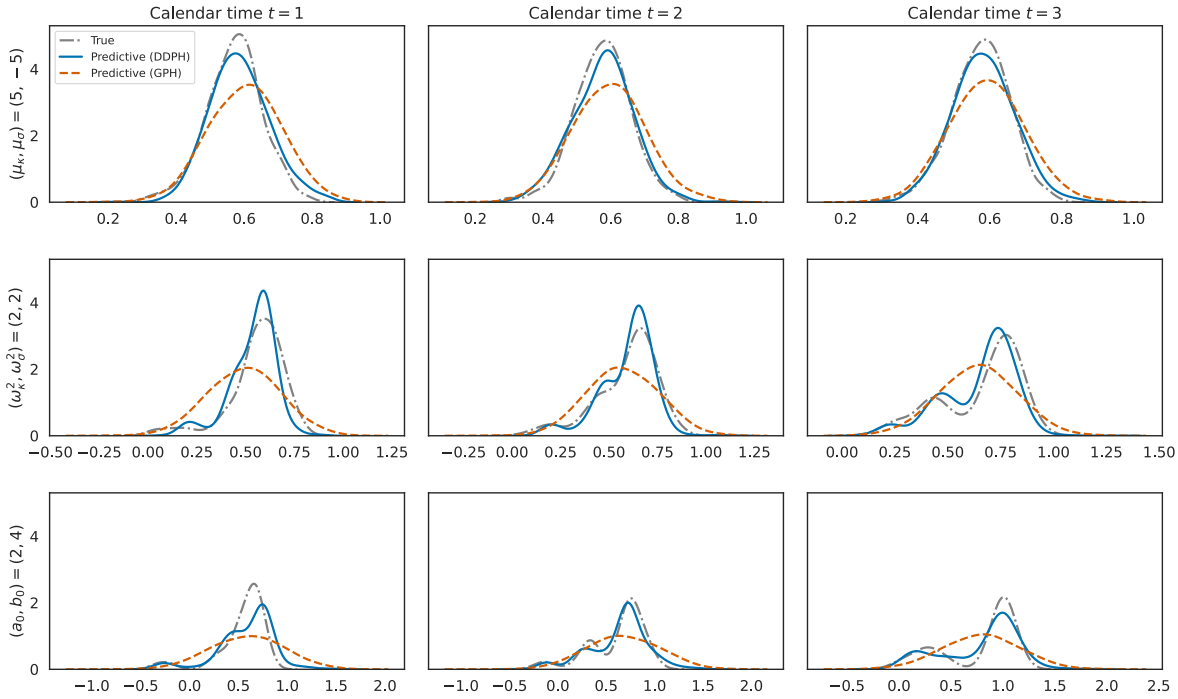

Figure 19: Posterior predictive distributions of the first coefficient $\beta_i^{(1)}(t)$ at time periods $t = 1, 2, 3$ (columns) under three DGP hyperparameter specifications. Top row: $(\mu_\kappa, \mu_\sigma) = (5, -5)$, middle row: $(\omega_\kappa^2, \omega_\sigma^2) = (2, 2)$, and bottom row: $(a_0, b_0) = (2, 4)$. Each subplot overlays the true distributions (gray, dash-dotted) with the DDPH (blue) and the GPH posterior predictive distributions (orange, dashed).

Convergence of the MCMC Chain

We provide the trace plots for the parameters across the MCMC iterations for one run of our model on D3 in simulation. Figure 20 shows these trace plots for selected parameters based on 2,000 thinned MCMC draws from an entire chain consisting of 20,000 iterations obtained by retaining every 10th iteration. The vertical dashed line in the plots marks the burn-in cutoff. We show these plots for the following quantities: the kernel length scale κ_p and variance σ_p^2 for each of the four Gaussian-process coefficients $p \in \{0, 1, 2, 3\}$ (top row); the DP concentration α and the active-cluster count $|\{z_i\}|_{i=1}^N$ (second row); posterior predictive draws of the first coefficient $\beta^{(1)}(t)$

at $t = 1$ and $t = 8$ (third row); and posterior predictive draws of the first intercept $\beta^{\text{icept}_1}(t)$ at the same two time points (fourth row).

