Modeling Dynamic Heterogeneity using Gaussian Processes

Ryan Dew*
The Wharton School
University of Pennsylvania

Asim Ansari
Columbia Business School
Columbia University

Yang Li
Cheung Kong Graduate
School of Business

October 7, 2019

Marketing research relies on individual-level estimates to understand the rich heterogeneity that exists in consumers, firms, and products. While much of the literature focuses on capturing static cross-sectional heterogeneity, little research has been done on modeling dynamic heterogeneity, or the heterogeneous evolution of individual-level model parameters. In this work, the authors propose a novel framework for capturing the dynamics of heterogeneity, using individual-level, latent, Bayesian nonparametric Gaussian processes. Similar to standard heterogeneity specifications, this Gaussian Process Dynamic Heterogeneity (GPDH) specification models individual-level parameters as flexible variations around population-level trends, allowing for sharing of statistical information both across individuals and within individuals over time. This hierarchical structure provides precise individual-level insights regarding parameter dynamics. The authors show that GPDH nests existing heterogeneity specifications, and that not flexibly capturing individual-level dynamics may result in biased parameter estimates. Substantively, they apply GPDH to understanding preference dynamics, and modeling the evolution of online reviews. Across both applications, they find robust evidence of dynamic heterogeneity, and illustrate GPDH's rich managerial insights, with implications for targeting, pricing, and market structure analysis.

Keywords: dynamics, heterogeneity, Bayesian nonparametrics, Gaussian processes, choice models, topic models, machine learning

*Ryan Dew is the lead and corresponding author. This paper is based on the second essay of his dissertation. The other authors are listed alphabetically. Ryan Dew (ryandew@wharton.upenn.edu) is an Assistant Professor of Marketing at the Wharton School of the University of Pennsylvania. Asim Ansari (maa48@gsb.columbia.edu) is the William T. Dillard Professor of Marketing at Columbia Business School. Yang Li (yangli@ckgsb.edu.cn) is an Associate Professor of Marketing at Cheung Kong Graduate School of Business in Beijing.

1 Introduction

The modeling of dynamic phenomena is central to marketing research. Marketers are interested in understanding the evolution of consumer perceptions, preferences, and their response sensitivities, as well as the success and failure of different brands over time. Marketing decisions that focus on temporal consequences of marketing actions necessarily rely on empirical models of such marketing dynamics (Xie et al., 1997; Pauwels and Hanssens, 2007; Naik, 2015). Often, these dynamics are heterogeneous across individual units. We use the term "individuals" to broadly refer to units over which the heterogeneity is defined, examples of which include consumers, brands, and products. For example, the pattern of evolution of preferences could vary across customers because of how they are differentially affected by economic conditions such as recessions. Similarly, how market perceptions evolve could vary across brands because of competitive activity. In such situations, the interest is in both the market-level evolution of preferences, as well as in the individual-level trajectories that may differ from each other and from how the market is evolving on average.

In this paper, we develop a modeling framework for representing such dynamic heterogeneity. Dynamic heterogeneity characterizes situations where individual-level model parameters evolve over time according to a stochastic process. More specifically, we allow individual level parameters to evolve flexibly in a fashion that does not force them to exactly mimic the dynamic evolution of the population mean. We do this by allowing the individual deviations from the population means to vary over time. At a given point in time, the collection of individual-level parameters forms a distribution of cross-sectional heterogeneity. The evolution of these individual-level parameters therefore results in a time-varying population distribution, in which the relative positions of individuals are changing over time. We illustrate the concept of dynamic heterogeneity in the top part of Figure 1.

While marketing researchers have modeled many different forms of heterogeneity (DeSarbo et al., 1997), most of the literature focuses on the variation in preferences across individuals. Variation within individuals over time has been relatively understudied. Modeling this within-individual variation has important managerial implications for understanding changes in markets over time, and for developing dynamic segmentation and targeting strategies. In addition, just as ignoring cross-sectional heterogeneity can induce estimation bias, not accounting for parameter evolution can also distort inferences about elasticities or response sensitivities and misinform managerial actions.

A number of marketing studies have used models that include time varying individual parameters. Examples include Kim et al. (2005), Liechty et al. (2005), Sriram et al. (2006), Lachaab et al. (2006), and Guhl et al. (2018). These studies have used a number of different specifications to capture the evolution of parameters. Kim et al. (2005), for example, use a vector autoregressive model to represent the evolution of the population mean, while Sriram et al. (2006) employ a dynamic linear model, and Guhl et al. (2018) rely on penalized splines. Crucially, while all these papers use a dynamic model to capture how parameters evolve on average, each imposes a *static* heterogeneity assumption: conditional on a time-varying mean model $\mu(t)$, the individual-level parameter at time t for individual i is given

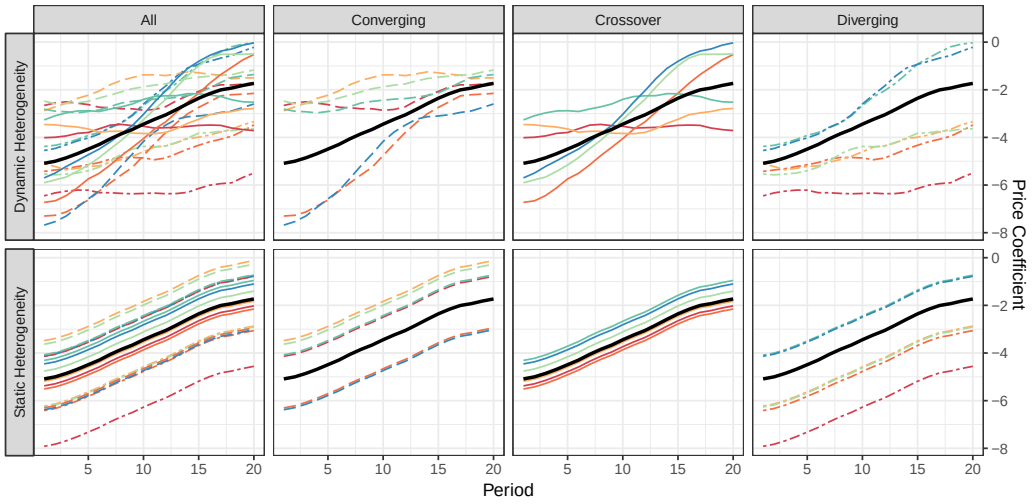

Figure 1: A synthetic example of dynamic heterogeneity, contrasting, in the top row, price sensitivities of individuals evolving dynamically relative to the population mean (black), and in the bottom row, the same individuals modeled under the fixed-offsets assumptions. In the leftmost panel, we see all of the individuals plotted together, while in the remaining panels, we illustrate some interesting dynamic heterogeneity patterns. Comparing the top and bottom sets, we see that the fixed-offsets assumption ignores the interesting individual-level dynamics that may be useful for making individual-level targeting decisions, and for characterizing changes in preferences over time.

by $\beta_{it} = \mu(t) + \theta_i$, where θ_i is a *time-invariant* offset for the individual. Such a *fixed-offsets* specification is restrictive: while it allows the population mean to change according to a model, the individual-level parameters are forced to maintain a fixed distance θ_i from that mean at any instant. Even though this results in time-varying individual-level parameters, their dynamic patterns exactly mimic the overall population dynamics, resulting in static heterogeneity. This fails to reflect the full richness of individual-level dynamics, as illustrated through the bottom part of Figure 1.

We propose a new methodological framework for modeling *dynamic* heterogeneity in hierarchical models. Drawing on the literature on Bayesian nonparametric models in statistics and machine learning (Rasmussen and Williams, 2005), we develop a novel Gaussian Process Dynamic Heterogeneity (GPDH) specification that characterizes heterogeneity over time-varying latent variables using individual-level random functions of time. These functions are estimated using Gaussian processes (GPs) that are centered around a common mean model. This model captures population-level dynamics, and is itself inferred from the data. The GPDH specification is a dynamic analogue to static random coefficient specifications, where the mean model plays the role of the population trend, and the individual-level functions capture time-varying heterogeneity around this trend. Similar to traditional heterogeneity specifications, our proposed dynamic heterogeneity specification allows for: (1) the sharing of statistical information across individuals by shrinking their trajectories toward a common mean trajectory, (2) the sharing of statistical information within individuals across time periods (i.e., intra-individual smoothing), (3) flexible inter-temporal evolution,

and (4) a principled probabilistic mechanism for projecting the evolution of individual and mean trajectories into other time periods. An important feature of our GPDH specification is that it can be used with *any* mean model, allowing the researcher to incorporate prior expectations, theory, or even specific drivers of dynamics. Our use of GPs to nonparametrically represent individual-level latent functions that are shrunk towards a common dynamic population model is novel to the econometric, marketing, and machine learning literatures.

Capturing dynamic heterogeneity yields many benefits. It can generate interesting insights about the different patterns of individual-level evolution. For example, as we show in our two applications, identifying individuals whose parameters shifted from one extreme of the population to the other, or who moved from being in the extremes of the distribution to the center, or vice versa, can enhance managerial and substantive understanding, and can be leveraged by managers for targeted marketing. However, the importance of capturing dynamic heterogeneity goes beyond such individual-level insights. Statistically, if dynamic heterogeneity is present but static heterogeneity is assumed, as is commonly done, we could obtain misleading estimates about both the *population-level* mean and the extent of heterogeneity, both of which can negatively impact targeting decisions. This can be true even if the correct functional form for the population mean model is used, as we illustrate through simulations.

In this paper, we present two applications of GP dynamic heterogeneity: the first and most extensive is in a choice modeling context, similar to our motivating examples, where GPDH is used to represent time-varying consumer preferences for consumer packaged goods, over a span of time that includes the Great Recession. Using both simulated and real purchasing data, we show that GPDH yields more accurate and statistically efficient population and individual-level estimates of preference evolution. On data from six CPG categories, GPDH outperforms static heterogeneity specifications in both fit and forecasting tasks, across a wide array of performance metrics. At the same time, GPDH also uncovers individual-level patterns that can be used to characterize and target customers, and to study individual-level responses to economic shocks. More specifically, we find both simulated and empirical evidence of an attenuation bias in estimating population-level parameters when assuming static instead of dynamic heterogeneity around a dynamic mean model. Moreover, we find that, across all categories, estimated individual-level elasticities are notably higher when estimated with dynamic versus static heterogeneity. These biases, and the ability to predict individual-level dynamics, can directly impact category manager decision-making. Finally, the individual-level dynamics uncover cross-category differences in response to the recession: while there are obvious aggregate-level changes in price sensitivity in many categories, GPDH also uncovers category-level differences in the degree of *individual-level* response to the Great Recession.

Apart from the modeling of preferences, our specification can be adapted to a number of different settings. In our second application, we focus on an entirely different substantive context: the modeling of product reviews. We develop a novel, GPDH-based dynamic topic model to summarize relevant topics that are discussed in customer reviews for different brands of tablet computers. Particularly, our GPDH topic model captures how the review

contents of individual products evolve relative to aggregate patterns. Empirically, we show how these product-level topic trajectories give insights about the dynamics of market structure in the tablet computer market. Such a granular set of results is not obtainable via aggregate models of market dynamics.

The rest of the paper is structured as follows: We first give an overview of the needed methodological background, before introducing our GPDH framework. We then discuss our two applications: choice modeling and topic modeling. We conclude by highlighting other GPDH applications, describing limitations of the current paper, and suggesting future research directions.

2 Methodological Background

2.1 Literature

Gaussian processes (GPs) are Bayesian nonparametric models that are popular in statistics and computer science (O'Hagan and Kingman, 1978; Williams and Barber, 1998; Rasmussen and Williams, 2005) for flexibly modeling temporal and spatial phenomena. Marketing researchers have used Bayesian nonparametrics to represent the heterogeneity in static model parameters via Dirichlet process priors (Ansari and Mela, 2003; Kim et al., 2004; Braun et al., 2006; Ansari and Iyengar, 2006; Braun and Bonfrer, 2011; Li and Ansari, 2014). While Dirichlet processes are most commonly used to model uncertainty over probability distributions, Gaussian processes are most commonly used to model uncertainty over spaces of continuous functions.[1]

GPs have been used in marketing applications by Dew and Ansari (2018) to decompose variation in purchase rates in a dynamic customer base analysis setting. In their application, GPs were used to represent a mean model of spending rates, but individual-level variation around that mean model was still assumed to be static. Gaussian processes are also related to kriging methods used in Bronnenberg and Sismeiro (2002) for predicting demand across markets. In the context of choice models, Girolami and Rogers (2006) use GPs to model the utility functions of multinomial probit models in a non-dynamic and non-heterogeneous context. Finally, our work is also closely related to the marketing literature that models individual-level dynamics (Liechty et al., 2005; Sriram et al., 2006; Lachaab et al., 2006; Ansari and Iyengar, 2006; Khan et al., 2009; Guhl et al., 2018) via fixed-offsets specifications. Our purpose in this paper is to show how GPDH offers a more flexible and precise alternative than those restricted forms of heterogeneity. We now briefly describe GPs.[2]

[1] In some applications, these distinctions are fuzzy: for instance, GPs can also be used to model the density function of a distribution, as in Adams et al. (2009), while DPs can also be used to specify mixing distributions over functions, leading to flexible function approximations, as in Kottas (2006).

[2] We refer the reader to Rasmussen and Williams (2005) for a comprehensive treatment, and to Dew and Ansari (2018) for an extensive overview in a marketing context.

2.2 Gaussian Processes

A Gaussian process is a stochastic process $f()$ defined over some input space, which in the present work, we take to be time, $t \in \mathbb{R}^+$. GPs are defined by a mean function, $m(t)$, and a covariance function or kernel, $k(t, t')$ over input pairs (t, t') such that $m(t) = E[f(t)]$, and $k(t, t') = \text{Cov}(f(t), f(t'))$. If $f \sim \mathcal{GP}(m(t), k(t, t'))$, then for any finite set of inputs, $\boldsymbol{t} = (t_1, \ldots, t_T)$, the collection of corresponding function values (outputs) over these inputs has a joint multivariate Gaussian distribution, i.e.,

$$f(\boldsymbol{t}) = (f(t_1), \ldots, f(t_T)) \sim \mathcal{N}(m(\boldsymbol{t}), K(\boldsymbol{t}))\,, \tag{1}$$

where $m(\boldsymbol{t}) = (m(t_1), \ldots, m(t_T))$ is the mean vector of the multivariate normal and K is the $T \times T$ covariance matrix with entries given by $K_{ij} = k(t_i, t_j)$. In this way, GPs specify a Gaussian distribution over outputs for any given set of inputs, and therefore provide a natural mechanism for specifying uncertainty over a function space.

The mean and the kernel determine the nature of the functions that a GP prior generates. Informally, the mean function encodes the expected location of the functions, whereas the kernel encodes function properties, such as smoothness, amplitude, and differentiability. Much of the GP literature assumes a constant mean function to reflect a lack of prior knowledge about the shapes of the unknown functions, and the kernel serves as the main source of model specification. Many different kernels have been proposed in the GP literature. In theory, the kernel can be any function $k : \mathbb{R}^2 \to \mathbb{R}$ such that $K(\boldsymbol{t})$ remains positive semidefinite. Kernels are specified via hyperparameters that control the traits of the functions that a GP prior generates. Estimation yields hyperparameters as well as function values corresponding to particular inputs. These are used to predict the function values for a new set of inputs, based on the conditional distribution of the multivariate normal. In practice, stationary kernels, such as the Matérn kernel that we use in this paper, are most commonly used. We describe this kernel's properties in more detail in the next section.

3 Gaussian Process Dynamic Heterogeneity

Given the above background we now introduce our Gaussian Process Dynamic Heterogeneity (GPDH) specification in the context of a general hierarchical non-linear modeling framework specified in multiple stages. The first stage models the individual-level data in terms of individual-specific latent functions of time, the second stage specifies how these latent functions vary across individuals according to a GP that is characterized by a mean-model and a covariance kernel, and the third stage specifies priors over any invariant parameters in the individual-level model and the hyperparameters for the mean model and the heterogeneity specification.

3.1 Stage 1: Individual-Level Model

Suppose that the data y_{it} for individual i, at time t, observation m, comes from

$$y_{itm} \sim \pi(\boldsymbol{\beta}_{it}, \boldsymbol{\theta}),$$

where the individual-level parameters $\boldsymbol{\beta}_{it} = (\beta_{i1t}, \beta_{i2t}, \ldots, \beta_{iPt})$ are time-varying, and the parameters in $\boldsymbol{\theta}$ are invariant, both across individuals and over time. The exact functional form and the distributional assumptions used in specifying $\pi(\cdot)$ can differ across applications. For example, $\pi(\cdot)$ is a multinomial logit in our Application I, in which case $\boldsymbol{\beta}_{it}$ contains brand intercepts and response coefficients. In Application II, $\pi(\cdot)$ represents a topic model, where $\boldsymbol{\beta}_{it}$ captures online chatter about a particular topic for a brand i at time t.

3.2 Stage 2: Heterogeneity Specification

The key conceptual innovation of our framework is considering the time-varying individual-level parameters, β_{ipt}, as *functions* of time, $\beta_{ip}(t)$. We can then use GPs to specify a distribution over the space of individual-level functions, such that for each dynamic parameter $p = 1, \ldots, P$,

$$\beta_{ip}(t) \sim \mathcal{GP}(\mu_p(t; \alpha_p), k(t, t'; \boldsymbol{\phi}_p)). \tag{2}$$

The mean function of this GP is a dynamic population model $\mu_p(t; \alpha_p)$ that specifies how the individual-level functions evolve on average, conditional on parameters α_p. The individual-level functions are centered around this population model, and are shrunk towards it in a Bayesian fashion. The properties of the individual-level departures from this mean model determine the properties of the dynamic heterogeneity, and are governed by the hyperparameters of the kernel, $\boldsymbol{\phi}_p$. These parameters control both the magnitude of inter-individual heterogeneity (i.e., the degree of inter-individual shrinkage), and the degree of intra-individual temporal pooling (or smoothing). We now describe these GP components in more detail.

Mean Model The focus of this work is on capturing dynamic heterogeneity around a focal model, and we thus assume that the researcher has a specific mean model in mind. The marketing literature on dynamic modeling includes several examples, such as state space models (e.g., dynamic linear models that are typically estimated via the Kalman filter in simpler settings), traditional time series models such as ARMA (Box et al., 2015), and parametric models capturing a specific dynamic phenomenon, as in latent force models within the machine learning literature (Alvarez et al., 2013) or models of advertising dynamics in marketing (Naik et al., 1998). The mean model could also be another GP. We use different examples in our applications. Which specification is appropriate depends on the modeling context. Again, our goal here is not to compare mean models, but to illustrate their use in understanding dynamic heterogeneity. Conditional on an appropriate (or sufficiently flexible) mean model, we have found that the bigger gain in performance comes from using dynamic versus static heterogeneity, rather than the choice of mean model.