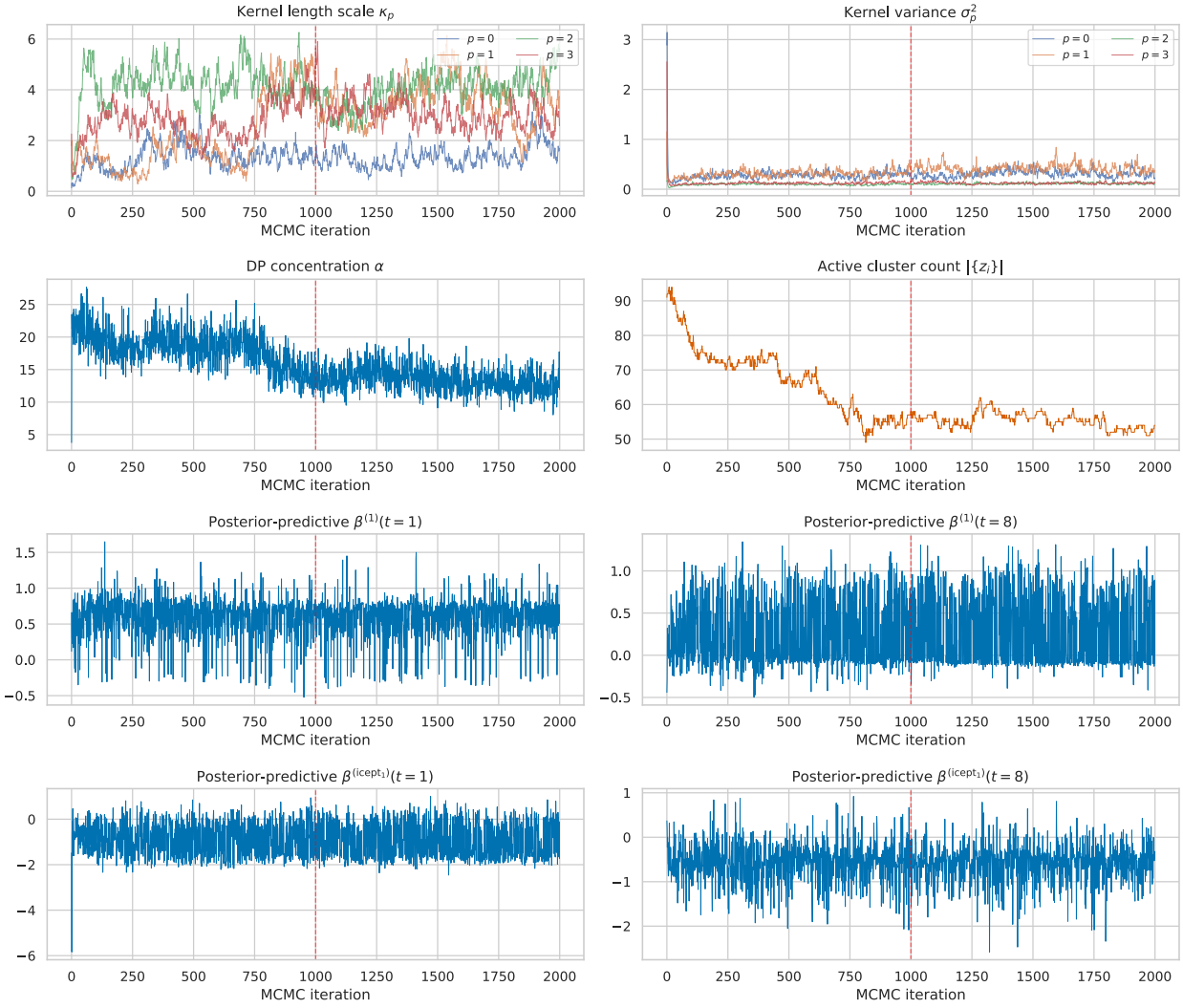

Figure 20: Trace plots from one run on D3. Top row: kernel length scale κ_p (left) and variance σ_p^2 (right) for $p = 0, 1, 2, 3$. Second row: DP concentration α (left) and active-cluster count $|\{z_i\}|$ (right). Third row: posterior predictive draws of $\beta^{(1)}(t)$ at $t = 1$ and $t = 8$. Fourth row: posterior predictive draws of $\beta^{\text{icept}_1}(t)$ at the same two time points. The chain is thinned by retaining every 10 iterations. The vertical dashed line marks the burn-in cutoff at iteration 10,000.

The chain mixes well: the kernel hyperparameters fluctuate within stable bands after burn-in, the DP concentration α and active-cluster count appear to settle into their stationary distributions before the burn-in cutoff, and the posterior predictive draws of $\beta^{(1)}(t)$ and $\beta^{\text{icept}_1}(t)$ show no

persistent trends. We use only the second half of the chain (post burn-in) for posterior inference.

WEB APPENDIX B: MORE ON EMPIRICAL APPLICATION

Data Splitting and Prediction Procedure

We construct three evaluation sets from the monthly scanner panel data along two dimensions: calendar time and consumers. We split the calendar time window into a training period and a holdout period: the last four months are held out for the future-period forecasting task ($T_{\text{test}} = 4$), while the preceding $T = 68$ months form the training period. We further split the data randomly on the consumer dimension. A training subset includes 80% of the consumers, and the holdout data consists of the remaining 20%. We denote N_{test} as the number of consumers in the holdout data. We use the three evaluation sets in the following three prediction tasks.

- **In-sample fit**: training period observations of training-set consumers ($T = 68$ months, $N =$ training consumers).
- **New consumer prediction**: training period observations of holdout consumers, who are excluded entirely when fitting the model.
- **Future purchase prediction**: holdout period observations of training-set consumers, projecting forward beyond the period over which the model was fit.

For the new-consumer task, we draw individual-level preference trajectories for each holdout consumer from the posterior predictive distribution of the trained model: under DDPH, this corresponds to drawing a cluster assignment from a discrete distribution based on the stick-breaking weights, while under GPH it corresponds to sampling from the consumer-level GP whose hyperparameters were learned at training time. For the future-purchase task, we similarly draw the training-set consumers' preference trajectories on the holdout calendar months from the posterior predictive of the GP, conditioned on their training period trajectories. Details on derivation appear in Appendix A.

Population Mean Evolution

We report the estimated population mean trajectories of individual-level parameters for the chips category in the main body of the manuscript. Here, we present these trajectories for the remaining product categories: detergent, toilet paper, peanut butter, soup, and coffee.

Across all five categories, the population mean of the parameters is generally estimated to evolve quite differently between the two models, as shown in Figures 21 to 25. For example, we see some qualitative differences, focusing on the recession period highlighted by the shaded region in the figures. For the detergent category, the GPH model significantly overestimates the brand 5 intercept and the feature coefficient. The two models suggest significantly different trajectories for the brand 4 intercept during the recession period in the peanut butter category, as well.

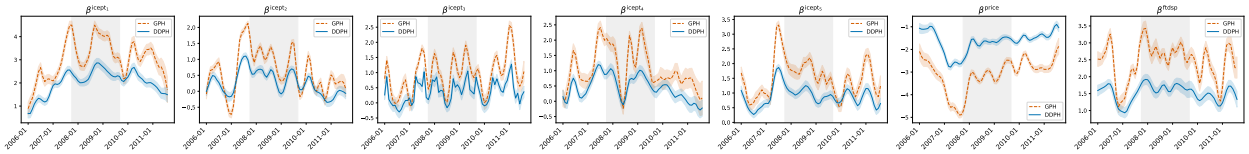

Figure 21: Population mean of the individual-level parameters over time for the detergent category.

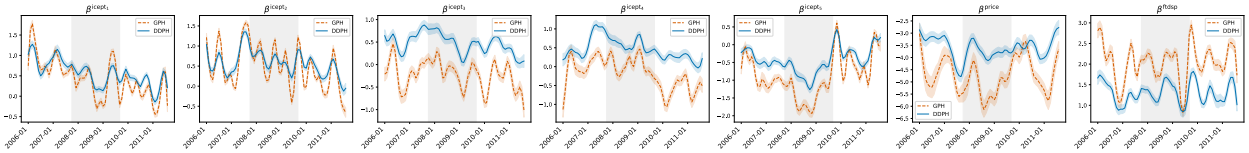

Figure 22: Population mean of the individual-level parameters over time for the toilet paper category.

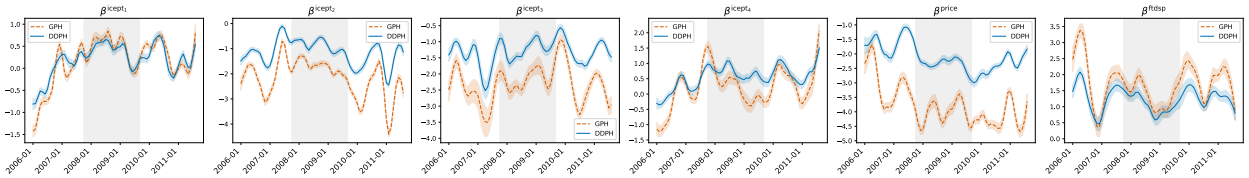

Figure 23: Population mean of the individual-level parameters over time for the peanut butter category.

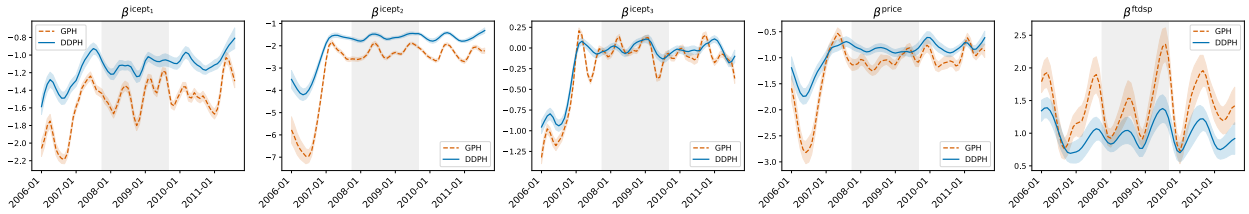

Figure 24: Population mean of the individual-level parameters over time for the soup category.

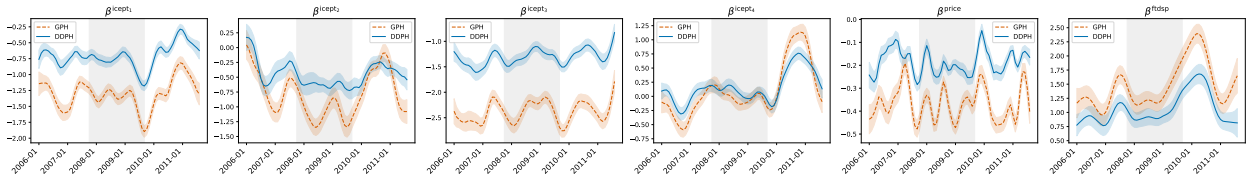

Figure 25: Population mean of the individual-level parameters over time for the coffee category.

Posterior Predictive Distribution of the Price Coefficient

Here, we show a visual summary of the posterior predictive distributions of the price coefficient between 2007 Q2 and 2009 Q2 for the categories of detergent, toilet paper, peanut butter, soup, and coffee. Figures 26 to 30 show the posterior predictive distributions for these categories. The DDPH model reveals complex, multimodal natures of the price coefficient distributions, suggesting the existence of multiple consumer segments with varying price sensitivities. Our model also captures meaningful shifts in these distributions over time during the Great Recession. For example, in the peanut butter category, the distributions captured by the DDPH model show a clear leftward shift as the recession deepens (i.e., from 2007 Q4 to 2008 Q2). This movement towards more negative price coefficients is consistent with the intuition that consumers become more price-sensitive during periods of economic distress.

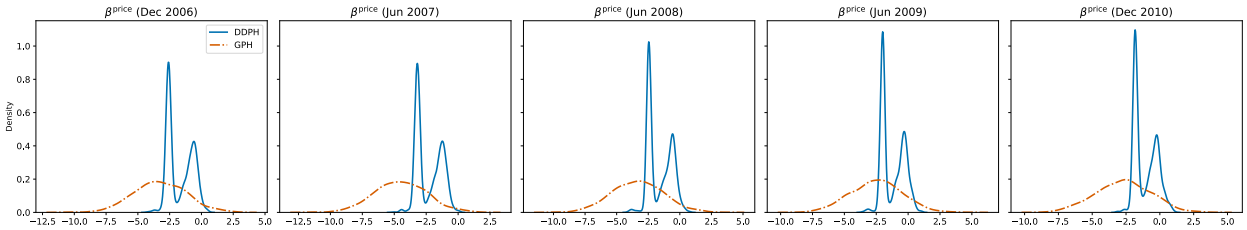

Figure 26: Posterior predictive distribution of the price coefficient for the detergent category.

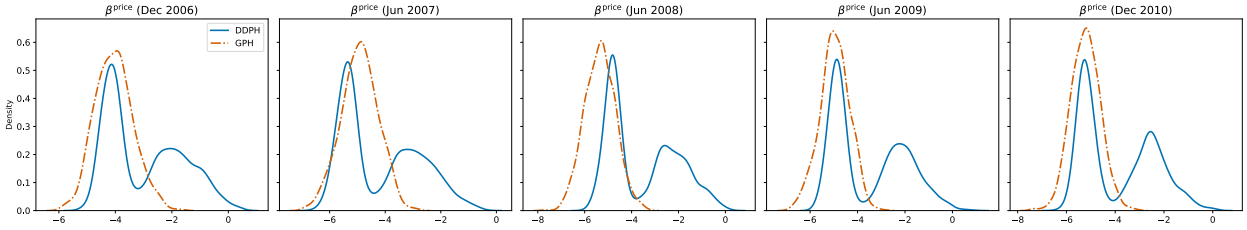

Figure 27: Posterior predictive distribution of the price coefficient for the toilet paper category.

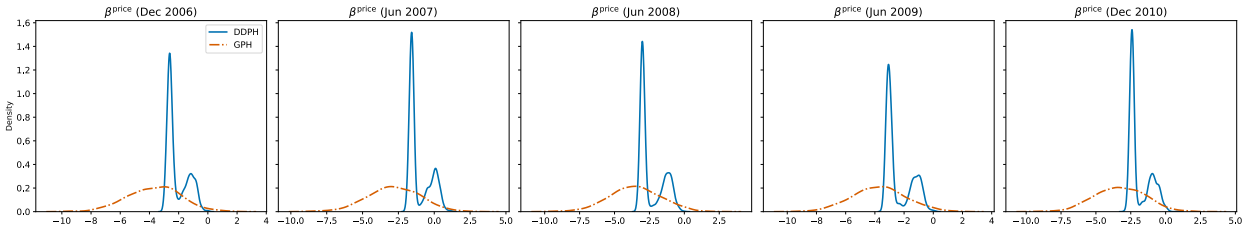

Figure 28: Posterior predictive distribution of the price coefficient for the peanut butter category.

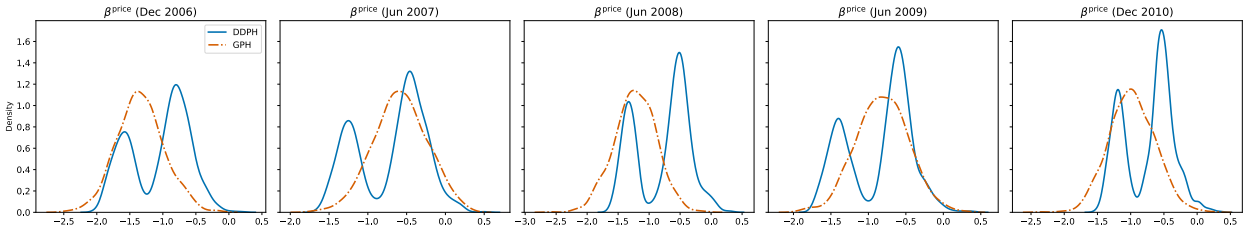

Figure 29: Posterior predictive distribution of the price coefficient for the soup category.

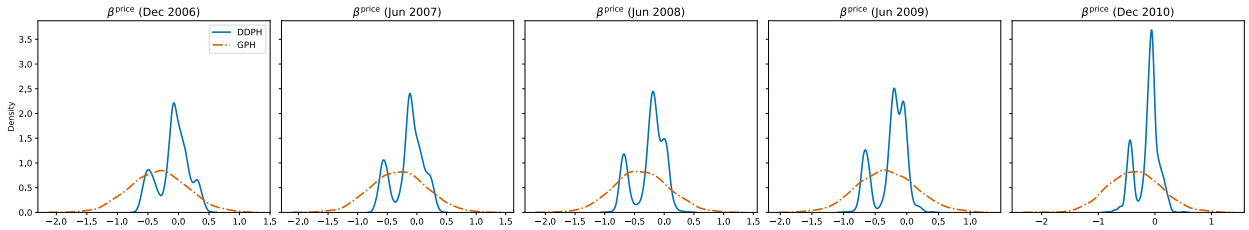

Figure 30: Posterior predictive distribution of the price coefficient for the coffee category.