Kernel Choice The kernel captures the properties of the dynamic heterogeneity. In this work, we use the rich class of Matérn kernels, which has a general form given by:

$$k(t, t'; \eta, \kappa, \nu) = \eta^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\kappa \left|t - t'\right|\right)^{\nu} K_{\nu}\big(\kappa \left|t - t'\right|\big), \tag{3}$$

where $\eta > 0$, $\kappa > 0$, and $\nu > 0$ are the kernel hyperparameters that govern the characteristics of the function draws, $\Gamma(\cdot)$ is the gamma function, and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind. While the functional form of the kernel is unintuitive, its hyperparameters have straightforward meanings: the amplitude η controls the variability of the individual function draws around the mean function, while κ, the inverse length-scale, determines the smoothness of those function draws.[3] The degree ν also determines the smoothness of the functions by determining their level of differentiability, as draws from a GP with a Matérn kernel are $\lceil \nu - 1 \rceil$ times differentiable, where $\lceil \cdot \rceil$ is the ceiling function. Thus, the amplitude η determines the *magnitude of dynamic heterogeneity*, as it reflects how far the individual-level curves can be from the mean curve, where as κ captures the *degree of intra-individual pooling across time*.

Prior work has shown that the Matérn kernel hyperparameters cannot all be consistently estimated, and in particular, ν cannot be separately identified from κ (Zhang, 2004; Kaufman and Shaby, 2013). Hence, ν is typically fixed to a value that reflects the supposed smoothness of the underlying process. Moreover, when the degree is fixed to a half integer ($\nu = n + 1/2,\ n \in \mathbb{N}$), the complicated functional form in Equation 3 simplifies to a product of a $\lceil \nu - 1 \rceil$ degree polynomial and an exponential. For example, when $\nu = 3/2$, the kernel simplifies to:

$$k(t, t'; \eta, \kappa) = \eta^2 \left(1 + \kappa \left|t - t'\right|\right) \exp\big(-\kappa \left|t - t'\right|\big). \tag{4}$$

Fixing ν to a half integer thus makes kernel estimation more tractable. This is especially important when inference methods rely on gradients that involve the kernel function, as derivatives of the Bessel function can be computationally intensive. Furthermore, when the degree $\nu \to \infty$, the kernel converges to the squared exponential kernel that is used in Dew and Ansari (2018). Consistent with the previous literature (Rasmussen and Williams, 2005), we limit ourselves to Matérn kernel with $\nu = 1/2, 3/2, 5/2$, and ∞ (i.e., the squared exponential kernel).[4]

We use the Matérn kernel class in this work for several reasons. First, it is easier to control the smoothness of the function draws from this kernel such that momentary temporal fluctuations can be captured while still representing the underlying smoothness of

[3] A more typical form of the Matérn is:

$$k(t, t'; \eta, \rho = 1/\kappa, \nu) = \eta^2 \, 2^{1-\nu} \Gamma(\nu)^{-1} \left[\sqrt{8\nu} \left|t - t'\right| / \rho\right]^{\nu} K_{\nu}\Big(\sqrt{8\nu} \left|t - t'\right| / \rho\Big).$$

We use an inverse length-scale, slightly rescaled parametrization, such that our parameter $\kappa = \sqrt{8\nu}/\rho$ from the more typical parametrization. This follows the discussion of Fuglstad et al. (2018). Using an inverse length-scale allows us to nest the fixed offsets model as a special case, and is amenable to our choice of prior for the hyperparameters. The rescaling also helps with the interpretability of the prior.

[4] For more detailed discussion of the degree parameter, and the restriction to these four values, we refer readers to Rasmussen and Williams (2005), pp. 84-85.

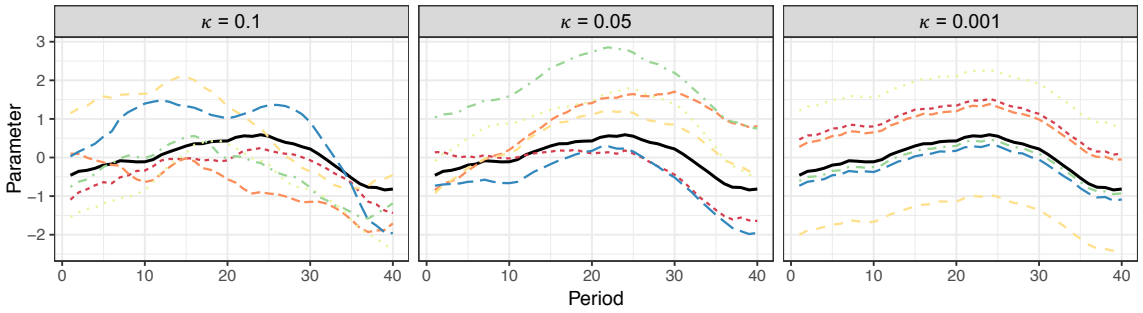

Figure 2: Draws from a GPDH model with a fixed mean function. Each panel uses a different value of κ with a Matérn-3/2 kernel. The mean function is denoted by the bold solid line. We can see that, as $\kappa \to 0$, the curves more closely mirror the mean function.

the process. This is especially suitable for the preference data in Application 1. Second, this class nests the squared exponential kernel — the typical workhorse of the GP literature and also used by Dew and Ansari (2018) — as a limiting case. Third, as we describe in the next section, the GPDH specification with the Matérn kernel nests more common heterogeneity specifications as special cases. Finally, Matérn kernels allow the use of complexity penalizing priors which facilitates fully Bayesian inference in a principled manner.

Link with Static Heterogeneity Specifications With this kernel specification, GPDH nests the static (fixed offsets) heterogeneity specification as a special case. Mathematically, the fixed offsets model assumes $\beta_{it} = \mu_t + \theta_i,\ \theta_i \sim \mathcal{N}(0, \sigma^2)$. For a fixed set of time periods, $t = 1, \ldots, T$, this is equivalent to assuming: $(\beta_{i1}, \ldots, \beta_{iT}) \sim \mathcal{N}((\mu_1, \ldots, \mu_T), \sigma^2 \mathbf{1}\mathbf{1}')$, where $\mathbf{1}$ is a T vector of ones. That is, assuming static heterogeneity around a dynamic mean model is equivalent to assuming that the full vector of parameters, $(\beta_{i1}, \ldots, \beta_{iT})$, has a multivariate normal distribution with a rank one covariance matrix, where each entry is given by σ^2. It can be shown that as $\kappa \to 0$, the Matérn-3/2 kernel, given in Equation 4 degenerates to $k(t, t'; \eta) = \eta^2$, yielding a rank one covariance matrix, which is equivalent to the fixed offsets case. In other words, as $\kappa \to 0$, GPDH converges to fixed offsets heterogeneity. This relationship holds for any member of the Matérn family of kernels. We demonstrate this convergence in Figure 2. This relationship also explains why we use the inverse length-scale parametrization, as this parametrization allows us to place a sizable prior mass on models converging to the fixed-offsets model, and thus allows us to add a prior tendency toward that restricted model. Therefore, if the posterior places a sizable mass away from zero, we can be confident that the data rejects the fixed offsets restriction. Moreover, we can use the magnitude of κ as a proxy for the extent to which individuals typically vary over time, relative to a static heterogeneity assumption.

3.3 Stage 3: Hyperpriors

We employ a fully Bayesian strategy for estimating the GPDH hyperparameters. In particular, we leverage the Penalized Complexity (PC) prior for Matérn Gaussian random fields introduced by Fuglstad et al. (2018). The PC prior is a weakly informative prior, based on the idea of penalizing the complexity induced by the kernel hyperparameters in the resultant Gaussian process. Complexity in classical Gaussian process models refers to functions with high amplitude (large η) and small length-scale (small ρ, equivalent to large inverse length-scale, κ). In GPDH, these hyperparameters have distinct meanings: the individual-level amplitude governs the degree of inter-individual shrinkage, while the inverse length-scale captures the degree of individual-level dynamics. Thus, by penalizing high amplitudes and high inverse length-scales, the PC prior encourages shrinkage across individuals, and places significant prior mass on the nested fixed offsets model. The density of the PC prior is:

$$p(\eta, \kappa) = \frac{1}{2}\lambda_1\lambda_2\kappa^{-1/2}\exp(-\lambda_1\sqrt{\kappa} - \lambda_2\eta); \quad \lambda_1 = -\log\alpha_\rho\sqrt{\frac{\rho_0}{\sqrt{8\nu}}}, \quad \lambda_2 = \frac{\log\alpha_\eta}{\eta_0}. \tag{5}$$

Despite the unintuitive functional form, another advantage of this prior is that the parameters $\eta_0, \rho_0, \alpha_\eta$, and α_ρ can be set in an intuitive way to take into account expectations on the magnitude of heterogeneity and the degree of intertemporal information sharing. Specifically, as derived in Fuglstad et al. (2018), this prior yields the following tail probabilities for η and $\rho = \sqrt{8\nu}/\kappa$:

$$P(\eta > \eta_0) = \alpha_\eta, \quad P(\rho < \rho_0) = \alpha_\rho. \tag{6}$$

In our work, we fix $\eta_0 = 5$, $\alpha_\eta = .01$, reflecting a diffuse prior assumption that the magnitude of heterogeneity will not be too large, and $\rho_0 = 1$, $\alpha_\rho = .001$, reflecting a prior assumption that the length-scale will not fall below 1.[5]

3.4 Estimation

Given the generality of our framework, the details of the estimation procedure for a hierarchical model that uses GP dynamic heterogeneity depend upon the specific individual-level model used in stage 1. We discuss our application specific strategies in the following sections. As a general point, a number of different inferential strategies have been proposed in the GP literature. These include the use of Laplace approximations, variational Bayesian, and expectation propagation methods (Rasmussen and Williams, 2005; Girolami and Rogers, 2006). Often, approximate inference techniques are used with GPs to overcome the computational complexity in estimating the function values and the hyperparameters of a GP,

[5]These values are appropriate for the choice setting considered here, where the utility is defined on a logit scale, the inputs (e.g., prices) are standardized, and time intervals are discrete months (e.g. $t = 1, 2, 3, ...$). If a larger amount of heterogeneity is expected, or the inputs are not standardized, then the tail probability for the amount of heterogeneity, η, can be adjusted by choosing a higher value of η_0, or a larger tail probability threshold α_η. The assumed values $\rho_0 = 1$, $\alpha_\rho = 0.001$ place prior mass away from length-scales, ρ, that are not properly identified: since the data is spaced evenly, very small length-scales are not identified from one another. If the data are not integer-spaced, then the value of ρ_0 can be adjusted to reflect the smallest gap between inputs that is expected.

when T, the number of time periods, is large. In our applications, as the number of time periods is not very large, we use MCMC methods for exact inference. Filippone et al. (2013) and Filippone and Girolami (2014) perform a comparative evaluation of different MCMC estimation strategies for GP models. In particular, we use the No-U-Turn sampler (NUTS) variant of Hamiltonian Monte Carlo (Hoffman and Gelman, 2014). We have found in our GPDH applications that it is important to jointly sample both the function values and the hyperparameters in one go, as the strong dependency between these sets of parameters makes HMC-within-Gibbs strategies very slow to converge.

3.5 Distinctions from Previous Work

Finally, we reiterate two important features of our approach that makes it distinct from previous work. The first is that in our specification, the GPs are used to estimate *individual-level* functions, which is distinct from using GPs to estimate mean dynamics, as in Dew and Ansari (2018). Secondly, while recent work by Yang et al. (2016) appears similar to ours in the use of collections of Gaussian processes, they model *observed* variables using GPs. In contrast, we model *latent* individual-level model parameters via Gaussian processes. Since the quantities of interest in our work are latent, we must impose more restrictions than Yang et al. (2016) on the nature of the covariance. Specifically, we assume a parametric form for the covariance kernel, which allows us to estimate the model without needing to directly observe the quantity of interest. This assumption also lets us mathematically link the GPDH method to existing heterogeneity specifications as special subcases of our specification.

4 Application I: Dynamic Preference Heterogeneity

We now apply our modeling framework to study the evolution of individual-level preferences over time in a multinomial logit choice model. We first estimate the model on synthetic data to illustrate the relative merits of GPDH and the potential pitfalls of not capturing dynamic heterogeneity. We then shift our focus to real data of grocery store purchasing during the Great Recession.

4.1 GPDH Multinomial Logit Model

We consider discrete choice data y_{it}, from individuals, $i = 1, \ldots, N$, who make choices over time $t = 1, \ldots, T$ from a choice set of $j = 1, \ldots, J$ alternatives. The choices can be explained in terms of a set of observed covariates $\mathbf{x}_{ipjt}$, indexed by $p = 1, \ldots, P$, including brand intercepts. We assume a linear utility specification, with independent, standard

extreme value (EV) errors, and with parameters modeled by GPDH,

$$u_{ijt} = \sum_{p=1}^{P} \beta_{ip}(t)\, x_{ipjt} + \epsilon_{ijt},\ \epsilon_{ijt} \sim \text{EV}(0,1), \tag{7}$$

such that consumers choose the alternative with the highest utility. For identification, we normalize the intercept of the brand with the highest market share to zero. As we order brands by market share, such normalization effectively forces Brand 1s intercept to zero. This yields the standard softmax specification for the logit choice probabilites in terms of the individual-level time-varying intercepts and sensitivities in $\beta_{ip}(t)$.[6] We then model these individual-level functions using the GPDH specificaton with a Matérn kernel:

$$\beta_{ip}(t) \sim \mathcal{GP}(\mu_p(t), k_p(t,t')), \quad \text{where} \quad k_p(t,t') = k_{\text{Matern}}(t,t';\eta_p,\kappa_p,\nu_p). \tag{8}$$

For both the simulations and the real data, we fix $\nu_p = 3/2$. We choose this value based on cross-validation using the real data. However, we also found that, in general, predictive performance was only marginally affected by the degree parameter. We include a brief discussion of kernel degree selection in Web Appendix A.

Mean Models Our emphasis is in modeling the evolution of heterogeneity around a given mean model. As such, and to illustrate the flexibility of GPDH, we test four different mean models in this application, corresponding to four common specifications in the literature:

1. Random walk state space (RW): The random walk is the simplest linear state space model that is used in the Kalman filtering literature. Our implementation is given by:

$$\mu_p(t) = \mu_p(t-1) + \zeta_{pt},\ \zeta_{pt} \sim \mathcal{N}(0, \alpha_p^2). \tag{9}$$

2. Gaussian process (GP): As in Dew and Ansari (2018), we can assume a GP as the population model:

$$\mu_p(t) \sim \mathcal{GP}(c_p, k_{0p}(t,t';\eta_{0p},\kappa_{0p},\nu_{0p})). \tag{10}$$

   We assume a constant mean c_p and a Matérn kernel, with the degree parameter ν_{0p} of this upper level kernel to be the same as in the GPDH kernel.[7] This is the mean model we assume in the simulations.

3. Autoregressive moving average (ARMA) time series: Time series models are especially common in econometric applications and can easily be incorporated into our GPDH framework. We test an ARMA(1) mean model specification, given by:

$$\mu_p(t) = \mu_{pt} = \alpha_{0p} + \alpha_{1p}\,\mu_{pt-1} + \alpha_{2p}\,\zeta_{pt-1} + \zeta_{pt},\ \zeta_{pt} \sim \mathcal{N}(0, \tau_p^2). \tag{11}$$

[6] Correlated Gaussian errors could also be used here, leading to a variant of the multinomial probit model. We favor logit choice probabilities for computational convenience.

[7] This is merely a simplifying assumption: there is no theory-based reason to fix both to have the same smoothness.

4. Parametric: A theory-driven parametric model can also serve as the mean model. In this case, one interesting question is the degree to which the Great Recession is associated with changes in consumers' preference parameters. Thus, to illustrate how a parametric model could be used in conjunction with GPDH, we use a mean function given by the PDF of a generalized inverse Gamma distribution:

$$\mu_p(t) = \alpha_{0p} + \alpha_{1p}\left[(\alpha_{2p})^{\alpha_{3p}} t^{-\alpha_{3p}-1} \exp\left(-\alpha_{2p}/t\right)/\Gamma(\alpha_{3p})\right], \tag{12}$$

with $\alpha_2, \alpha_3 > 0$. This parametric mean function allows for a unimodal pattern, with different pre- and post-peak function asymptotes, thus allowing us to isolate the impact of the recession.

For each of these, we subsequently denote the collection of parameters of the mean model generically as α. Note that α varies across different mean models.

Extensions There are many possible extensions and alternatives to the utility and mean model specifications that can incorporate other potentially desirable features alongside dynamic heterogeneity, depending on the available data and choice context. For instance, if the researcher has access to a set of potential drivers of shifts in preferences, such as individual-level events like job loss or changes in income, or market-level events like an indicator for the Great Recession, these can be incorporated directly in the mean model. Specifically, denoting these drivers generically as $\mathbf{z}_{it}$, an additive linear specification could be used, such that:

$$\beta_{ip}(t) \sim \mathcal{GP}(\gamma_p' \mathbf{z}_{it} + \mu_p(t), k(t, t'; \phi_p)), \tag{13}$$

where γ_p captures the expected effect of these drivers on preferences. In our choice modeling application, we prefer to estimate the effect of the Great Recession nonparametrically, through a flexible mean function, and we do not have other covariates available to include. However, we include a simulated example of using such parametric drivers in Web Appendix J.

A second important consideration in many choice modeling contexts is endogeneity, particularly price endogeneity. While in this work we focus just on the modeling of heterogeneity, we note the GPDH specification can be used in conjunction with methods for controlling for endogeneity. For instance, in the case of price endogeneity, the two stage control function method of Petrin and Train (2010), or the semiparametric approach of Li and Ansari (2014), could be seamlessly incorporated in the utility specification in Equation 7, together with an additional equation for the price setting process.

Estimation We estimate all variants of our GPDH logit model via HMC, using the NUTS algorithm. Specifically, we jointly sample all model parameters, including the individual-level functions, the shared mean function, and the hyperparameters. For the parameters of the mean model, we use weakly informative priors. The joint density for the full model is

given by:

$$p(y, \beta, \mu, \alpha, \phi | X) = \prod_{m=1}^{M} p(y_m | X_m, \{\beta_{i_m p}(t_m)\}_{p=1}^{P}) \times \prod_{i=1}^{I} \prod_{p=1}^{P} p(\beta_{ip}(t) | \mu_p(t), \boldsymbol{\phi}_p)\, p(\mu_p(t) | \boldsymbol{\alpha}_p)\, p(\boldsymbol{\phi}_p)\, p(\boldsymbol{\alpha}_p), \tag{14}$$

where $m = 1, \ldots, M$ indexes observations, y_m is the choice, t_m is the time period, and i_m is the individual associated with the mth observation. $\boldsymbol{X}_m$ contains the price and feature/display variables across all brands. We standardize the variables over the calibration data and report standardized results below. We run the sampler for 400 iterations (200 warmup), and measure convergence through the $\hat{R}$ statistic (Gelman et al., 1992). In all cases, we achieve $\hat{R} \approx 1$. We include more estimation details in Web Appendix B, and a discussion of computation time in Web Appendix G. The Stan code to implement the model is contained in Web Appendix I.

4.2 Simulations

In this section, we briefly describe a simulation exercise that illustrates the benefits of modeling dynamic heterogeneity. In Web Appendix C, we include additional simulations, to help understand the shrinkage properties, and computational complexity of GPDH.

To understand the benefits of capturing dynamic heterogeneity, and the potential limitations of competing approaches, we simulate multiple sets of choice data from the GPDH multinomial logit with a GP mean model. We then estimate the following three models on each of these data sets: (1) the true model (GPDH logit with GP mean); (2) a fixed offsets (FO) model that uses the GP mean model, but with static heterogeneity; and (3) an independent periods (IP) mixed logit specification that estimates a mixed logit model in each period, with only the variance of the random coefficients shared across periods, which therefore does not directly allow for within individual shrinkage across time. By simulating data with GPDH, we ensure the presence of dynamic heterogeneity. Moreover, by assuming a GP mean as the true data generating process, we nest both the FO and IP[8] specifications as limiting cases.

There are two key results from these choice model simulations. First, by sharing information both within and across individuals, GPDH yields highly efficient estimates, relative to models that assume independence across time periods. By efficient, we mean small credible intervals, while still recovering the true curve. We illustrate this in Figure 3, which shows examples of true individual-level curves and their recovery by the three specifications. GPDH, as expected, correctly recovers the true curves, and does so with a reasonable

[8]The IP specification is equivalent to the case where the length-scales of the GP mean model and the GPDH heterogeneity specification go to zero, implying no cross-period correlations for either the mean or the individual-level trajectories.

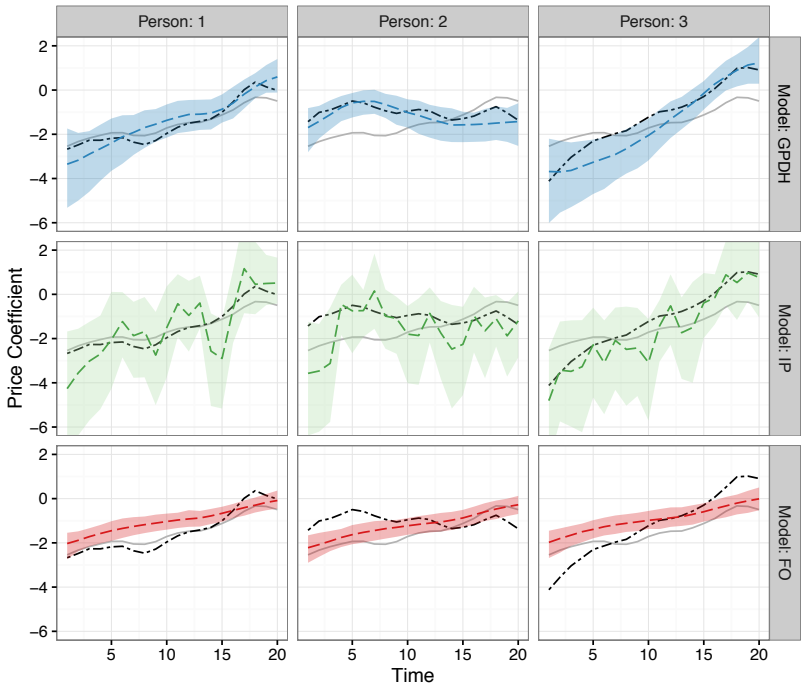

Figure 3: Three illustrative curves: In the columns, we plot three individuals' true price parameters (black/dot-dash), simulated from the GPDH model, relative to the (true) simulated population mean trajectory (grey/solid). In the rows, we then show the estimated curves (color/dashed) for those same three individuals for each of the three specifications. In the first row, we see the GPDH recovery is accurate and precise, leveraging the inter and intra-individual pooling of information to yield reasonable error bars. In the second row, using an independent periods assumption (IP), we see the estimated curves are jagged, and the error bars large, reflecting no smoothing or inter-temporal information sharing. In the last row, using a fixed offsets (FO) assumption, we see the curves have narrow error bars, but are wrong, each one reflecting the shape of the (estimated) population mean.

amount of precision, as shown by the 95% credible intervals, relative to the curves recovered by IP. Under IP, there is no inter-temporal sharing of information, leading to estimates that are quite jagged, and with much wider credible intervals. Finally, under FO, the recovered curves are simply wrong: since the FO assumes individuals are always at a fixed distance from the mean trajectory, the interesting patterns of individual-level variation are missed.

The second key result is that, if dynamic heterogeneity exists in the data, but static heterogeneity is assumed as in the FO model, the population-level estimates under FO are biased toward zero. This is the case even when the true data generating mean model is used in the FO model, as it is in our simulations. In Figure 4a, we illustrate this bias for a single simulated dataset, plotting the recovered marginal distribution of the point estimates of the coefficients at different points in time. We can see that the posterior median as recovered by FO is always biased toward zero, and that the estimated distribution of effects has significantly less variation than the truth. In Figure 4b, we show the same result, but

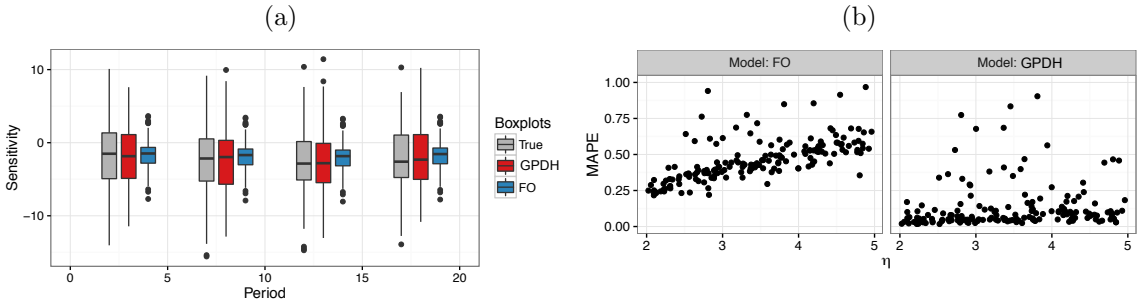

Figure 4: (a) Boxplots showing the distribution of the individual-level effects in specific periods, evaluated at four evenly spaced time periods, (b) MAPE of the recovery of the population mean dynamics across the two models for the "price" coefficient. The truncation at MAPE of 1 omits 16 observations out of 172 simulations from the plot.

from 172 repeated simulations, where we varied the η parameter of the true GPDH data generating process. For each simulated data set, we again estimated both GPDH and FO heterogeneity specifications around the same (true) mean model. Then, we computed the mean absolute percentage error in recovering the true population mean. We see that the error is higher in the FO model and increases with η, which represents the magnitude of dynamic heterogeneity in the data generating process. Taken together, these simulations suggest that the popular approach of assuming static heterogeneity around dynamic mean models may lead to biased estimates of the population mean, thereby distorting managerial decisions.

4.3 Consumer Packaged Goods in the Great Recession

We now turn our attention to modeling real choices. Specifically, we model brand choice in the IRI consumer packaged good (CPG) panel data, from January 1st, 2006 to December 31st, 2011 (Bronnenberg et al., 2008). We chose this span because it includes the Great Recession, which according to the National Bureau of Economic Research, began in December 2007, and ended in June 2009. Thus, analyzing this time period has the potential to yield purchasing dynamics of interest to both economists and managers. Specifically, we study the evolution of consumers' individual-level brand preferences, price sensitivities, and feature/display sensitivities across six different product categories: peanut butter, coffee, potato chips, laundry detergent, tissues, and toilet paper. We model the time variation at the monthly level. We retain all panelists who spent at least five times during the data and save the last four months of data for holdout validation. Summary statistics for the categories are displayed in Table 1.

Case Study: Preferences for Tissues To begin, we will focus our analysis on just one category and one model: the tissues category, and the GPDH logit model with an ARMA

Category	Brands	People	Purchases (Total)	Avg. Months (Per Person)	Avg. Purchases (Per Person)	Price Mean (SD)	% Ft/Dsp
Chips	4	1552	36152	28.29	45.45	4.12 (0.79)	96
Coffee	5	912	14298	21.73	32.31	5.38 (2.19)	91
Detergent	6	1117	16784	19.96	24.70	1.20 (0.78)	90
Peanut Butter	5	1085	16212	19.41	25.37	1.95 (0.46)	86
Tissues	4	979	15005	22.26	34.02	1.59 (2.79)	69
Toilet Paper	6	1512	26958	24.03	34.23	.61 (0.17)	83

Table 1: Summary statistics for the consumer packaged goods data, by category. *% Ft/Dsp* is the percentage of observations in which there was at least one brand featured or displayed).

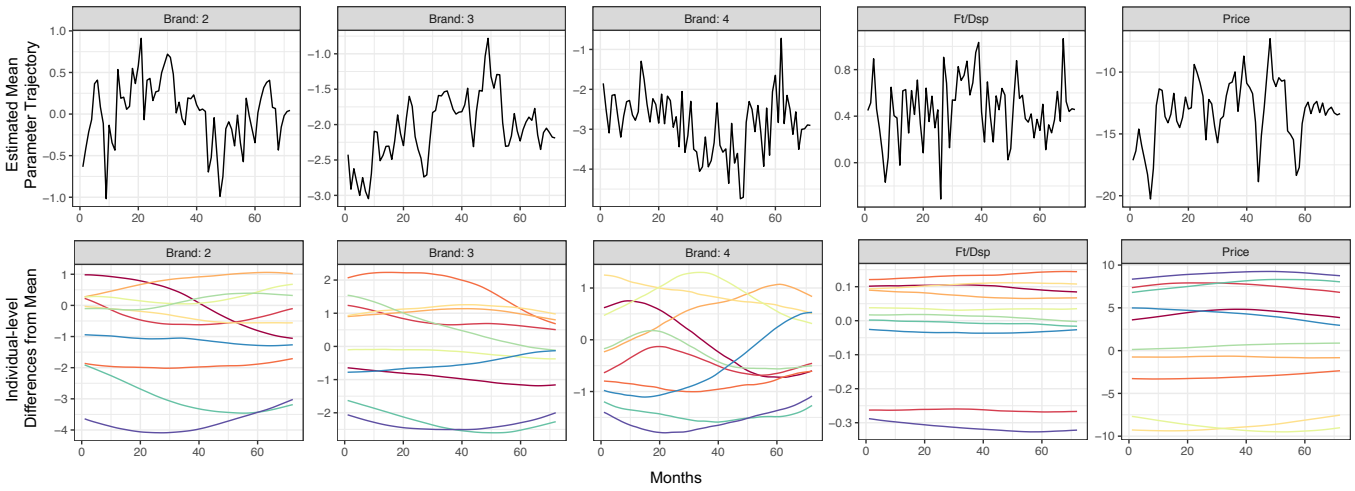

Figure 5: At top, we plot posterior mean estimates of the population means, $\mu_p(t)$, in the tissues category. The last four periods (months) are forecasts. At the bottom, we plot the difference between the individual-level curves and the mean model for a few randomly selected individuals.

mean model. We will use this specific example to illustrate the output and insights about dynamic heterogeneity that can be generated from a GPDH specification. The tissue category, in particular, generates interesting patterns of dynamic heterogeneity, and we use the ARMA mean model here as it tended to perform the best among all mean models studied. We defer a discussion of the results across all categories to the next section.

We start with the posterior estimates for the mean model $\mu_p(t)$. The top portion of Figure 5 shows these estimates for tissues. The five panels show obvious monthly dynamics. On average, the intercepts for brands 2 and 3 tended to move in opposite directions to each other, while the intercept for Brand 4 appears to track that of Brand 2 to some degree. These intercept dynamics are relative to the normalized intercept of Brand 1. Both price sensitivity and feature/display parameters also exhibit some monthly dips and spikes.

While the mean patterns are certainly interesting, the primary focus of this paper is on capturing how individuals changed *relative* to those mean trends. In the bottom portion of

Figure 5, we show the difference between the individual-level curves and the estimated mean model, $\text{Diff}_{ip}(t) = \hat{\beta}_{ip}(t) - \mu_p(t)$, for a few individuals who spent consistently throughout the data.[9] From this, we can see that, while some individuals followed the mean trajectory, resulting in flat differenced curves, others moved significantly relative to the mean function. Capturing this movement is the goal of GPDH.

The nature of the individual-level deviations is determined by the estimated hyperparameters, η_p and κ_p. As the amplitude η_p grows, the individual-level curves are allowed to spread further from the mean. As κ_p grows, the individual-level curves become less smooth. For the tissues category, the estimated posterior mean GPDH hyperparameters imply that the feature/display coefficient has a very low degree of heterogeneity, reflected in its low $\eta = .298$.[10] The feature/display coefficient also bears the closest resemblance to the fixed offsets assumption, with $\kappa = .01 \approx 0$. The price coefficient has a relatively large degree of heterogeneity, with $\eta = 6.943$, and the deviations from the mean are relatively smooth, with $\kappa = .021$. Brand 4 exhibits the least smooth variation, with the highest $\kappa = .068$. These effects are also evident from Figure 5.[11]

We now zero in on a few interesting cases of individual-level evolution that highlight the nuanced insights made possible by considering dynamic heterogeneity. We do this in Figure 6, by focusing on only a single parameter, the Brand 2 intercept, which captures the intrinsic preferences for that brand, relative to the baseline, Brand 1. We showcase individuals who spent consistently, and whose curves exhibit three interesting patterns:[12]

- Converging: In the leftmost panel, we plot a set of individual curves that converge toward the population mean. These customers started in one extreme of the distribution for the Brand 2 intercept, but by the end of the observation window, were in the middle of the distribution. Under a fixed offsets model, these individuals would be estimated as being moderately above or below the population mean, which is true only in the middle of the observation window, and does not reflect current or expected future behavior.

- Crossover: In the middle panel, we plot a set of customer curves that cross over the population mean. That is, these individuals started out liking/disliking Brand 2 (relative to others), and moved to disliking/liking (respectively) by the end of the observation. Under a fixed offsets model, these individuals would be classified as falling

[9] We define consistent by dividing the data timespan into four parts: months 1-18, 19-36, 37-54, 55-72. A consistent purchaser is one who spent during each one of these time periods. Selecting individuals in this way is important because the GPDH model exhibits mean reversion in periods where a customer does not make purchases. By selecting individuals who spent throughout the span of our data, we ensure that the patterns in this figure reflect true dynamics and not mean reversion. For the tissues category, there were 212 consistent spenders.

[10] The η parameter depends on the scale of the variables: since price and feature/display are standardized (mean zero, variance one), η can be compared across them, but cannot necessarily be compared to the intercept parameters.

[11] To contextualize these values of κ, and build intuition as to how different values of κ shape individual-level curves, we refer readers back to Figure 2.

[12] See footnote 9 for our definition of a consistent purchaser, and the rationale for restricting the sample in this way.

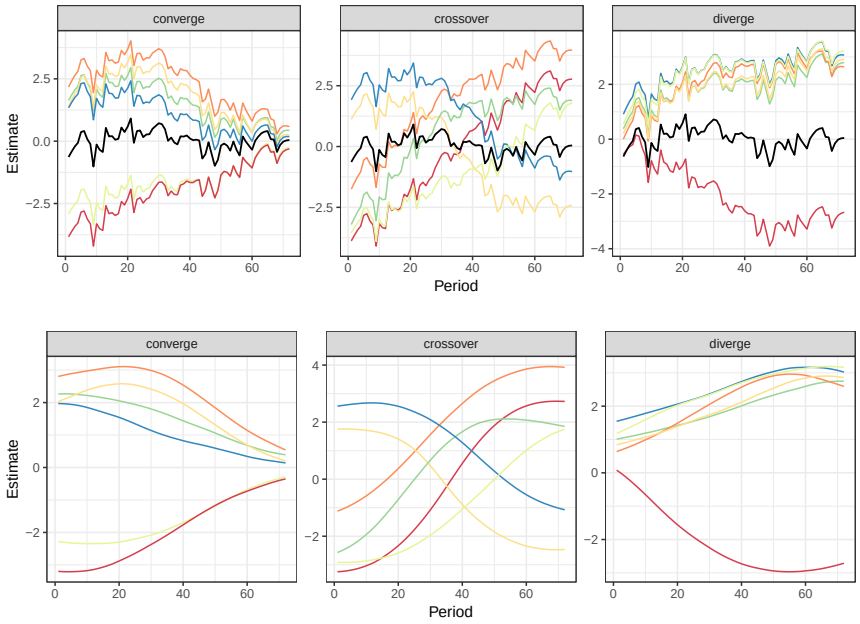

Figure 6: At top, we plot a sample of interesting individual-level curves in the tissues category, overlaid on the estimated mean model (in bold). Specifically, we isolate individuals whose curves converge toward the population mean, cross over the population mean, and diverge from the population mean. At bottom, we plot the difference between those same individual-level curves and the mean model, more clearly illustrating these changes.

near the population mean; in fact, they are perhaps the least average consumers, from a marketing research perspective, as they reflect a strong change in preferences.

- Diverging: Similar to the converging case, in the rightmost panel, we plot individual curves that diverge away from the population mean. These customers started out relatively average in their tastes for Brand 2, but moved to the extremes of the distribution over time. Under a fixed offsets model, they would be estimated as being moderately above or below the population mean, which is only true in the middle of the observation window, and again does not reflect current or expected future behavior.

Model Fit We now focus on the results across all six categories, so as to make some generalizations. The key result is that dynamic heterogeneity is pervasive across the six categories. On comparing the fixed offset model to our dynamic heterogeneity model, we find that GPDH fits the data better across all metrics, both in the calibration data, and in forecasting tasks, including on metrics that penalize model complexity. We include detailed definitions of these statistics, together with the full set of fit statistics and Bayesian measures like WAIC, in Web Appendix D. In Figure 7, we plot a subset of these measures, including in-sample and forecast sensitivity, specificity, and F1 (the harmonic mean of precision and recall), expressed as the lift from using GPDH versus static heterogeneity, across all mean models and categories. The superior fit of GPDH across nearly all of these metrics, both in Figure 7 (lift > 0) and in the appendix, strongly supports our claim that dynamic heterogeneity is present, even in relatively simple panel datasets like grocery store purchases.

Parameter Estimates and Attenuation Bias The hyperparameters of GPDH capture both the magnitude of dynamic heterogeneity for a given parameter, and how much within-individual variation there is, over time. They also allow us to assess the degree by which individual-level trajectories differ from the fixed offsets restriction. Across categories, we find that the magnitude of dynamic heterogeneity, η, is typically large, especially for brand intercepts and price sensitivity: for intercept parameters, the mean $\eta = 2$ (SD $= .64$), while for price, the mean $\eta = 2.23$ (SD $= 2.38$). For feature/display, the mean $\eta = .29$ (SD $= .11$).[13] Moreover, GPDH soundly rejects the fixed offsets model: the distribution across all categories and coefficients of κ, the inverse length-scale, is centered away from zero, with a mean $\kappa = .03$ (SD $= .02$), and with some values as high as $\kappa = .09$.[14]

We found in our simulations that not accounting for dynamic heterogeneity can lead to attenuation bias in both the mean model estimates, and in the overall extent of heterogeneity. We also find empirical evidence of the bias in our real data. Specifically, we find that the empirical standard deviation of individual-level parameters within a given time period

[13] It is difficult to directly compare η across coefficients, as it is not invariant to the scaling of the predictors: brand intercepts are binary whereas the other features are standardized (mean zero, variance one).

[14] We report all posterior mean estimates for the hyperparameters in Web Appendix H.