Predictive Performance: Macro-averaged Recall and Specificity

We supplement the overall prediction accuracy reported in the Model Comparison and Fit subsection with macro-averaged recall (sensitivity) and macro-averaged specificity. We use the same definitions of these metrics as in Web Appendix A. The same three prediction tasks as in Table 6 (in-sample fit, new-consumer prediction, and future-purchase prediction) are evaluated. Tables 17 and 18 report these metrics across the six categories. The performance of DDPH relative to GPH is preserved under both metrics for nearly every (category, task) cell, confirming that the results from the accuracy metric are robust to the choice of a classification metric.

Table 17: Macro-averaged recall by category and prediction task.

	In-sample fit		New consumer prediction		Future purchase prediction	
Category	DDPH	GPH	DDPH	GPH	DDPH	GPH
Detergent	0.696	**0.705**	**0.436**	0.431	**0.480**	0.469
Chips	**0.541**	0.539	**0.372**	0.367	**0.492**	0.482
Toilet paper	**0.681**	0.656	**0.354**	0.340	**0.463**	0.387
Peanut butter	0.713	**0.716**	**0.483**	0.462	**0.381**	0.318
Coffee	**0.648**	0.642	**0.265**	0.260	**0.500**	0.486
Soup	**0.443**	0.429	**0.307**	0.302	0.397	**0.401**
Average	**0.620**	0.614	**0.370**	0.360	**0.452**	0.424

Higher value in each (category, task) pair is in bold.

Table 18: Macro-averaged specificity by category and prediction task.

	In-sample fit		New consumer prediction		Future purchase prediction	
Category	DDPH	GPH	DDPH	GPH	DDPH	GPH
Detergent	0.944	**0.945**	**0.889**	0.888	**0.910**	0.909
Chips	**0.854**	0.848	**0.795**	0.794	**0.850**	0.829
Toilet paper	**0.937**	0.932	**0.876**	0.873	**0.890**	0.874
Peanut butter	**0.929**	0.928	**0.872**	0.866	**0.852**	0.829
Coffee	**0.914**	0.910	**0.818**	0.818	**0.873**	0.867
Soup	**0.829**	0.813	**0.775**	0.773	**0.804**	0.795
Average	**0.901**	0.896	**0.838**	0.835	**0.863**	0.851

Higher value in each (category, task) pair is in bold.

Posterior Predictive Checks for Shares

We assessed the adequacy of both models via posterior predictive checks of the market shares of the different brands for the entire duration of the data and for each time period. We simulated replicated consumer choices from the posterior predictive distribution under each model for each retained MCMC draw r. From these replicates, we computed the market share of each brand c, defined as the proportion of observations in which brand c was chosen in the replicated datasets. The R retained draws yield a posterior predictive distribution for the share of each brand, which we compared against the observed share in the data. We also report the posterior predictive p-value (PPP), the fraction of replicated shares that are at least as large as the observed share. PPP values near 0.5 indicate that the observed share sits in the bulk of the predictive distribution, which is desirable. Values close to 0 or 1 indicate that the model systematically over- or under-predicts the share. Figures 31 to 36 show the replicated share histograms for the GPH and DDPH models with the observed share (dashed vertical line) for each of the six categories. Both models track the observed shares closely for most brands, with non-extreme PPP values.

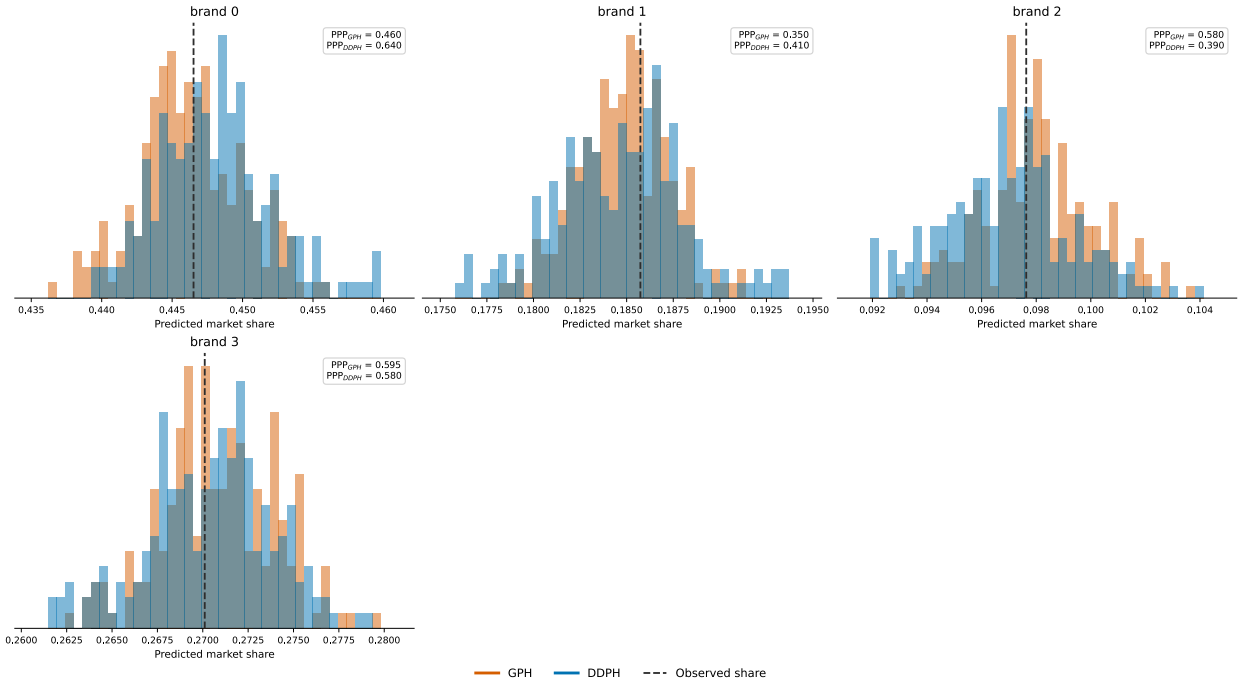

Figure 31: Posterior predictive check of market share for the chips category.

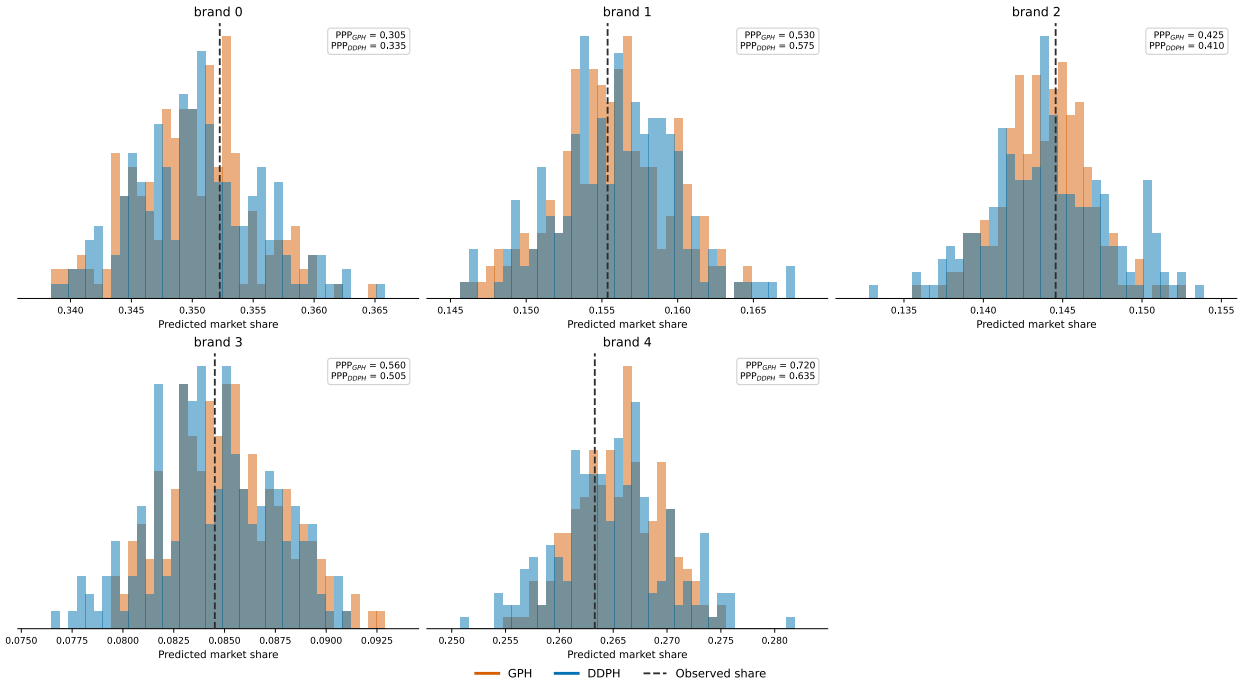

Figure 32: Posterior predictive check of market share for the coffee category.

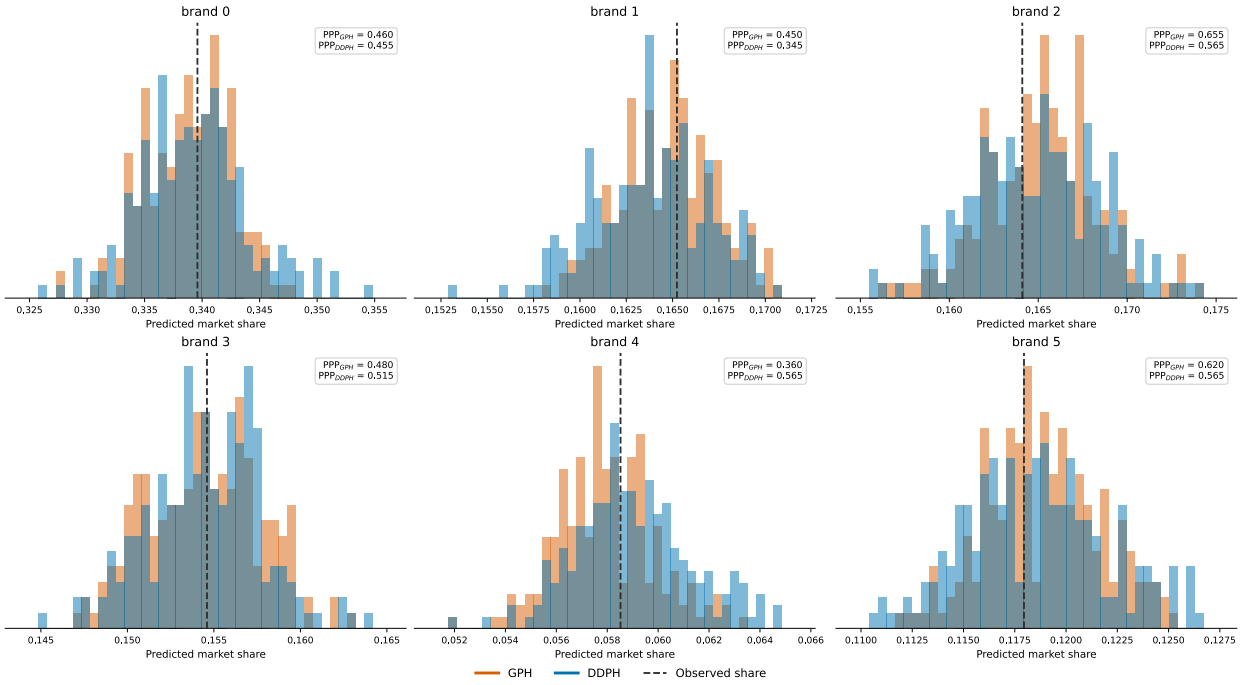

Figure 33: Posterior predictive check of market share for the detergent category.

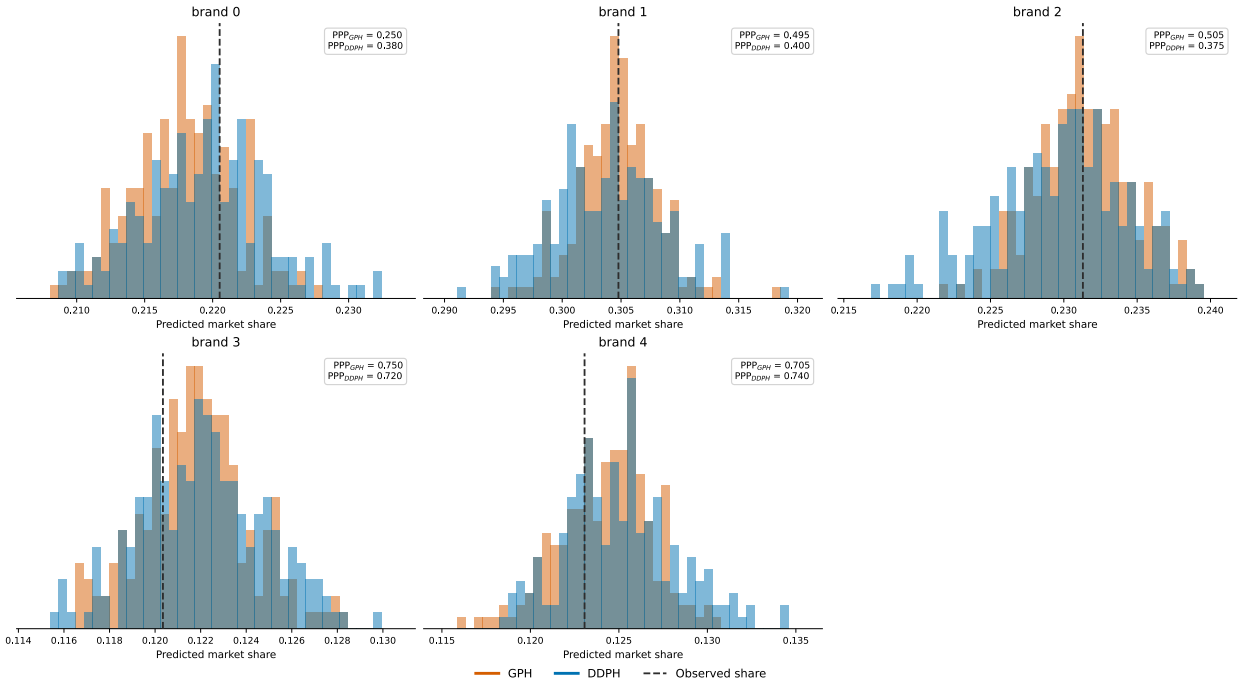

Figure 34: Posterior predictive check of market share for the peanut butter category.