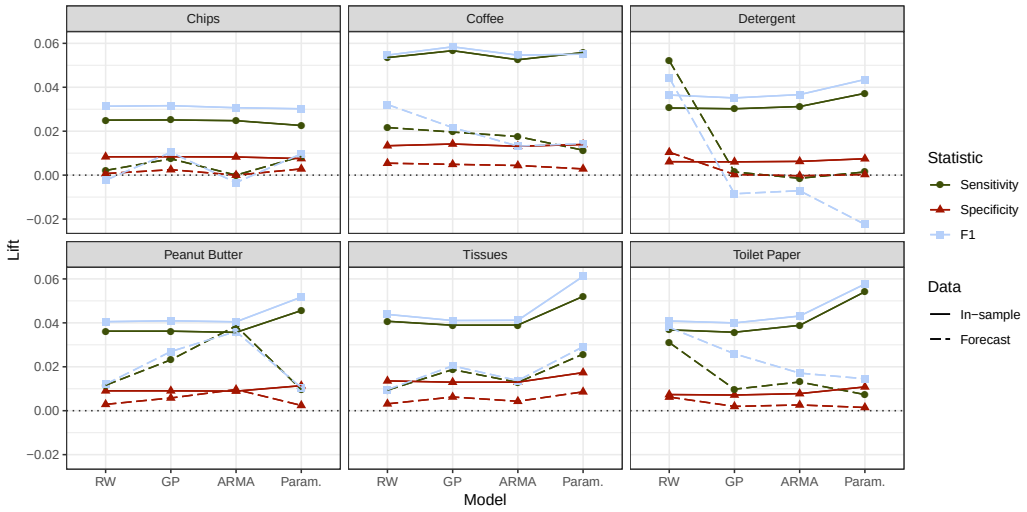

Figure 7: We compare the lift from using GPDH over static heterogeneity (i.e., fixed offsets) in in-sample and forecast fit statistics in all categories (panels), for our four population mean specifications (x-axis), both in-sample (solid lines) and forecasting ahead four months (dotted lines) using three measures of fit: micro-averaged sensitivity (dark green circles) and specificity (red triangles), and macro-averaged F1 (light blue squares), which is the harmonic mean of precision and recall. A lift greater than zero means GPDH is performing better than FO on the given fit measure. For more details about these statistics, please see Web Appendix D.

is lower when using a fixed offsets model than when using GPDH in 75% of all cases, with a maximum difference (GPDH SD - FO SD) of .244 and a minimum difference of only -.034. These results indicate a robust and oftentimes significant downward bias in the spread of FO estimates versus those from GPDH.

Moreover, when we contrast the mean curves recovered from a GPDH specification with those from the FO specification, we see the FO mean curves are biased toward zero. To illustrate this we develop what we call the *signed relative difference* (SRD) statistic:

$$\text{SRD}_p = \frac{1}{T}\sum_{t=1}^{T} \text{sign}\big(\hat{\mu}_p^{\text{GPDH}}(t)\big) \times \frac{\hat{\mu}_p^{\text{GPDH}}(t) - \hat{\mu}_p^{\text{FO}}(t)}{1 + |\hat{\mu}_p^{\text{GPDH}}(t)|}, \tag{15}$$

where $\hat{\mu}_p^{\text{GPDH}}(t)$ is the estimated value of the mean model at time t under the GPDH specification, $\hat{\mu}_p^{\text{FO}}(t)$ the estimated value of the mean model at time t under a fixed-offsets (static) heterogeneity assumption, and $\text{sign}(x) = 1$ if $x \geq 0$ and -1 if $x < 0$. This statistic will always be positive when $\hat{\mu}_p^{\text{GPDH}}(t)$ is farther away from zero than $\hat{\mu}_p^{\text{FO}}(t)$. Moreover, its magnitude reflects how much further $\hat{\mu}_p^{\text{GPDH}}(t)$ is away from zero than $\hat{\mu}_p^{\text{FO}}(t)$, on average, on a relative basis. In Figure 8, we plot the estimated signed relative differences, as a function of η, the magnitude of dynamic heterogeneity. From this, we see first that all but one of the SRD statistics are positive, across all categories and parameters, lending strong empirical support to an attenuation bias in mean parameter estimates when static heterogeneity is assumed around a dynamic mean model. Moreover, we argued previously

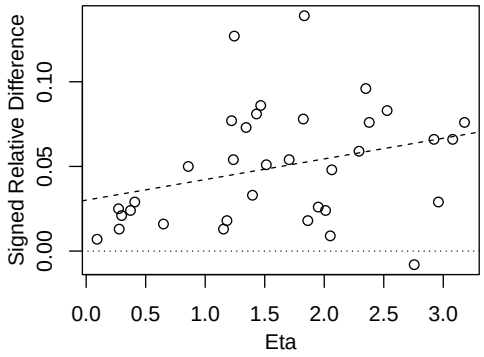

Figure 8: The signed relative difference (SRD) statistic as a function of the posterior mean estimate of the hyperparameter η. A positive SRD indicates that the mean function estimated using a fixed-offsets assumption is closer to zero than when using GPDH. The magnitude of SRD indicates how different these values are, with a higher SRD indicating a bigger standardized difference between FO and GPDH. We expect the attenuation bias to grow as the magnitude of dynamic heterogeneity (η) grows, and see this reflected in the increasing SRD statistics. We omit one outlier with $\eta > 7$ to aid visualization.

that, as the magnitude of dynamic heterogeneity (η) grows, the attenuation bias worsens. The upward trend in Figure 8 is consistent with this prediction.

Individual-level Elasticities Accounting for dynamic heterogeneity is important for accurately computing decision-relevant quantities, including time-varying price elasticities. By both correcting for the attenuation bias, and estimating intra-individual dynamics, the individual-level decision variables inferred from GPDH may be dramatically different than those based on a static heterogeneity specification. To illustrate this, we consider own price elasticity of demand across static and dynamic heterogeneity specifications. For each observation in our data, for each brand b, we compute the price elasticity using the standard multinomial logit formula, $\xi_{ibtm} = \hat{\beta}_i^{\mathrm{P}}(t) \times \mathrm{Price}_{ibtm} \times [1 - p_{ibtm}]$, where $\hat{\beta}_i^{\mathrm{P}}(t)$ is the estimated posterior mean of the price parameter for person i at time t, and p_{ibtm} is the probability that person i chooses brand b at time t, observation m, under the model. For individuals with multiple observations per time period, we average the elasticities, yielding a final elasticity estimate, $\bar{\xi}_{ibt}$. We compute the elasticities for both of the models.

First, we consider an illustrative case of a tissues consumer, selected to showcase the differences in elasticities estimated by dynamic versus static heterogeneity. In Figure 9, we present two sets of plots: in the first, we show the same consumer's choice parameters under both dynamic (GPDH) and static (FO) heterogeneity assumptions. In the second, we show the implied elasticities over time, for all periods in which the consumer was active. Comparing GPDH to FO heterogeneity in the top panel, we see two things: first, the consumer's brand intercepts deviated significantly from the pattern implied by FO, due to individual-level dynamics. This effect is especially interesting for Brand 2, where the consumer went from negative to positive. Second, we see that the price curve is *significantly*

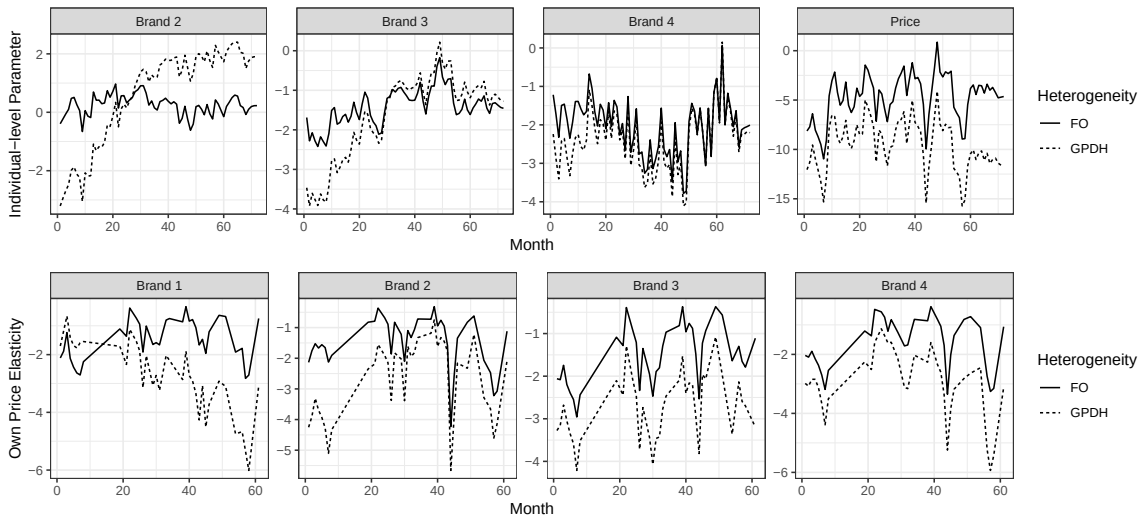

Figure 9: Top: The individual-level parameters for an illustrative consumer for the three brand intercepts and the price coefficient in the tissues category, across the two heterogeneity specifications. Bottom: the implied own price elasticities for the same consumer.

underestimated using FO, which is likely driven by the attenuation bias. Taken together, these effects produce two effects in the elasticities: first, in almost all cases, the price elasticity is underestimated by roughly 50%. Second, we see the brand intercept dynamics spill over into the price elasticities, with very different patterns implied, especially for Brands 1 and 2.

This example builds intuition around why we expect to see differences between decision variables under dynamic versus static heterogeneity assumptions. Such differences in elasticities are not limited to special cases. In fact, they are widespread across all categories. To assess these differences more generally, we compute the percentage difference in elasticities from assuming static versus dynamic heterogeneity: $\text{PD}_{ibt} = (\bar{\xi}_{ibt}^{\text{GPDH}} - \bar{\xi}_{ibt}^{\text{FO}})/\bar{\xi}_{ibt}^{\text{GPDH}}$. We present summary statistics for the distribution of PD_{ibt} across individuals, brands, and time periods in Table 2. We can see that, on average, individual-level elasticities are underestimated by using static instead of dynamic heterogeneity specifications. Moreover, the tails on the distribution are large, indicating that, for some people, the difference in estimated price elasticity between static and dynamic heterogeneity specifications is quite significant.

The Great Recession In the previous sections, we showed the applicability of GPDH to targeting and pricing. In this section, we show how GPDH can also be used by researchers to nonparametrically understand the impact of events, like the Great Recession, on individual-level consumer preferences. In particular, we use our GPDH estimates to understand the changes in individual and market-level preferences during the Great Recession. Prior literature has documented how price sensitivity within categories varies with business cycles (Gordon et al., 2013), and more generally how preferences for CPG shifted, on average, during the Great Recession (Cha et al., 2015). Similar to this previous research, we can

Category	Mean	SD	5%	25%	50%	75%	95%
Chips	3.31	4.18	-3.43	1.08	3.35	5.75	9.72
Coffee	16.89	9.59	1.33	10.82	17.42	22.74	30.59
Peanut Butter	10.58	5.24	2.20	7.31	10.47	13.76	19.80
Detergent	12.00	4.30	4.47	10.21	12.11	14.06	18.19
Tissues	6.48	4.60	-.96	3.80	6.29	9.56	14.42
Toilet Paper	9.00	4.80	2.53	5.82	8.22	11.38	18.04

Table 2: Summary statistics for the distribution of PD_{ibt}, the percentage difference between GPDH and fixed-offsets elasticity estimates, relative to the GPDH estimate, across people, brands, and time periods.

use the individual-level GPDH estimates to compute how the average price elasticity of demand changed over time during the recession. We, too, find differences in the effects of the recession on average own price elasticities across the categories, and we include a full discussion of average price elasticities over time in Web Appendix E.

Beyond mean-level analyses, a key benefit of GPDH is that we can also analyze individual-level parameter trajectories. By studying how individuals' curves deviated from the mean trajectory during the Great Recession, we can nonparametrically analyze how preferences appear to have changed during that time period. To illustrate how GPDH individual-level parametric trajectories can be used in this fashion, we created two metrics, related to the timing and impact of the Recession:

1. *Individual-level Maximal Rates of Change.* The first metric aims to understand when preferences changed most rapidly over the course of the data. To measure this, we again consider the differenced individual-level estimates $\text{Diff}_{ip}(t) = \hat{\beta}_{ip}(t) - \mu_p(t)$, which capture how each individual changed relative to the population over time, and which are displayed for the tissues category in the bottom of Figure 5. To isolate periods in which individuals changed most dramatically relative to the population, we then consider the derivative of Diff_{ip}, which we approximate using the slope of locally linear regressions. Finally, for each consumer, in each category, we select the period in which the absolute value of this numeric derivative is highest, retaining only those cases which exhibited significant variation (estimated slope $> .05$). Mathematically, this procedure approximates finding:

$$t^*_{ip} = \arg\max_t \frac{d}{dt}\text{Diff}_{ip}(t) = \arg\max_t \frac{d}{dt}\left[\hat{\beta}_{ip}(t) - \mu_p(t)\right]. \tag{16}$$

   The distribution of the timing of these maximal rates of change then serves as a metric by which we can assess the timing of distributional shifts in preferences.

2. *Timing of Crossovers.* Our second metric isolates the timing of crossovers; that is, the periods in which individual-level curves crossed over the mean curve by either going from the bottom part (half) of the distribution to the top part (half), or vice versa.[15]

[15]In the case of symmetric marginal distributions, which we often find, "bottom part" is equivalent to

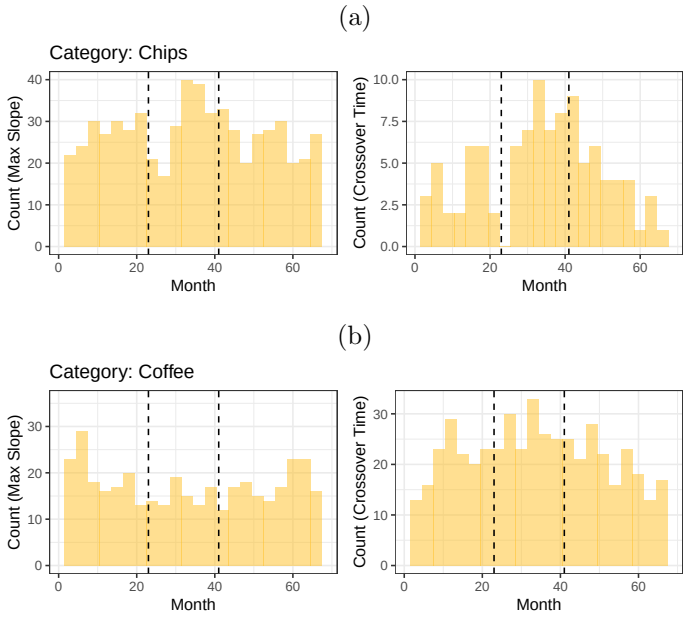

Figure 10: We present curve timing results for (a) the chips category, and (b) the coffee category. In each subfigure: At left is the distribution of the timings of maximal slopes for individual-level curves, with the recession bounded by the dashed lines. At right is the distribution of the timings of crossovers, again with the recession bounded by dashed lines.

The distribution of the timings of crossovers then allows us to assess the periods in which preferences appear to have been changing in interesting ways.

Using these two metrics, we find an apparent impact of the recession on individual-level dynamics and the distribution of heterogeneity which differs by category. In Figure 10a, for instance, we plot the result for the chips category, where there are striking peaks in both metrics associated with the beginning and the end of the recession. Similarly, we find evidence of such peaks in tissues. In other categories, most notably coffee, we find no evidence of a recession-era effect, as shown in Figure 10b. In fact, in coffee, as well as in detergent, the most rapid changes in the distribution of parameters appears to be concentrated toward the ends of the observation window.[16] While understanding the reasons behind these cross-category effects is beyond the scope of the current work, these findings illustrate the types of analyses enabled by our dynamic heterogeneity framework.

5 Application II: Dynamic Topic Heterogeneity

While heterogeneity in marketing has most often been discussed in the context of consumer preferences, GPDH is widely applicable. In this section, we apply it to a very different

"bottom half."

[16] We include the full results for curve timings across all categories in Web Appendix F.

domain: modeling the product-level evolution of review content. In particular, we fuse the latent Dirichlet allocation (LDA) topic model (Blei et al., 2003) with GPDH to capture dynamic heterogeneity in the evolution of reviews for different products. We apply our model on a dataset of time-stamped reviews for tablet computers to address questions such as, (1) how the topics used to discuss tablets have changed over time; (2) how the discourse about a focal product is affected by the introduction of new products in the marketplace; and (3) how deviations in product-level topic trajectories reflect the success or failure of the product. Our focus here is on illustrating how our framework can be used across different types of data, contexts, and models, and we therefore do not dwell at great length on the substantive conclusions in this application.

5.1 LDA-GPDH

Our model extends the standard LDA model of Blei et al. (2003) to the case where documents pertaining to different groups (e.g., products) evolve over time. In particular, we define a document as the review content of a specific product in a given calendar time period. We model the evolution of the reviews of these products, indexed $i = 1, \ldots, N$, in calendar time, $t = 1, \ldots, T$. Products are introduced at different times within those T periods, with each product's introduction time denoted by $t_i^{\min}$. We assume that there are D topics that summarize the entire discourse across all brands. Topics are probability distributions over words and capture groups of words that commonly co-occur in the reviews within a time period. We assume that the topics themselves remain static, but, over time, the topics which are emphasized in the reviews of a given product i may change. Specifically, we posit the following generative model:

- Generate each topic $d = 1, \ldots, D$, from a Dirichlet distribution, i.e., $\boldsymbol{\nu}_d \sim \text{Dirichlet}(\alpha)$, where ν_{dv} is the probability of seeing word v under topic d.

- For each topic $d = 1, \ldots, D-1$, draw the mean rate of seeing that topic over time using a GP with two length-scales of variation and a periodic component:

$$\mu_d(t) \sim \mathcal{GP}(0, k_{\text{Long}} + k_{\text{Short}} + k_{\text{Per}}).$$

  This specification mirrors the calendar time structure used in Dew and Ansari (2018). It captures momentary fluctuations in the prevalence of a given topic, as well as longer run trends and cyclical variation. We use the squared exponential kernel, which is the limiting case of the Matérn kernel as $\nu \to \infty$, for the long k_{Long} and short-run kernels k_{short}. For the periodic kernel, we use the periodic variant of the squared exponential kernel, given by:

$$k_{\text{Per}}(t, t'; \omega, \eta, \kappa) = \eta^2 \exp\left\{-\kappa \sin^2\left(\pi(t-t')^2/\omega\right)\right\},$$

  with a cycle length, $\omega = 12$, to capture monthly cyclicality. The mean rate for the Dth topic is normalized to zero for identification. Note that any of the population models described earlier could be used here; we choose the GP mean model to illustrate both

the flexibility of the mean specification, and to capture short-term and periodic spikes in chatter at certain times of the year (e.g., holidays).