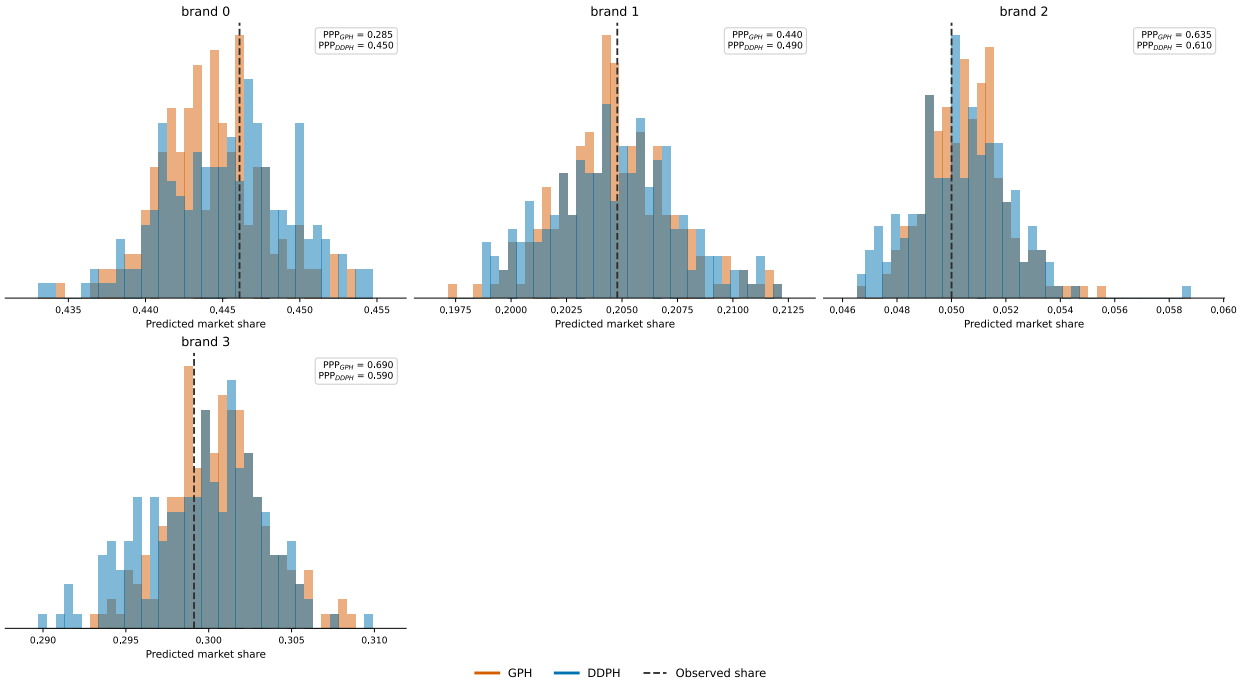

Figure 35: Posterior predictive check of market share for the soup category.

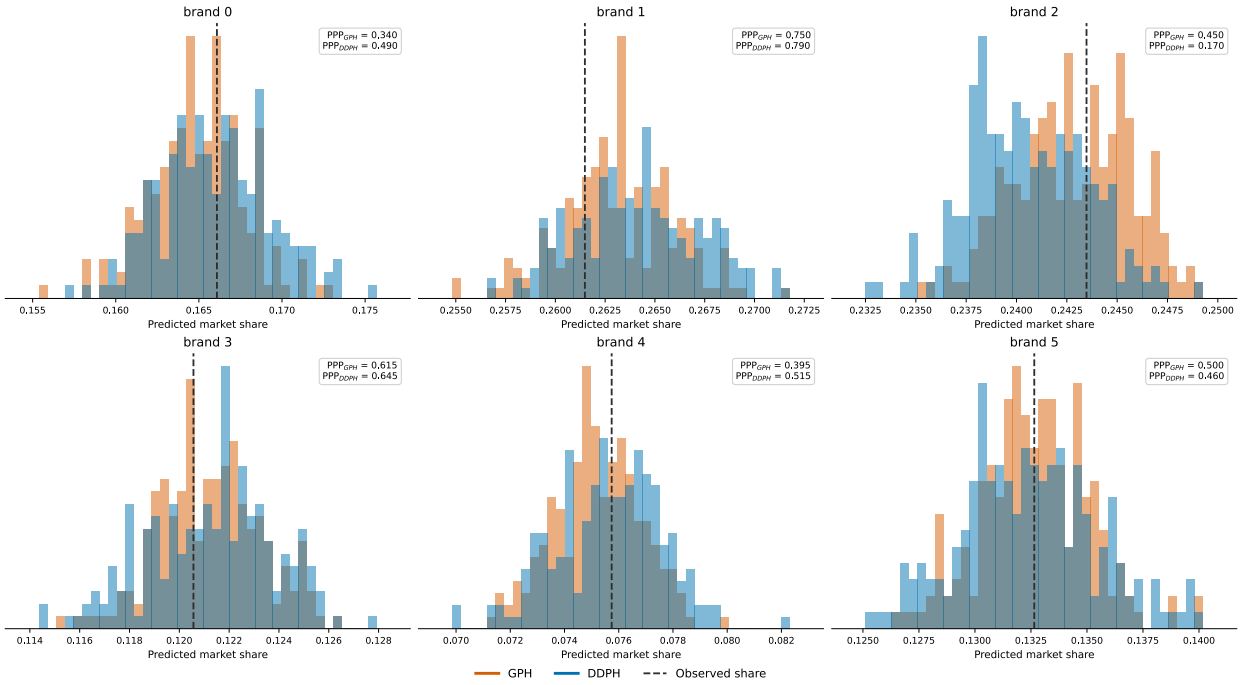

Figure 36: Posterior predictive check of market share for the toilet paper category.

Posterior Predictive Checks for Shares Over Time

We extended the aggregate share check by computing the monthly market share for each brand from the replicated and observed choices. The resulting per-period replicated shares trace a time-varying trajectory for each MCMC draw. We summarize the posterior predictive distributions of these trajectories by their median and a 95% credible interval band and compare them to the observed per-period share. The credible intervals from a model capturing the temporal dynamics correctly should tend to include the observed share trajectory. Figures 37 to 42 display these temporal posterior predictive checks for both models across the six categories.

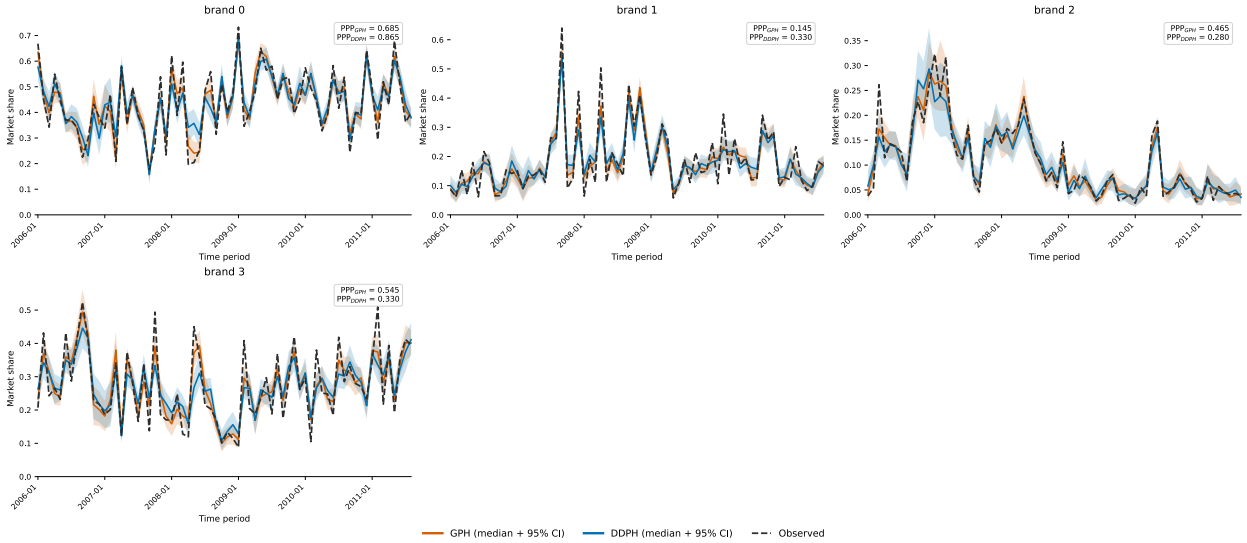

Figure 37: Posterior predictive check of temporal market share for the chips category.

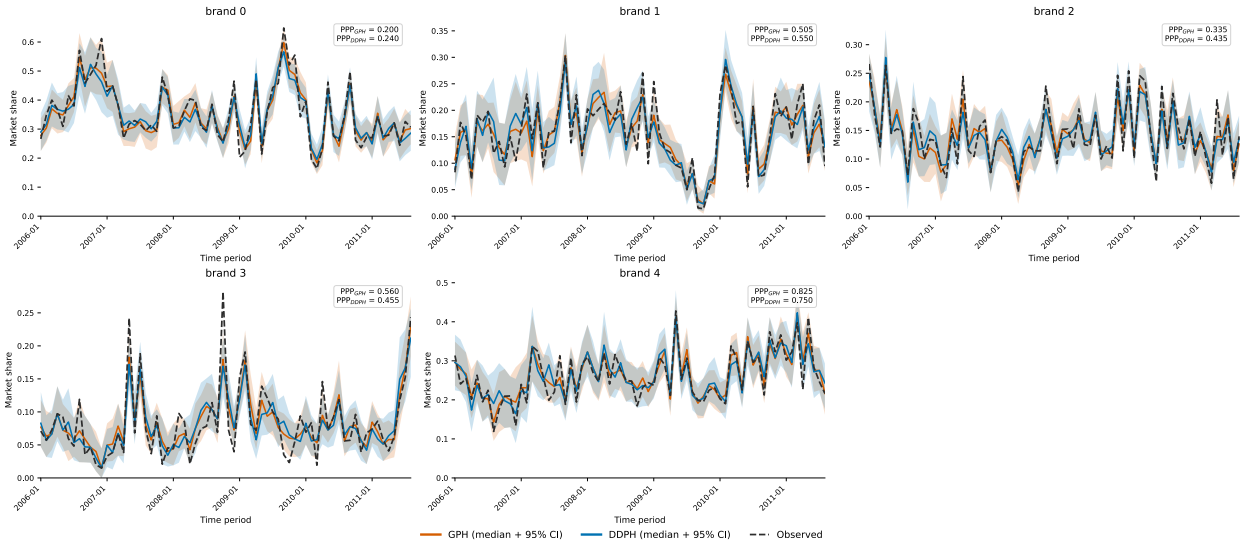

Figure 38: Posterior predictive check of temporal market share for the coffee category.

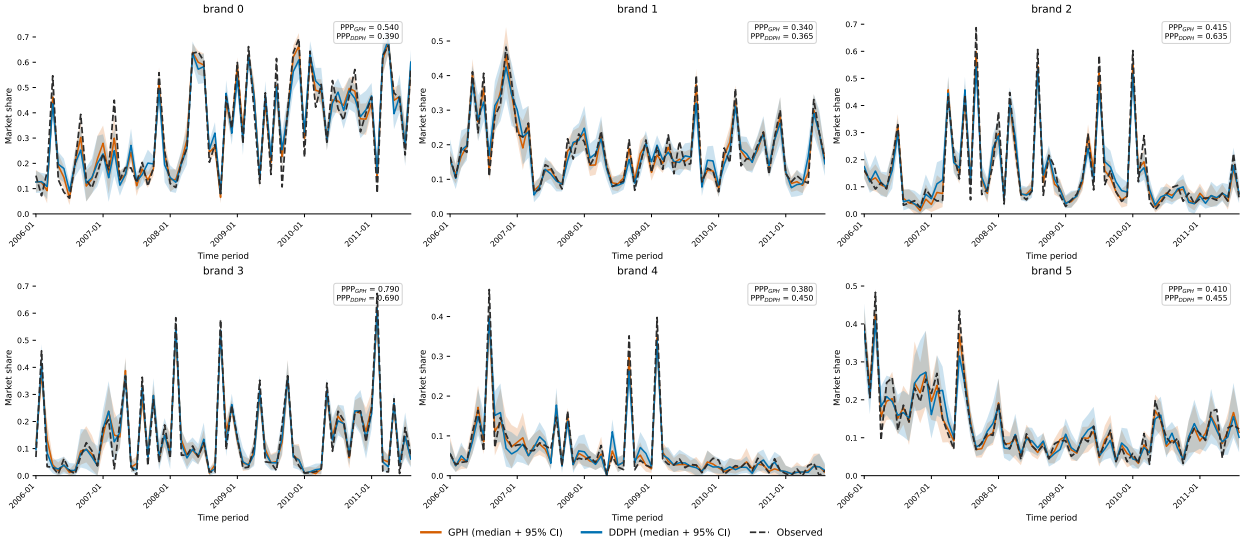

Figure 39: Posterior predictive check of temporal market share for the detergent category.