- For each product $i = 1, \ldots, N$, for each time period $t = 1, \ldots, T$:
  - For each topic $d = 1, \ldots, D-1$, using a Matérn-3/2 kernel, $k_d(.)$, draw the unnormalized topic weights for product i using GPDH to capture product-level departures from population-level trends:

$$u_{id}(t) \sim \mathcal{GP}(\mu_d(t), k_d(t, t'; \eta_d, \kappa_d)).$$

  - Set the Dth topic unnormalized weight to zero: $u_{iD} = 0$
  - Compute the normalized topic assignment probabilities:

$$\beta_{id}(t) = \frac{\exp(u_{id}(t))}{\sum_{j=1}^{D} \exp(u_{ij}(t))}.$$

  - For each word token $m = 1, \ldots, M_{it}$ in product i's reviews in period t, draw a topic assignment for that word: $a_{itm} \sim \text{Categorical}(\beta_{i1}, \ldots, \beta_{iD})$.
  - Draw the actual word token from the assigned topic's word weights: $w_{itm} \sim \text{Categorical}(\boldsymbol{\nu}_{a_{itm}})$.

In some periods, a given product i may not have any reviews. For that period, the parameters are interpolated or extrapolated.

5.2 Comparison with Existing Models

The most common dynamic topic model is that of Blei and Lafferty (2006), which is often referred to simply as *the* dynamic topic model. In this model the topics evolve over time, but documents are static, and there is no accounting of heterogeneity. The focus of this model is solely on modeling the dynamics of content within one group of documents. The LDA-GPDH model is distinct from this classic dynamic topic model in that it focuses on the dynamic evolution of content for multiple groups of documents, but assumes that topics are static. The LDA-GPDH framework is thus suitable for the case where new documents are added within each group over time: in the case of reviews, we consider the unit of analysis a single product, where new reviews are continually added over the lifespan of the product. For other applications, like the modeling of scientific documents within a collection (e.g., theoretical physics papers), considered by Blei and Lafferty (2006), the documents are static. However, the words that are used in documents may change, hence requiring the evolution of the topics themselves. In other words, LDA-GPDH captures *heterogeneity* in discourse between groups of documents over time, while the dynamic topic model captures the evolution of content in a single group.

A simpler approach to model the evolution of reviews would be to apply the basic LDA model to documents defined as the composite of all of the reviews posted for a given product in a given month. Unlike LDA-GPDH, such an approach treats the reviews of a product as independent across time periods, rather than assuming some consistency of topics within products over time, thus disregarding the primary unit of analysis (the product). As a result, the topics identified by the two approaches are substantially different, with LDA-GPDH finding topics that are consistent within products. For instance, in the case of tablet computers, LDA-GPDH finds many more topics associated with specific brands, while independent LDA finds more topics associated with usage and liking. Moreover, since GPDH shares information across time periods, the topic evolutions estimated using GPDH are much smoother, allowing researchers to better separate noise from true parameter dynamics.

5.3 Estimation

Just like in the choice modeling application, we estimate LDA-GPDH using NUTS. As before, we jointly sample all model parameters, including the individual-level function coefficients, the shared mean function, and the hyperparameters. Unlike in the choice modeling application, LDA-GPDH has discrete parameters, namely the topic assignments, which cannot be sampled by NUTS. Hence, during estimation, we marginalize out the topic assignments, by computing:

$$p(w_{itm} = v \,|\, \boldsymbol{\beta}_i(t), \boldsymbol{\nu}) = \sum_{d=1}^{D} p(w_{itm} = v \,|\, a_{itm} = d) p(a_{itm} = d \,|\, \beta_{id}(t)) = \sum_{d=1}^{D} \nu_{dv} \beta_{id}(t),$$

where $\boldsymbol{\beta}_i(t) = (\beta_{i1}(t), \ldots, \beta_{iD-1}(t))$ and $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_D)$. With this marginalization, the joint distribution is given by:

$$p(w|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \prod_{i=1}^{N} \prod_{t=t_i^{\min}}^{T} \prod_{m=1}^{M_{it}} p(w_{itm} \,|\, \boldsymbol{\beta}_i(t), \boldsymbol{\nu}) \prod_{d=1}^{D-1} p(u_{id}(t) \,|\, \mu_d(t), \kappa_d, \eta_d) \times$$
$$p(\mu_d(t)|\boldsymbol{\kappa}_{0d}, \boldsymbol{\eta}_{0d}) p(\eta_d, \kappa_d, \boldsymbol{\eta}_{0d}, \boldsymbol{\kappa}_{0d}) p(\boldsymbol{\nu}_d). \quad (17)$$

As before, we run the sampler for 400 iterations (200 warmup).

5.4 Data

We apply our LDA-GPDH formulation to model the evolution of reviews in a single product category: tablet computers. We use the data from Wang et al. (2013), which contains the full set of reviews from Amazon for the tablet computer category, from September 2003 to July 2012.[17] We limit our sample to the 43 month span from January 2009 to July 2012,

[17] For a thorough overview of the raw data, we refer readers to Wang et al. (2013).

which contains the bulk of the reviews (for context, the first Apple iPad was released in April, 2010). We further restrict our sample to products that have at least 10 reviews. We aggregate these reviews at the product-month level to form our evolving document stream for each product. For the review content, we follow standard text processing procedures: we first stem the text and eliminate stopwords. We then retain all words appearing in at least 5% of observations, but not in more than 75%, where observations are period (month)/product pairs. Finally, we retain the 1000 words with the highest average term frequency-inverse document frequency (TF-IDF) scores across documents. This resulted in a dataset of 2,686 observations across 265 products.

We ran LDA-GPDH on this data using $D = 15$ topics. We selected the number of topics by running the standard LDA model multiple times with different number of topics. We found that 15 topics is the most that can be used before the topics become redundant or difficult to interpret.

5.5 Aggregate Results

We start by describing the topics learned by the model, and how their prevalence varies, on average, over time. In Table 3, we show the 10 words with the highest posterior probabilities for each topic. We see that the LDA-GPDH learns meaningful topics that tend to fall into three broad categories: functional topics, capturing aspects of how the products are used or function, especially topics 1, 3, 7, 10, 12, 13, 14; experiential topics, capturing consumers experiences with their purchases, especially topic 5; and brand topics, discussing distinct brands and products, especially topics 2 (Motorola XOOM), 4 (ASUS Transformer), 6 (Windows), 8 (Samsung), 9 (Apple), 11 (Amazon Kindle), and 15 (HP Touchpad). Note, however, that these distinctions are not always clear: topic 10, for instance, primarily discusses apps, reading, and downloads, but also has the word Kindle; topics 1 and 9 mention Apple products, but also functional words; and topic 12 has "archo," which is the stemmed form of ARCHOS, a tablet manufacturer, in addition to a number of functional words.

While the topics are static, the prevalence with which they are discussed changes. Figure 11, plots the mean model $\mu_d(t)$ for a selection of topics, reflecting the degree to which those topics are emphasized, relative to the baseline topic (Topic 15: HP Touchpad). We also plot a normalized version, given by:

$$m_d(t) = \frac{\exp(\mu_d(t))}{\sum_{\ell=1}^{D} \exp(\mu_\ell(t))},$$

which corresponds to the topic weights of the "average" product. We can see in Figure 11 that the Experiential topic (topic 5), has remained the predominant topic over time, for the average product. Other topics have waxed and waned in their prevalence. For instance, we see the relatively recent emergence of topic 10 about Reading, reflecting the increasing prevalence of this use-case in the market. We see the sharp decline in discussion of Netbooks and the Windows operating system, reflecting the growing acceptance of tablets as their own

Topic	η_d	κ_d	Terms with the highest posterior probabilities
1. Early Apple Features	.07	.09	devic, book, iphon, ipod, kindl, netbook, easi, flash, look, pdf
2. XOOM Android	2.25	.22	xoom, android, app, market, honeycomb, devic, rom, flash, googl, motorola
3. Netbook Comparisons	2.84	.33	screen, netbook, touch, keyboard, batteri, comput, lenovo, dell, upgrad, mode
4. ASUS Transformer	2.63	.21	asus, android, app, transform, thrive, acer, issu, playbook, keyboard, usb
5. Experiential	.80	.09	screen, can, like, will, just, one, good, great, time, want
6. Windows	3.22	.10	window, X7, keyboard, need, devic, comput, slate, can, laptop, pen
7. Android Apps	2.00	.16	android, market, input, devic, cobi, googl, download, flash, kyro, amazon
8. Samsung Galaxy	2.53	.36	app, tab, android, samsung, galaxi, devic, phone, X7, card, camera
9. Apple and iOS	2.47	.13	io, appl, can, X2, app, devic, like, laptop, one, new
10. Reading	2.03	.32	app, one, book, read, download, kindl, love, bought, io, can
11. Kindle Fire	2.35	.07	fire, kindl, amazon, book, read, love, like, can, devic, great
12. Media Playback	3.36	.10	archo, devic, android, app, video, player, music, firmwar, touch, file
13. Features and Development	.51	.06	develop, electron, audio, ad, charg, dollar, beauti, check, come, intern
14. Usage	2.56	.71	devic, app, will, io, amazon, one, screen, video, web, download
15. HP Touchpad	-	-	touchpad, hp, app, android, price, os, great, devic, electron, mani

Table 3: Summary of the LDA-GPDH Results: In the first two columns, we show the GPDH hyperparameter estimates across the 14 unnormalized topics. Higher values of η_d reflect more spread around the mean function, while higher values of κ_d reflect less smooth departures from the mean model, reflecting more brand-specific deviations from the mean trends. In the final column, we show the top 10 words characterizing each topic, selected by sorting posterior term probabilities.

product class, with distinct uses from netbooks and personal computers. We also see the the rise of the Apple and Samsung topics around the times of their tablet introductions. While these market dynamics make intuitive sense, understanding how individual products evolved *relative* to these trends is more interesting. We thus turn our attention to understand product-level deviations from these mean trends, which can be recovered from the GPDH specification.

5.6 Dynamic Heterogeneity in Topic Weights

Just as in the choice application, GPDH in our topic model captures individual-level departures from the mean patterns, reflecting in this case product-specific discourse trajectories. The properties of those departures are captured by the two GPDH hyperparameters, which we present for each topic in Table 3. In particular, we find that there is significantly less dynamic heterogeneity for topics 1 and 5 than for the others. As can be seen in the table, topic 1 discusses fairly generic tablet-related words, in addition to discussion of the iPhone and iPod. Moreover, as we saw in Figure 11, this topic sharply declined toward the end of the observation window. An interpretation of these patterns is that this topic captures comparisons to iPhones and iPods, which were prevalent before tablets became mainstream; after the popularization of the iPad, these topics were no longer discussed, and thus there was minimal discussion across all brands. Topic 5 captures fairly generic experiential words, and so it is again not surprising that these words occur somewhat more uniformly across brands than other topics. Topics 6 and 12 have the most heterogeneity. Both of these topics reflect somewhat technical language, as well as words associated with niche brands in the

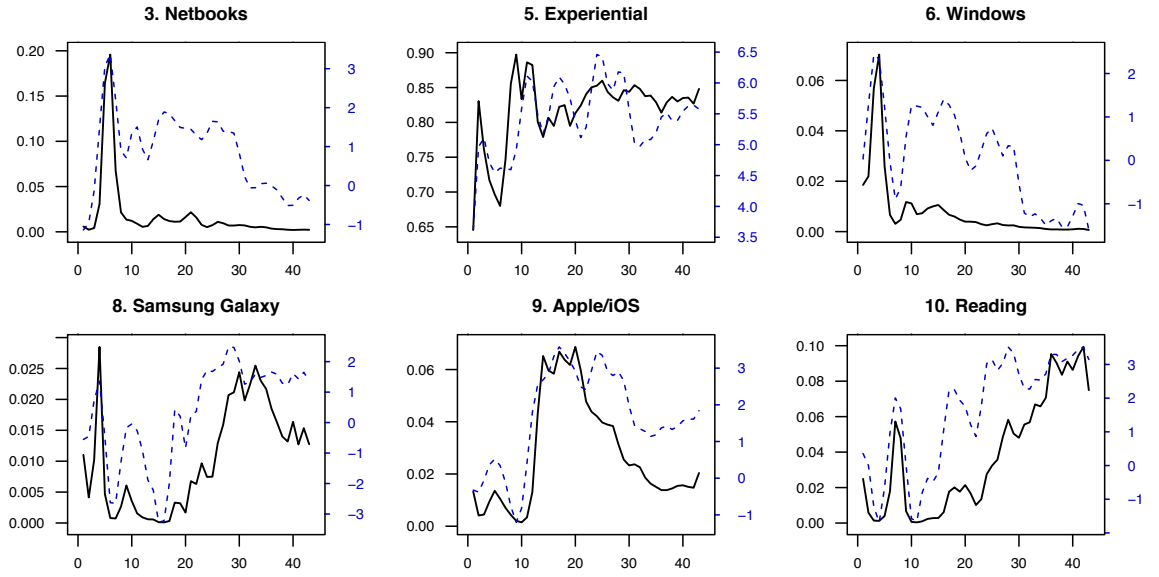

Figure 11: We plot the $\mu_d(t)$ and $m_d(t)$ functions, where each panel is one topic, for a selection of the 15 topics. The solid line is $m_d(t)$, corresponding to the scale on the left axis; the dashed line is $\mu_d(t)$, corresponding to the scale on the right axis. The x-axis is months.

tablet space (ARHOS, Windows). Thus, a large degree of variation in discussion is to be expected.

5.7 Market Structure Analysis

The key benefit of using GPDH in topic modeling is that we are able to obtain product-specific topic trajectories: for a given product, how did discourse for that product change, relative to how discourse, in general, changed? These product-level dynamics can shed light on market structure, by studying changes in the discourse for one product during time periods in which potentially competing products were introduced. For example, how did the introduction of Amazon's Kindle Fire, a highly anticipated Android tablet related to their popular Kindle e-reader, change the discourse in the reviews of existing products? Were niche products affected differently than mainstream products? Was the change in discourse in these products primarily related to brands and products, or did it relate to the functional aspects of the products, too? Answering each of these questions requires understanding how the discourse surrounding individual products changed over time.

In this analysis, we focus on the years 2011-2012, which is toward the end of our observation window. During this time, many "next generation" products were introduced, including a new generation of popular Android-based tablets, as well as Amazon's Kindle Fire, and two new versions of Apple's iPad. In particular, we saw the introduction of Amazon's Kindle Fire tablet at the end of September 2011 (period 33 in our data), and the introduction of the short-lived "new iPad" (or iPad 3) in April 2012 (period 40 in our

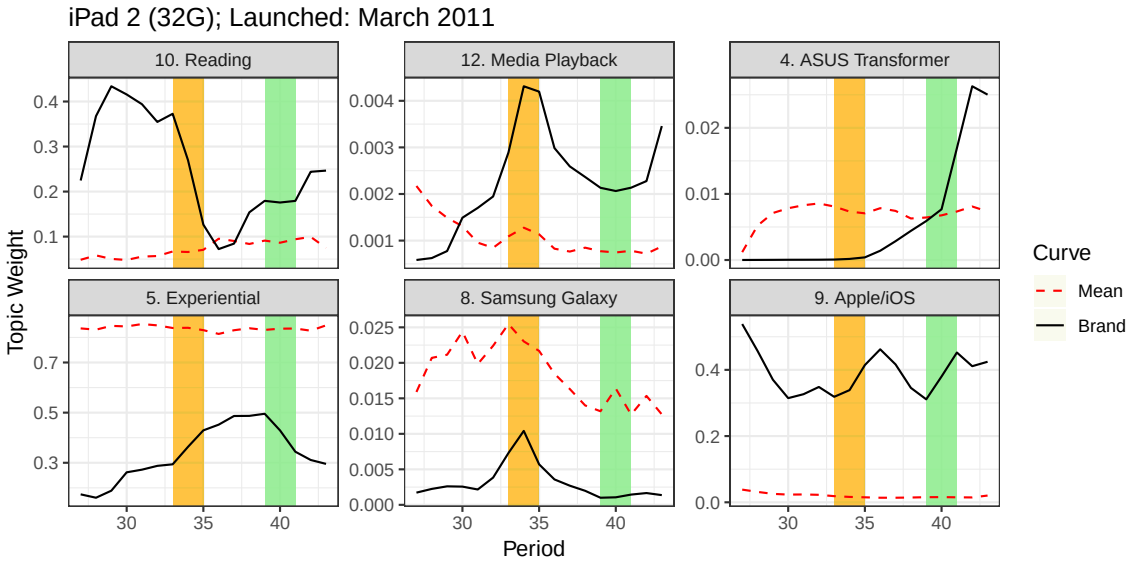

Figure 12: Dynamic topic weights for the iPad, estimated by GPDH, for six selected topics. We also plot the estimated mean probabilities for the topics. The yellow (leftmost) overlaid rectangle is the release of the Amazon Kindle Fire tablet, while the green (rightmost) rectangle is the release of the Apple iPad 3.

data).[18] To understand how these introductions affected product-level chatter, and what this may imply about market structure, we begin by considering a couple of case studies, before reporting results across products.

Case Study: Reviews of the iPad 2 Figure 12 shows several of the dynamic topics learned for the iPad 2 (32GB version). In the figure, we highlight in yellow (the leftmost overlaid rectangle) the release window of the Kindle Fire, and in green (the rightmost rectangle) the release date of the iPad 3. The first thing of note is the significant dynamics present during the release of the Kindle. We see that topic weights during that time shifted from Reading to Experiential, as well as to topics about the Apple brand and the iPad. There was also a notable uptick in chatter at that time about Samsung products, reflecting the launch of a new Samsung tablet simultaneous with the Kindle Fire, and an uptick in chatter about Media Playback. After the release of the newer generation iPad 3, we also find changes, this time in chatter about Apple/iOS, about some competing Android-based products (e.g., ASUS products), and again about Reading.

These patterns reflect different aspects of the tablet market structure. First, while the Kindle Fire was Android-based, it appears to have attracted significant attention even among iPad users, especially when considering reviews that focus on reading. Previous

[18]The period of the release of a product and the period of its first review are sometimes not the same, most often due to products being reviewed before the official release, or slightly different release dates of different versions of the same product. We thus consider a 3 period (month) window of time around the official release dates as the "release window" of a focal product.

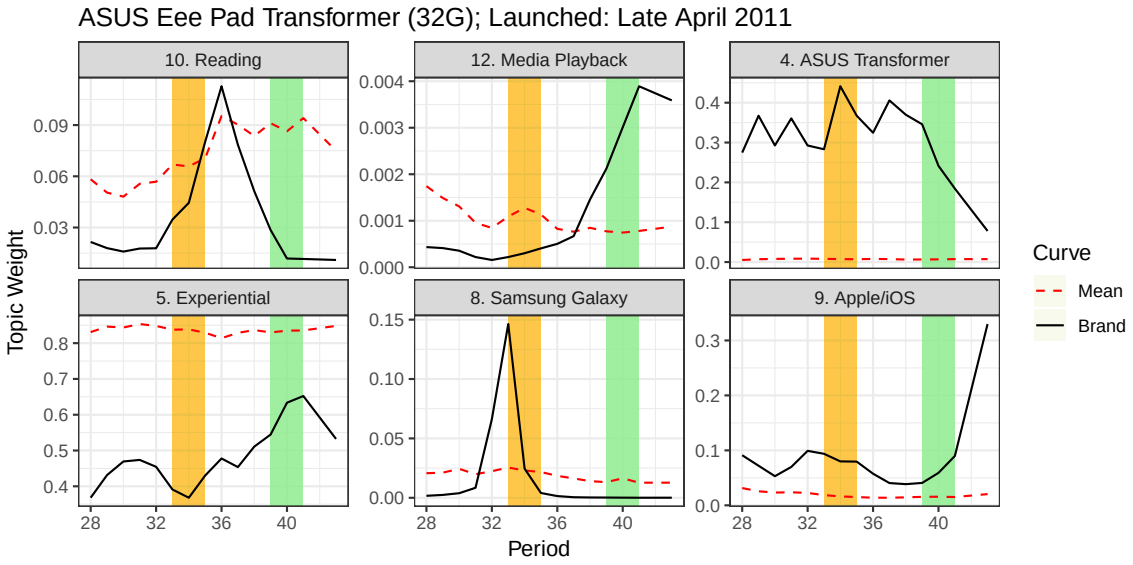

Figure 13: Dynamic topic weights for the ASUS Eee Pad Transformer, estimated by GPDH, for six selected topics. We also plot the estimated mean probabilities for the topics. The yellow (leftmost) overlaid rectangle is the release of the Amazon Kindle Fire tablet, while the green (rightmost) rectangle is the release of the Apple iPad 3.

versions of the Kindle were electronic readers, but not full fledged tablets. With the Kindle Fire, Amazon entered the tablet market, but retained its emphasis on reading. In the ways in which the iPad reviews changed during this period, it appears that this move did attract attention, with customers who previously reviewed the iPad for its reading capacity largely having vanished after the Fire's introduction. The changes after the introduction of the newest iPad also reveal aspects of the market structure. The uptick in discussion of competing Android products and reading are consistent with a change in the customer base after the new iPad release: customers who continued buying the older version are likely customers who are more price sensitive, or are looking for a tablet with more basic functionality. Thus, we see an increase in chatter about competing but lower priced brands, as well as a focus on a more basic functionality (reading).

Case Study: Reviews of the ASUS Transformer Figure 13 plots the dynamics over the topics for the ASUS Eee Pad Transformer (32GB version). Immediately, we can see differences between this and the iPad. First, there are clear "fixed offset" style differences versus the iPad example. For instance, the ASUS brand topic is consistently higher for the ASUS Transformer's reviews than it was for the iPad's reviews. However, there are also clear differences in the dynamics of the topic weights, and departures from the mean-level trends, which are captured by the flexible GPDH framework. For example, similar to the iPad, we also see a rise in discussion of Experiential aspects over the product's lifecycle, reflecting a shift from more functional descriptions of the product to more experiential ones. In terms of responses to product releases, the ASUS topics appear to have been significantly

affected by the release of the new iPad, but not as much by the release of the Kindle Fire. While there was a significant spike in chatter about Reading after the release of the Kindle Fire, it was short-lived. However, after the release of the iPad 3, we see a huge bump in chatter about Apple and the iPad, and a corresponding drop in chatter about the ASUS brand.

Topic dynamics again reflect the tablet market structure. In addition to the dynamics plotted in Figure 13, several other topics are high toward the start of the ASUS Transformer's lifecycle, capturing different competing brands or products: topic 2, about the Motorola XOOM and Honeycome Android operating system, for instance, started out high. Similarly, topic 3, about netbooks and touchscreens, started out high, which is especially relevant since the Transformer had an attachable keyboard option. Likewise, topic 6 about Windows started out high. Early reviewers emphasized comparisons with these products, reflecting the market position of the Transformer as a mix of these products. The lack of impact of the Kindle Fire's introduction, together with the seemingly large impact of the iPad's also reflects aspects of the product's use: the Transformer was largely aimed at replacing laptops as a mobile computing system, and did not emphasize as much the reader aspect. When the new iPad was released, it likely attracted significant attention from this customer base, as reflected in these changes in topics. Finally, the significant but short-lived spike in chatter about the Galaxy also appears to reflect comparisons to that product upon its release, but with no lasting impact, suggesting perhaps different user bases.

Results Across Products Finally, we consider patterns of product-level topic evolution across all of the products in our dataset. Just as in the choice modeling application, where GPDH allowed us to define new metrics to capture interesting dynamics during the recession, we can also consider new metrics based on dynamic heterogeneity for topic weights, to systematically characterize product-level review dynamics. In particular, we consider a question that was raised by our case studies: during which periods of time did individual products exhibit the most change in topics? To answer this, we consider the following metric:

$$\delta_{it} = \frac{1}{K} \sum_{k=1}^{K} \left| \hat{\beta}_{ik}(t) - \hat{\beta}_{ik}(t-1) \right|,$$

which reflects the average per period change (from $t-1$ to t) in the estimated product-specific topic weights, $\hat{\beta}_{ik}(t)$. We find that the shape of the empirical distribution of δ_{it} across i does not change very significantly with time. However, the positions of individual products within that distribution vary considerably. To better understand when products undergo the biggest changes in topic weights, we look again at two focal periods: the period after Kindle's introduction ($t = 34$), and the period after iPad 3's introduction ($t = 40$). Among the top 20 δ_{it} statistics for $t = 34$, we find many eReader products, like the *Pandigital Planet Android 7-Inch Multimedia Tablet* and *Color eReader with Kindle*. In fact, we find four Pandigital products among the top 20 δ_{it} statistics in that period, indicating substantial competition between that brand and the Kindle Fire. For $t = 40$, we most notably find a variety of previous iPad versions, which is intuitive, as the release of a new version of the iPad likely changed the discourse about prior versions. Although many of these results are

intuitive, this analysis highlights the types of analyses afforded by dynamic, product-level topic trajectories, which can be estimated by a GPDH specification.

6 Conclusion

We developed a novel methodology for capturing dynamic heterogeneity in models of parametric evolution. Across two applications of our GPDH framework to essential marketing tasks, we showed the rich insights that come from modeling the evolution of the distribution of cross sectional heterogeneity. In our first application, we illustrated the importance of capturing dynamic heterogeneity in choice models with evolving sensitivities, and the managerial and economic insights uncovered by our GPDH specification. In our second application, we showcased the wide applicability of GPDH, by employing it in a very different context: a topic model capturing the product-level evolution of review content. We applied this model to reviews of tablet computers, and used these product-level topic trajectories to shed light on aspects of market structure.

Both applications demonstrate the versatility of the GPDH specification. As GPDH is a way of specifying heterogeneity for the parameters of a focal likelihood, around a particular mean model, it can be used with a number of different likelihoods and dynamic mean models of interest. In this work, we showcased five distinct mean models, ranging from Kalman filter-like state space models, to time series specifications, to multi-component Gaussian process models. As an extension, we also suggested how to incorporate the drivers of dynamics within mean models, though we did not have the appropriate data to demonstrate that use case. Additionally, we discussed briefly how endogeneity concerns can be handled via standard control function methods. While we focused only on the benefits of dynamic heterogeneity, and thus did not fully explore these extensions, we note that they may be important for researchers interested in adapting our framework in other substantive contexts.

Our work has several limitations that suggest opportunities for future research. First, especially in Application I, we only observe existing customers. As a result, we cannot rule out that the observed patterns of heterogeneity are driven by different customer lifetimes (i.e., left censoring). However, the patterns of dynamics we uncover still correctly reflect the changes that occur at a given point in *calendar* time, regardless of the underlying source of those dynamics. Moreover, given the mature and common nature of the studied CPG categories, we do not expect left censoring to be the primary driver of our results. In addition, due to our emphasis on the methodological contribution of GPDH, we did not fully explore some of the interesting substantive phenomena that were revealed by our GPDH specification. In particular, in Application 1, we noted the prevalence of shifts in brand intercepts versus price coefficients during the Great Recession, as well as the heterogeneous impact of the recession on different categories. In Application 2, we uncovered associations between discourse and product lifecycles. Understanding the mechanisms behind these phenomena is beyond the scope of this work, but may prove interesting topics for future research. Finally,

from a computational perspective, our implementation of GPDH using MCMC methods is somewhat slow. Recent advances in Bayesian inference, including variational methods (e.g., Ansari et al., 2018), may prove valuable in accelerating the computation time for these models.

Lastly, it is worth noting that while we used GPDH in the context of dynamic heterogeneity, our framework is generally applicable for modeling collections of functions defined on *any* index, not just time. Other use cases may include spatial modeling and functional modeling of variables. As the modeling of both heterogeneity and dynamics is crucial to marketing, we hope that GPDH will be used and extended for research across a wide variety of domains.

References

Adams, R. P., Murray, I., and MacKay, D. J. (2009). Nonparametric bayesian density modeling with gaussian processes. *arXiv preprint arXiv:0912.4896*.

Alvarez, M. A., Luengo, D., and Lawrence, N. D. (2013). Linear latent force models using gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2693–2705.

Ansari, A. and Iyengar, R. (2006). Semiparametric thurstonian models for recurrent choices: A bayesian analysis. *Psychometrika*, 71(4):631.

Ansari, A., Li, Y., and Zhang, J. Z. (2018). Probabilistic topic model for hybrid recommender systems: A stochastic variational bayesian approach. *Marketing Science*, 37(6):987–1008.

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

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Braun, M. and Bonfrer, A. (2011). Scalable inference of customer similarities from interactions data using dirichlet processes. *Marketing Science*, 30(3):513–531.

Braun, M., Fader, P. S., Bradlow, E. T., and Kunreuther, H. (2006). Modeling the pseudo-deductible in insurance claims decisions. *Management Science*, 52(8):1258–1272.

Bronnenberg, B. J., Kruger, M. W., and Mela, C. F. (2008). The IRI Marketing Data Set. *Marketing Science*, 27(4):745–748.

Bronnenberg, B. J. and Sismeiro, C. (2002). Using multimarket data to predict brand performance in markets for which no or poor data exist. *Journal of Marketing Research*, 39(1):1–17.

Cha, W. M., Chintagunta, P. K., and Dhar, S. K. (2015). Food Purchases During the Great Recession. *SSRN ID #2548758*.

DeSarbo, W. S., Ansari, A., Chintagunta, P. K., Himmelberg, C., Jedidi, K., Johnson, R., Kamakura, W., Lenk, P., Srinivasan, K., and Wedel, M. (1997). Representing Heterogeneity in Consumer Response Models. *Marketing Letters*, 8(3):335–348.

Dew, R. and Ansari, A. (2018). Bayesian Nonparametric Customer Base Analysis with Model-based Visualizations. *Marketing Science*, 37(2).

Filippone, M. and Girolami, M. (2014). Pseudo-marginal bayesian inference for gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226.

Filippone, M., Zhong, M., and Girolami, M. (2013). A comparative evaluation of stochastic-based inference methods for gaussian process models. *Machine Learning*, 93(1):93–114.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2018). Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, pages 1–8.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Girolami, M. and Rogers, S. (2006). Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8):1790–1817.

Gordon, B., Goldfarb, A., and Li, Y. (2013). Does Price Elasticity Vary with Economic Growth? A Cross-category Analysis. *Journal of Marketing Research*, 50(February):4–23.

Guhl, D., Baumgartner, B., Kneib, T., and Steiner, W. J. (2018). Estimating time-varying parameters in brand choice models: A semiparametric approach. *International Journal of Research in Marketing*, 35(3):394–414.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Kaufman, C. G. and Shaby, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484.

Khan, R., Lewis, M., and Singh, V. (2009). Dynamic Customer Management and the Value of One-to-One Marketing. *Marketing Science*, 28(6):1063–1079.

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2004). Assessing heterogeneity in discrete choice models using a dirichlet process prior. *Review of marketing Science*, 2(1).

Kim, J. G., Menzefricke, U., and Feinberg, F. M. (2005). Modeling Parametric Evolution in a Random Utility Framework. *Journal of Business & Economic Statistics*, 23(3):282–294.

Kottas, A. (2006). Dirichlet process mixtures of beta distributions, with applications to density and intensity estimation. In *Workshop on Learning with Nonparametric Bayesian Methods, 23rd International Conference on Machine Learning (ICML)*.

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

Li, Y. and Ansari, A. (2014). A bayesian semiparametric approach for endogeneity and heterogeneity in choice models. *Management Science*, 60(5):1161–1179.

Liechty, J. C., Fong, D. K. H., and DeSarbo, W. S. (2005). Dynamic Models Incorporating Individual Heterogeneity: Utility Evolution in Conjoint Analysis. *Marketing Science*, 24(March 2015):285–293.

Naik, P. A. (2015). Marketing Dynamics: A Primer on Estimation and Control. *Foundations and Trends in Marketing*, 9(3):175–266.

Naik, P. A., Mantrala, M. K., and Sawyer, A. G. (1998). Planning media schedules in the presence of dynamic advertising quality. *Marketing science*, 17(3):214–235.

O'Hagan, A. and Kingman, J. F. C. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–42.

Pauwels, K. and Hanssens, D. M. (2007). Performance regimes and marketing policy shifts. *Marketing Science*, 26(3):293–311.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. *Journal of marketing research*, 47(1):3–13.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Sriram, S., Chintagunta, P. K., and Neelamegham, R. (2006). Effects of Brand Preference, Product Attributes, and Marketing Mix Variables in Technology Product Markets. *Marketing Science*, 25(5):440–456.

Wang, X., Mai, F., and Chiang, R. H. (2013). Database submissionmarket dynamics and user-generated content about tablet computers. *Marketing Science*, 33(3):449–458.

Williams, C. K. and Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.

Xie, J., Song, X. M., Sirbu, M., and Wang, Q. (1997). Kalman filter estimation of new product diffusion models. *Journal of Marketing Research*, 34(3):378–393.

Yang, J., Zhu, H., Choi, T., and Cox, D. D. (2016). Smoothing and mean-Covariance estimation of functional data with a bayesian hierarchical model. *Bayesian Analysis*, 11(3):649–670.

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

Web Appendix

A: Kernel Degree Selection

As described in the body of the paper, not all parameters of the Matérn kernel are consistently estimable. Hence, we follow common practice and fix the degree parameter ν to a half-integer value, or to ∞, which corresponds to the squared exponential kernel. By fixing ν to a half-integer, the functional form of the kernel reduces to a product of a polynomial term and an exponential term, which facilitates computation, compared to the original Bessel function formulation. Specifically, we consider the values $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$. This follows the advice of Rasmussen and Williams (2005), who argue that values of $\nu > \frac{5}{2}$ are difficult to distinguish from ∞, given typical data sizes. In the GPDH settings, this lack of distinction is even more the case, as the GPs are being specified several levels away from the data, as governing the parameters of a latent utility (or the mean of those parameters).

The ν parameter controls the level of differentiability of the function draws, as illustrated in Figure 14. Thus, if a less smooth process is desired, or theorized a priori, the researcher may choose to use a lower value ν (e.g., $\nu = 1/2$). Alternatively, cross-validation may be used to set the value of ν. In our Application I, we find little difference both in terms of fit and prediction across different values of the degree parameter. As an example, we plot a comparison of fit and forecasting accuracy across ν values in Figure 15 for the Peanut Butter category. The value $\nu = 3/2$ does marginally better in forecasting tasks, and imposes less stringent assumptions on smoothness, assuming only once differentiable function draws. Hence, we use $\nu = 3/2$ as our primary specification throughout the paper.

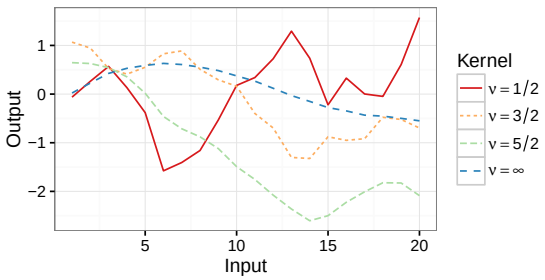

Figure 14: The impact of the choice of the degree parameter on the level of differentiability or "smoothness" of the function draws.

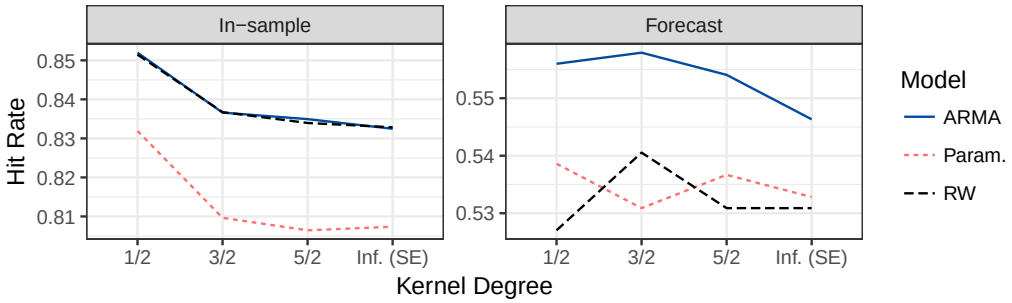

Figure 15: A comparison of the fit and forecasting ability of the GPDH-logit model across different values of the kernel degree parameter ν, across all non-GP mean models, on the peanut butter data. We limit consideration to non-GP mean models because of our assumption that, when using a GP mean, the degree parameter is the same as in the GPDH kernel.

B: Estimation Details

In this section, we give the explicit forms of the densities used in Equations 14 and 17 for estimating the models in our two applications.

Application 1 The joint density for the full model given in Equation 14 is reproduced here:

$$p(y, \beta, \mu, \alpha, \phi|X) = \prod_{m=1}^{M} p(y_m|X_m, \{\beta_{i_m p}(t_m)\}_{p=1}^{P}) \times \prod_{i=1}^{I}\prod_{p=1}^{P} p(\beta_{ip}(t)|\mu_p(t), \boldsymbol{\phi}_p)\, p(\mu_p(t)|\boldsymbol{\alpha}_p)\, p(\boldsymbol{\phi}_p)\, p(\boldsymbol{\alpha}_p).$$

The individual-level model is a multinomial logit such that

$$p(y_m = j|X_m, \{\beta_{i_m p}(t_m)\}_{p=1}^{P}) = \frac{\exp\left(\sum_{p=1}^{P} \beta_{i_m p}(t_m)\, x_{i_m pjt}\right)}{\sum_{\ell=1}^{J} \exp\left(\sum_{p=1}^{P} \beta_{i_m p}(t_m)\, x_{i_m p\ell t}\right)}.$$

The GP heterogeneity specification is given by

$$p(\beta_{ip}(\boldsymbol{t})|\mu_p(t), \boldsymbol{\phi}_p) = \text{MVN}(\mu_p(\boldsymbol{t}), K(\boldsymbol{t}, \boldsymbol{t}; \boldsymbol{\phi}_p)),$$

where $\boldsymbol{t} = \{1, 2, \ldots, T\}$ represents the vector of time points on which the GP is defined. The probabilistic representation for $p(\mu_p(t)|\boldsymbol{\alpha}_p)$ depends upon the specific mean-model used. For the ARMA(1) specification, we have

$$p(\mu_p(t)|\boldsymbol{\alpha}_p) = N(\alpha_{0p} + \alpha_{1p}\, \mu_{pt-1} + \alpha_{2p}\, \zeta_{pt-1}, \tau_p^2).$$

Finally, $p(\boldsymbol{\phi}_p)$ represents the PC prior, such that

$$p(\eta, \kappa) = \frac{1}{2}\lambda_1\lambda_2\kappa^{-1/2}\exp(-\lambda_1\sqrt{\kappa} - \lambda_2\eta); \;\; \lambda_1 = -\log\alpha_\rho\sqrt{\frac{\rho_0}{\sqrt{8\nu}}}, \;\; \lambda_2 = \frac{\log\alpha_\eta}{\eta_0}.$$

The last term, $p(\boldsymbol{\alpha}_p)$, represents the prior over the parameters in the specific mean-model used. We chose appropriate diffuse priors for these parameters.