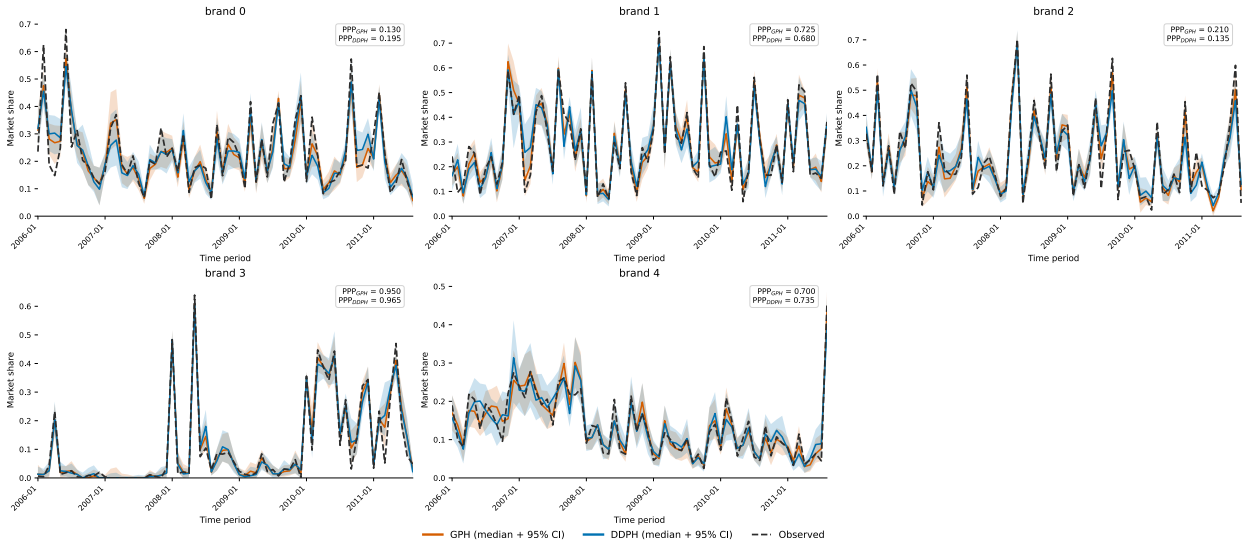

Figure 40: Posterior predictive check of temporal market share for the peanut butter category.

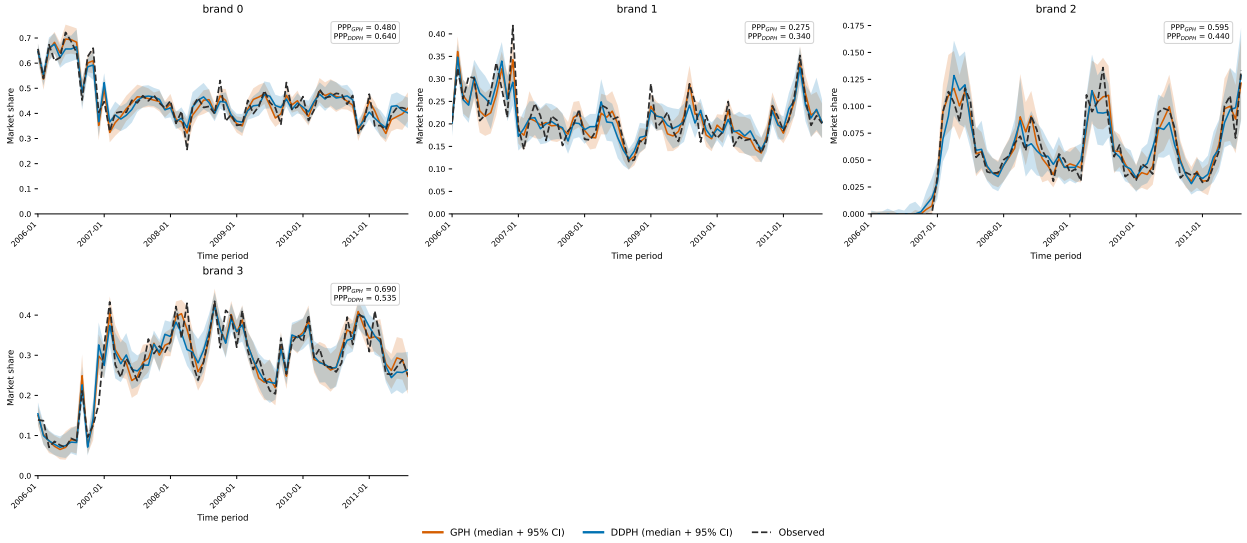

Figure 41: Posterior predictive check of temporal market share for the soup category.

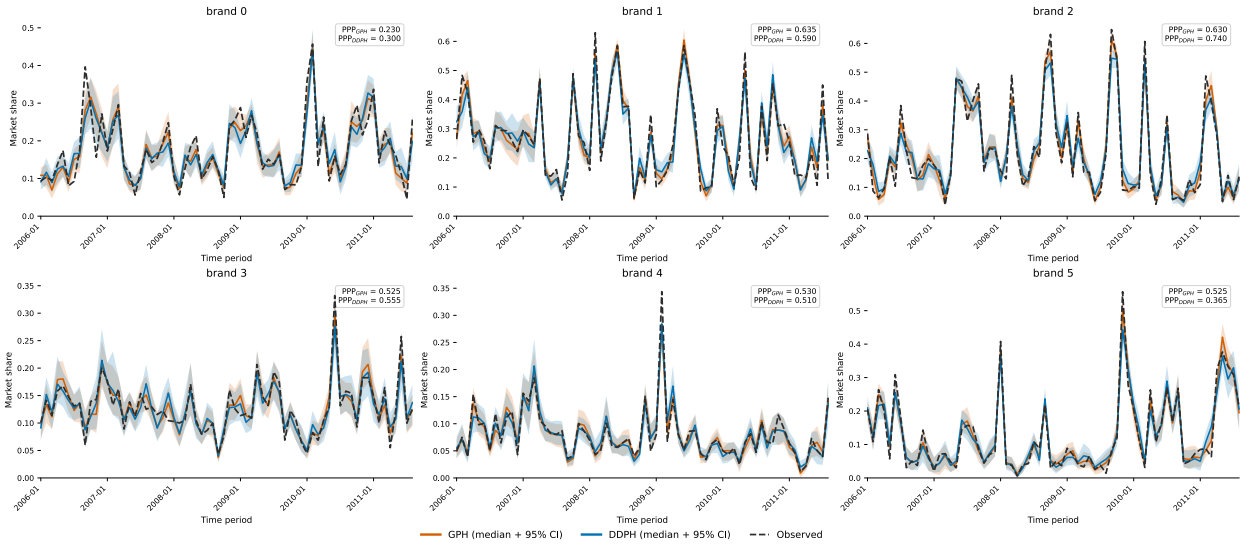

Figure 42: Posterior predictive check of temporal market share for the toilet paper category.