Application 2 The joint density is given by

$$p(w|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \prod_{i=1}^{N}\prod_{t=t_i^{\min}}^{T}\prod_{m=1}^{M_{it}} p(w_{itm}\,|\,\boldsymbol{\beta}_i(t), \boldsymbol{\nu}) \prod_{d=1}^{D-1} p(u_{id}(t)\,|\,\mu_d(t), \kappa_d, \eta_d) \times p(\mu_d(t)|\boldsymbol{\kappa}_{0d}, \boldsymbol{\eta}_{0d}) p(\eta_d, \kappa_d, \boldsymbol{\eta}_{0d}, \boldsymbol{\kappa}_{0d}) p(\boldsymbol{\nu}_d).$$

In the above, $p(w_{itm} = v \mid \boldsymbol{\beta}_i(t), \boldsymbol{\nu}) = \sum_{d=1}^{D} \nu_{dv}\beta_{id}(t)$, where $\boldsymbol{\beta}_i(t) = (\beta_{i1}(t), \ldots, \beta_{iD}(t))$ and $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_D)$. The GPDH term is given by

$$p(u_{id}(\boldsymbol{t}) \mid \mu_d(t), \kappa_d, \eta_d) = \text{MVN}(\mu_d(\boldsymbol{t}), K_d(\boldsymbol{t}, \boldsymbol{t}, \eta_d, \kappa_d)),$$

where $\boldsymbol{t} = \{1, 2, \ldots, T\}$ represents the vector of time points on which the GP is defined. The GP mean-model can be represented as

$$p(\mu_d(\boldsymbol{t}), \boldsymbol{\eta}_{0d}, \boldsymbol{\kappa}_{0d}) = \text{MVN}(\mathbf{0}, K_{Long}(\boldsymbol{t}, \boldsymbol{t}) + K_{Short}(\boldsymbol{t}, \boldsymbol{t}) + K_{Per}(\boldsymbol{t}, \boldsymbol{t})).$$

We used independent PC priors for each of the hyperparametes of the different GPs. The last term is given by $p(\boldsymbol{\nu}_d) = \text{Dirichlet}(\alpha)$.

C: Additional Choice Modeling Simulations

The simulations in the body of the paper assumed for simplicity that each individual purchased the same number of times. Here, we consider the more realistic case where individuals differ in the number of observations, with some spending often, and others very infrequently. Specifically, we vary four aspects of the data-generating process: (1) the number of time periods in the data; (2) the number of individuals in the data; (3) the minimum number of purchases needed for an individual to be included in the data; and (4), the variance of the number of purchases per person. We then study how the results of GPDH differ, in terms of fit, insights, and computation time.

To investigate the performance of GPDH as a function of all of these inputs, we ran a series of simulations, varying four aspects of the data generating process: the number of people ($N = 100, 200, 400$), the number of time periods ($T = 20, 40, 60$), the minimum number of spends per person ($m_{\min} = 1, 3, 5$), and the variance of the number of spends per person over the entire time window. We simulated the number of spends, m_i, as $m_i = m_{\min} + a_i$, $a_i \sim \text{Round}[\text{Gamma}(1, s)]$, where s is the scale parameter of the gamma, and varied $s = 2, 10, 20$. As before, we assume that the true data generating process is a GPDH multinomial logit with a GP mean model, and that there are three brands and a price variable.

Shrinkage Estimator Just like all Bayesian approaches to modeling heterogeneity, GPDH can be viewed as a shrinkage estimator, wherein individuals' parameter trajectories are shrunk toward the mean trajectory. For consumers with very few purchases, their estimated trajectories mirror the mean function, in terms of both shape and magnitude, with large credible intervals. For consumers with many purchases, their trajectories are estimated to be closer to the true, data-generating trajectories. We illustrate this in Figure 16, for a simulation with 200 people, and a minimum number of purchases of 3. Because of these shrinkage properties, GPDH can still perform well, even when the number of purchases per person is small. In Table 4, we show that, across all levels of simulated data sparsity, GPDH achieves superior in-sample hit rates, compared to the FO assumption.[19]

Computational Complexity In the GP literature, it is well established that the computation time for GP-based models scales at $O(T^3)$, where T is the number of unique inputs (Rasmussen and Williams, 2005). In our applications, the number of inputs is always fixed at the number of months in the data. In GPDH, there are two additional aspects of the data size, besides the number of inputs, which may affect scalability: the number of individuals, and the number of observations per individual.

Across these simulations, we found that $m_{\min}$ and s did not have a consistent effect on computation time. We plot the effect of N and T in Figure 17. When the number of time periods is small ($T = 20$), we see there is little difference in the computation time

[19]The full set of fit statistics is available from the authors upon request.

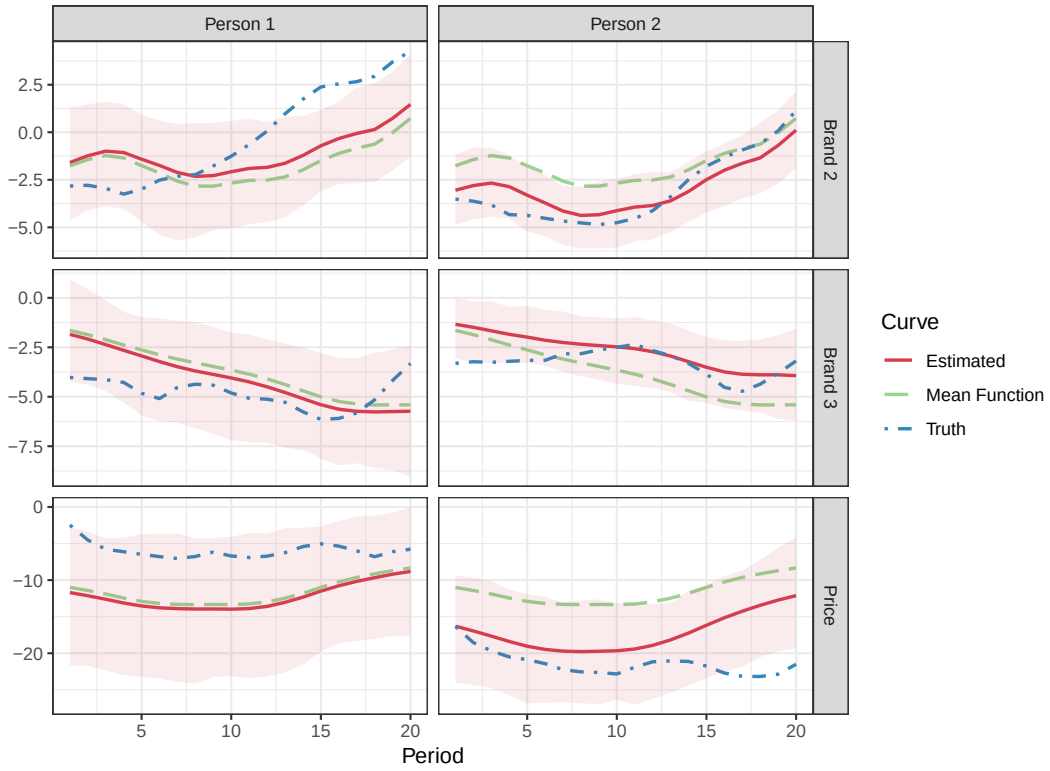

Figure 16: Shrinkage properties of GPDH: at left is a person with only 3 spends; at right is a person with 60 spends. We see that Person 1's estimated curves closely follow the mean function, while Person 2's estimated curves recover the truth, with some shrinkage toward the estimated mean function. The shaded bands are 95% credible intervals around the estimated individual-level trajectory.

Min Spends ($m_{\min}$)	1			3			5			Overall
Spend Variance (s)	2	10	20	2	10	20	2	10	20	
FO	.865	.849	.849	.870	.856	.842	.860	.848	.842	.853
GPDH	.896	.901	.902	.915	.911	.899	.900	.902	.891	.902

Table 4: Hit rate as a function of data sparsity, comparing the GPDH and FO specifications.

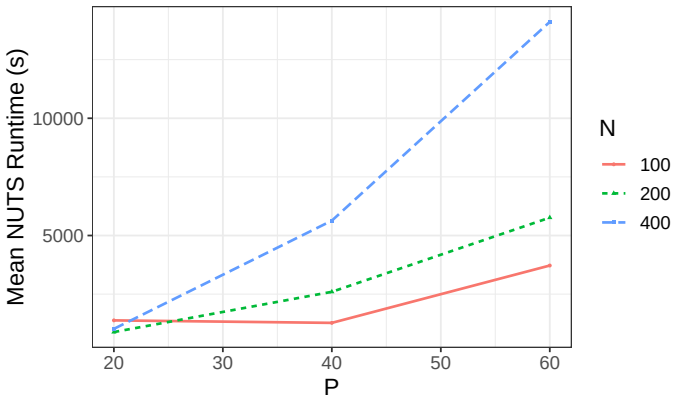

Figure 17: Average computation time, in seconds, for estimating GPDH on simulated data as a function of the number of people in the data N (the line colors and patterns), and the number of time periods T in the data (the x-axis). These results are averaged across m_{min} and s.

with respect to N. However, for larger number of time periods, the increase in computation time from increasing N becomes pronounced. Holding N fixed, we see that the increase in computation time from changing T is non-linear, consistent with the $O(T^3)$ scaling of GPs generally.

D: Full Fit Statistics (Application 1)

In this section, we present more fit statistics. As a whole, all fit statistics imply that GPDH significantly outperforms statistic heterogeneity, given the same mean model. In the main body of the paper, we presented several representative fit statistics in Figure 7. In this appendix, we also plot in Figure 18 the basic hit rates (accuracy) across specifications and data settings, and in Figure 19 the Watanabe-Akaike Information Criterion (WAIC), which is a Bayesian measure that measures model fit, penalizing for model complexity. We see that this measure again supports the idea that dynamic heterogeneity, as captured through GPDH, better describes the data, even taking into account the added complexity of the model. Interestingly, we find little difference in fit across mean models, except for a noted decrease in fit for the restrictive parametric model.

We also include here the full set of fit statistics, averaged across mean models, for all categories and heterogeneity specifications, in Table 5. Those statistics are based on the following counts, for a given brand b: True positives (TP_b) = the number of observations where the model predicted the consumer would choose brand b, and the consumer chose brand b ; False positives (FP_b) = the number of observations where the model predicted the consumer would choose brand b, but the consumer did not choose brand b; True negatives (TN_b) = the number of observations where the model did not predict the consumer would choose brand b, and the consumer did not choose brand b; and False negatives (FN_b) = the number of observations where the model did not predict the consumer would choose brand b, but the consumer chose brand b. From these, we compute the following statistics:

- Precision (Prec) - also called the hit rate, equal to $TP_b/(TP_b + FP_b)$,
- Sensitivity (Sens) - also called recall or the true positive rate, equal to $TP_b/(TP_b + FN_b)$,
- Specificity (Spec) - also called selectivity or the true negative rate, equal to $TN_b/(TN_b + FN_b)$,
- F1 - the harmonic mean of recall (Sensitivity) and precision.

Finally, we average these across brands in the following ways:

- Macro average: the average of each of the above rates. Intuitively, this aggregation treats all classes equally, ignoring potential class imbalance.
- Micro average: this aggregation computes the above statistics by summing over b at each step. Intuitively, this takes into account class imbalance, at the risk of showing good performance when one class dominates.
- Max: the max over b. Intuitively, this is the statistic for the class that was easiest to predict.

- Min: the min over b. Intuitively, this is the statistic for the class that was most difficult to predict.

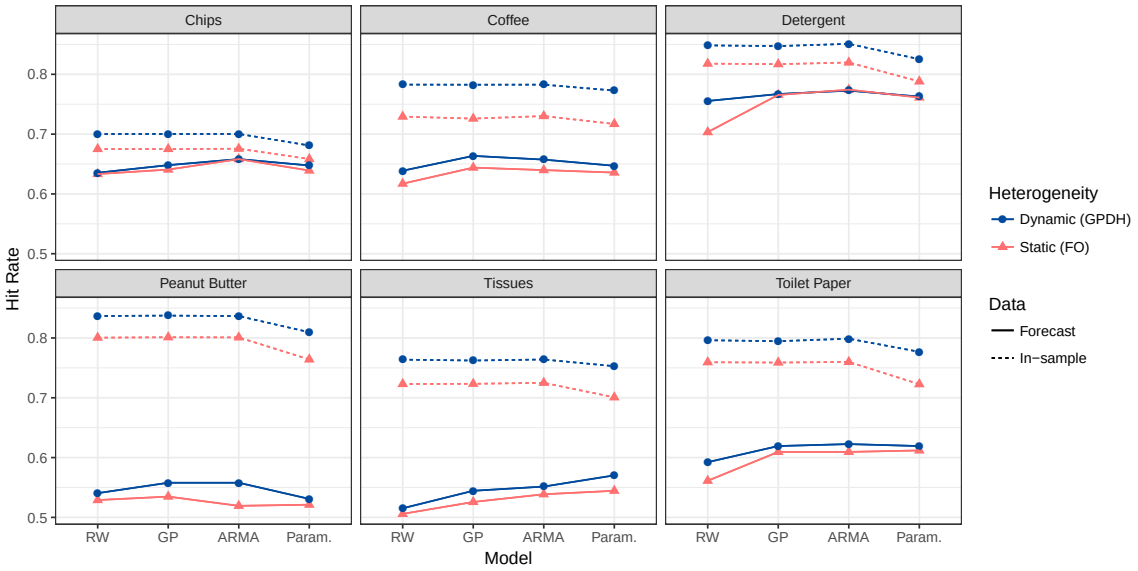

Figure 18: Hit rates (accuracy) across model specifications and data (in-sample and forecast), analogous to Figure 7

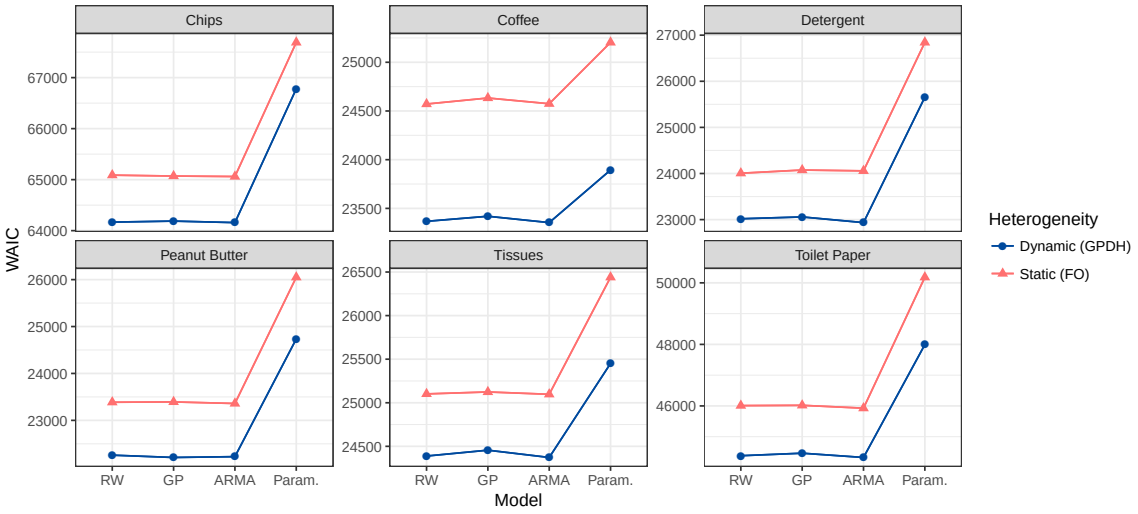

Figure 19: WAIC across model specifications. Lower indicates better fit, taking into account model complexity.

In-sample

		Macro			Micro			Max			Min		
Category	Heterogeneity	Prec	Sens	Spec	Prec	Sens	Spec	Prec	Sens	Spec	Prec	Sens	Spec
Chips	GPDH	.695	.629	.882	.695	.695	.898	.709	.832	.978	.665	.484	.724
Chips	FO	.665	.599	.873	.671	.671	.890	.689	.817	.975	.631	.434	.706
Coffee	GPDH	.782	.754	.940	.780	.780	.945	.804	.844	.985	.756	.675	.879
Coffee	FO	.729	.696	.926	.726	.726	.931	.763	.805	.983	.673	.608	.854
Detergent	GPDH	.836	.813	.967	.843	.843	.969	.868	.918	.992	.796	.732	.927
Detergent	FO	.801	.774	.960	.811	.811	.962	.844	.905	.991	.717	.665	.913
Peanut Butter	GPDH	.832	.819	.956	.830	.830	.958	.874	.872	.984	.810	.749	.929
Peanut Butter	FO	.789	.776	.946	.792	.792	.948	.858	.843	.982	.717	.632	.911
Tissues	GPDH	.762	.751	.916	.761	.761	.920	.776	.788	.970	.741	.703	.866
Tissues	FO	.717	.704	.901	.718	.718	.906	.735	.752	.963	.701	.630	.840
Toilet Paper	GPDH	.791	.781	.957	.792	.792	.958	.818	.846	.984	.742	.710	.924
Toilet Paper	FO	.746	.736	.948	.750	.750	.950	.789	.818	.981	.709	.665	.911

Forecast

		Macro			Micro			Max			Min		
Category	Heterogeneity	Prec	Sens	Spec	Prec	Sens	Spec	Prec	Sens	Spec	Prec	Sens	Spec
Chips	GPDH	.646	.528	.865	.647	.647	.882	.777	.769	.991	.524	.182	.704
Chips	FO	.618	.527	.863	.643	.643	.881	.778	.758	.987	.434	.200	.708
Coffee	GPDH	.631	.629	.904	.652	.652	.913	.756	.718	.967	.522	.521	.812
Coffee	FO	.615	.607	.900	.634	.634	.909	.753	.699	.969	.492	.521	.810
Detergent	GPDH	.717	.618	.947	.764	.764	.953	.878	.931	.996	.453	.177	.858
Detergent	FO	.720	.611	.943	.751	.751	.950	.877	.910	.996	.502	.218	.833
Peanut Butter	GPDH	.519	.545	.886	.547	.547	.887	.752	.650	.950	.331	.332	.822
Peanut Butter	FO	.492	.520	.879	.526	.526	.882	.673	.643	.931	.312	.279	.820
Tissues	GPDH	.558	.559	.846	.545	.545	.848	.712	.735	.924	.459	.470	.768
Tissues	FO	.541	.539	.840	.529	.529	.843	.699	.706	.923	.445	.458	.759
Toilet Paper	GPDH	.588	.574	.920	.613	.613	.923	.747	.782	.978	.369	.239	.859
Toilet Paper	FO	.562	.550	.917	.598	.598	.920	.722	.813	.980	.304	.169	.852

Table 5: Fit statistics average across mean model. The statistics are described above.

E: Average Elasticity Over Time

Similar to previous analyses of the Great Recession, our GPDH results can be used to nonparametrically study how price elasticity, on average, changed during the Great Recession, by simply averaging over individuals. Below, we present the full set of price elasticity plots over time. In the detergent, chips, and toilet paper categories, we find many brands experienced significant increases in average price elasticity. This is intuitive as the Great Recession negatively affected many people's earnings, which should lead to higher price sensitivity. Peanut butter and coffee, on the other hand, do not appear to have been significantly impacted. Finally, tissues appears to have behaved almost countercyclically during the recession: for all brands in tissues, the average recession-era price elasticity was smaller than before and after. There are several caveats to this population-level analysis, which may limit its interpretability or generalizability, and which also limits its comparability to previous studies, e.g., Gordon et al. (2013). Importantly, in this work, we only modeled choice conditional on the purchase decision, and do not capture effects like stockpiling. We also use a relatively lenient rule for retaining consumers in the panel, such that consumers that purchased at least five times were included. This means our estimates of average price elasticities may be subject to panelist attrition.

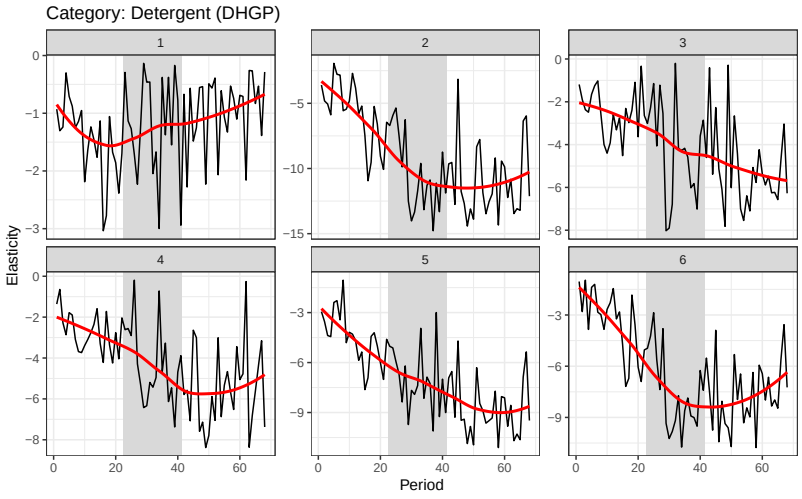

Figure 20: The average price elasticity of demand across detergent brands over time, as estimated by the GPDH logit model. The recession era, as defined by NBER, is marked by the grey rectangle. Overlaid on the estimated average price elasticities is a local linear smoothing (LOESS).

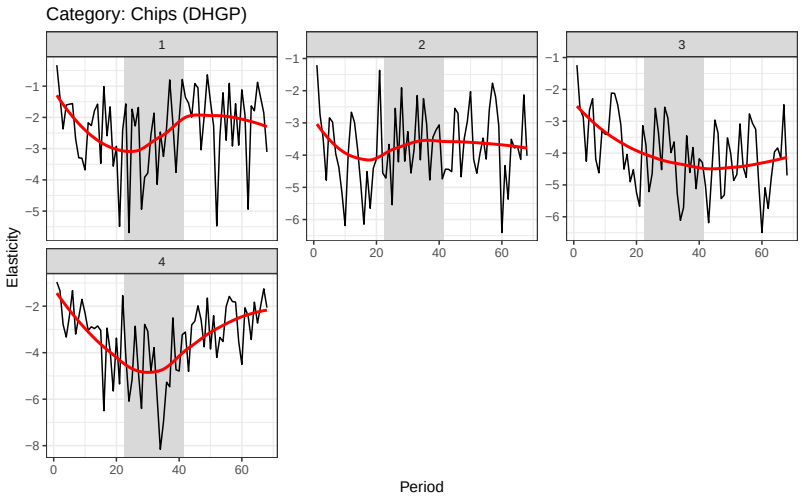

Figure 21: The average price elasticity of demand across chips brands over time, as estimated by the GPDH logit model. The recession era, as defined by NBER, is marked by the grey rectangle. Overlaid on the estimated average price elasticities is a local linear smoothing (LOESS).

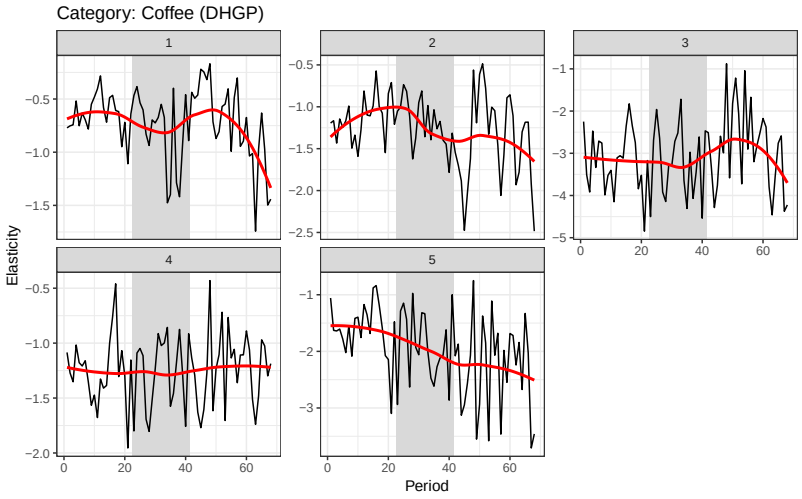

Figure 22: The average price elasticity of demand across coffee brands over time, as estimated by the GPDH logit model. The recession era, as defined by NBER, is marked by the grey rectangle. Overlaid on the estimated average price elasticities is a local linear smoothing (LOESS).

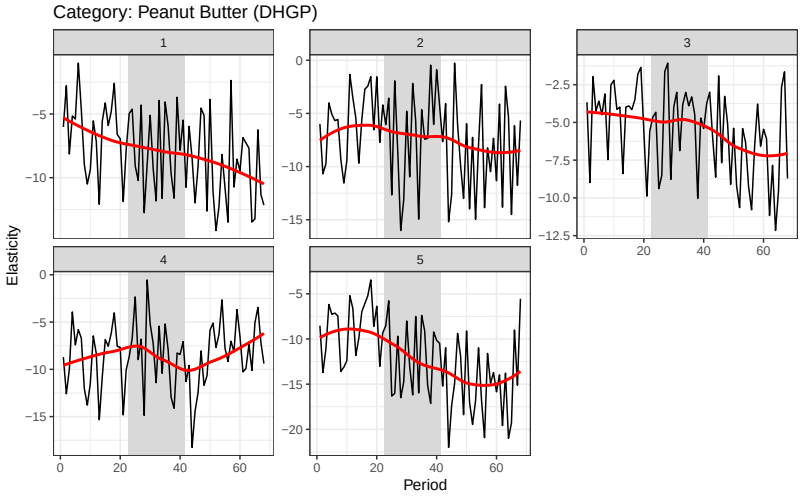

Figure 23: The average price elasticity of demand across peanut butter brands over time, as estimated by the GPDH logit model. The recession era, as defined by NBER, is marked by the grey rectangle. Overlaid on the estimated average price elasticities is a local linear smoothing (LOESS).

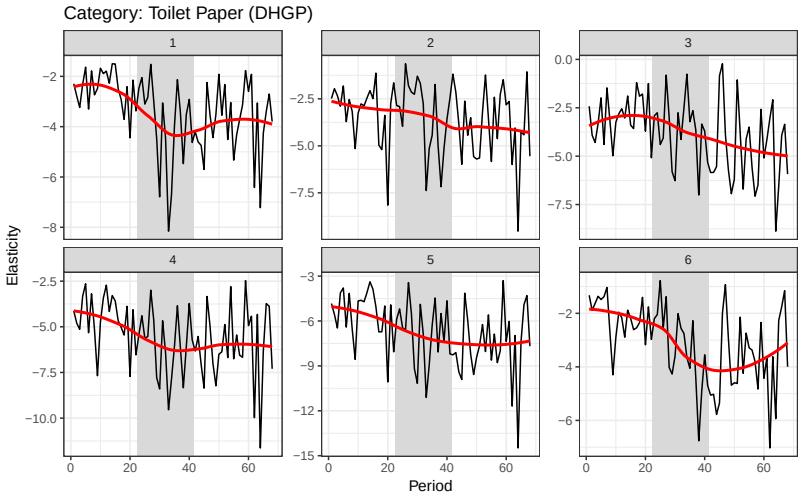

Figure 24: The average price elasticity of demand across toilet paper brands over time, as estimated by the GPDH logit model. The recession era, as defined by NBER, is marked by the grey rectangle. Overlaid on the estimated average price elasticities is a local linear smoothing (LOESS).

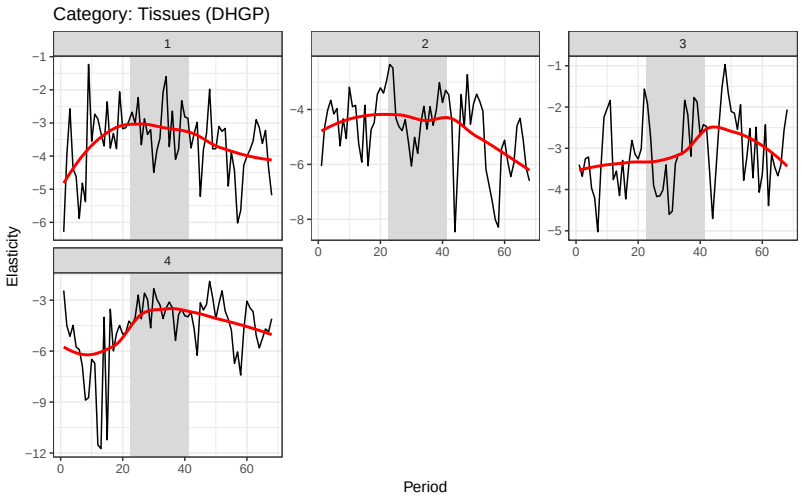

Figure 25: The average price elasticity of demand across tissue brands over time, as estimated by the GPDH logit model. The recession era, as defined by NBER, is marked by the grey rectangle. Overlaid on the estimated average price elasticities is a local linear smoothing (LOESS).

F: Curve Timing Plots

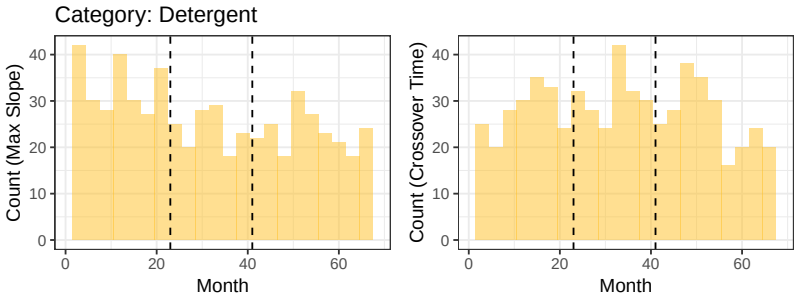

Figure 26: At left, the distribution of the timings of maximal slopes for individual-level curves in the detergent category, with the recession bounded by the dashed lines. At right, the distribution of the timings of crossovers in the chips category, again with the recession bounded by dashed lines.

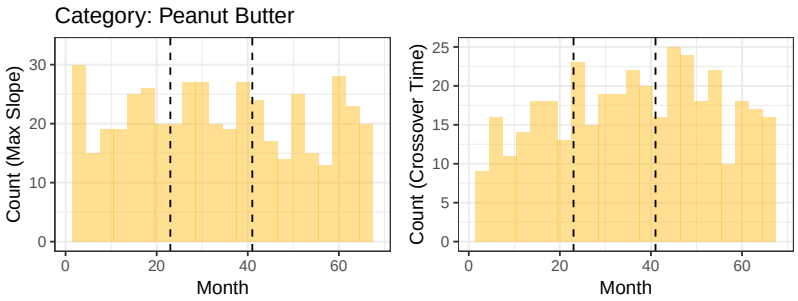

Figure 27: At left, the distribution of the timings of maximal slopes for individual-level curves in the peanut butter category, with the recession bounded by the dashed lines. At right, the distribution of the timings of crossovers in the coffee category, again with the recession bounded by dashed lines.

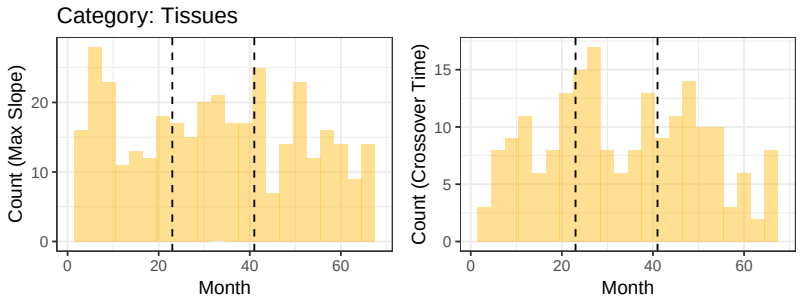

Figure 28: At left, the distribution of the timings of maximal slopes for individual-level curves in the tissues category, with the recession bounded by the dashed lines. At right, the distribution of the timings of crossovers in the chips category, again with the recession bounded by dashed lines.

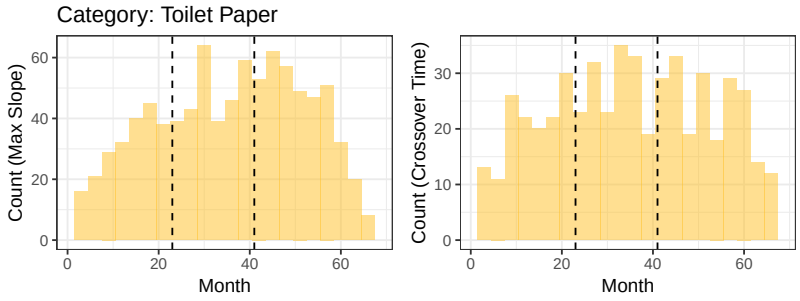

Figure 29: At left, the distribution of the timings of maximal slopes for individual-level curves in the toilet paper category, with the recession bounded by the dashed lines. At right, the distribution of the timings of crossovers in the coffee category, again with the recession bounded by dashed lines.

G: Computation Times

We report in Table 6 the computation times for each of the product categories from Application 1, for each of the mean model specifications. There is a positive correlation between the number of people, the number of purchases, and the computation time, although with only six categories, it is difficult to make further claims about scalability. The average time across all of the categories was 26 hours.

Category	# People	# Purchases	ARMA	GP	Param	RW	Average
Chips	1552	36152	33.6	13.8	25.4	16.9	22.41
Coffee	912	14298	17.6	22.4	15.5	12.2	16.92
Detergent	1117	16784	37.6	40.7	46.6	41.3	41.56
Peanut Butter	1085	16212	18.9	10.7	19.8	16.2	16.40
Tissues	979	15005	18.2	22.1	25.6	21.6	21.89
Toilet Paper	1512	26958	34.6	69.1	27.3	19.2	37.52

Table 6: Computation time, in hours, across the categories and mean model specifications.

H: Hyperparameter Estimates

Parameter		Chips	Coffee	Detergent	Peanut Butter	Tissues	Toilet Paper
Brand 2	η	1.18	2.29	3.18	1.86	2.01	2.05
	κ	.02	.02	.02	.03	.03	.02
Brand 3	η	1.71	2.76	1.47	1.82	1.51	1.95
	κ	.04	.02	.04	.03	.03	.02
Brand 4	η	1.15	2.53	1.24	1.40	1.43	2.96
	κ	.04	.03	.01	.00	.07	.03
Brand 5	η		2.38	1.83	3.08		2.92
	κ		.07	.03	.06		.02
Brand 6	η			2.06			1.24
	κ			.08			.09
Ft/Dsp	η	.09	.27	.37	.28	.30	.41
	κ	.04	.00	.01	.00	.01	.03
Price	η	.65	1.34	2.35	1.22	6.94	.86
	κ	.03	.02	.02	.02	.02	.02

Table 7: Hyperparameter estimates for application 1, across all categories and coefficients.

I: Stan Code

Here, we include the Stan code for the GPDH-ARMA choice model.

```
functions{
 real maternk(real x1, real x2, real eta, real kappa, int type){
    // NOTE ON THE TYPE INPUT:
    // type 0: matern-1/2
    // type 1: matern-3/2
    // type 2: matern-5/2
    // type 3: squared exponential (technically, should be type infinity)

    real r = fabs(x1-x2);
    real out;
    if (type == 0) {out = eta^2 * exp(-kappa*r); }
    if (type == 1) { out = eta^2 * (1+kappa*r) * exp(-kappa*r); }
    if (type == 2) { out = eta^2 * (1 + kappa*r + pow(kappa*r, 2)/3.0) * exp(-kappa*r); }
    if (type > 2) { out = eta^2 * exp(-pow(kappa*r, 2));}
    return out;
 }

 matrix Kmat(int P, real eta, real kappa, int type, real jitter){
    matrix[P, P] cov;

    for(i in 1:P){
        for(j in 1:P){
            cov[i,j] = maternk(i, j, eta, kappa, type);
         }
        cov[i,i] = cov[i,i] + jitter;
    }
    return cov;
 }

  // Penalize complexity prior for the hyperparameters of a matern kernel,
  // from Simpson et al., 2017
 real pc_prior_lpdf(vector hypers, real ktype, real eta_upper, real alpha_eta, real rho_lower,
   real alpha_rho){
    real degree = ktype + 0.5;
    real eta = hypers[1];
    real kappa = hypers[2];
    real lambda1 = -log(alpha_rho) * sqrt(rho_lower / sqrt(8.0*degree));
    real lambda2 = -log(alpha_eta) / eta_upper;

    return (log(0.5) + log(lambda1) - 0.5*log(kappa) - lambda1*sqrt(kappa) +log(lambda2) - lambda2*eta);
  }
}

data{
 int<lower=2> B;     // no. goods
 int<lower=1> N;     // no. customers
 int<lower=1> P;     // no. periods per customer
 int<lower=1> M;     // no. total choices

 int y[M];              // choice of customer on each choice occassion
 matrix[M,B] price;     // prices for each good at each choice occassion
 matrix[M,B] ftdsp;     // display/feature for each good at ...
 int<lower=1,upper=N> id[M];   // person id
 int<lower=1,upper=P> pd[M];   // period id
 int ktype;
}

parameters{
```

```
 // mean functions: ------------------------------------------------
 vector[P] mu_icept[B-1];
 vector[P] mu_price;
 vector[P] mu_ftdsp;

 // mf var parameters:
 real<lower=0> tau_icept[B-1];
 real<lower=0> tau_price;
 real<lower=0> tau_ftdsp;

 // mf arma parameters:
 real m_icept[B-1];
 real<lower = -1, upper = 1> phi_icept[B-1];
 real<lower = -1, upper = 1> theta_icept[B-1];

 real m_price;
 real<lower = -1, upper = 1> phi_price;
 real<lower = -1, upper = 1> theta_price;

 real m_ftdsp;
 real<lower = -1, upper = 1> phi_ftdsp;
 real<lower = -1, upper = 1> theta_ftdsp;

 // individual-specific GPs: ---------------------------------------
 vector[P] z_icept[N,B-1];
 vector[P] z_price[N];
 vector[P] z_ftdsp[N];

 // lower-level GP hyperparameters (shared across people):
 vector<lower=0>[2] hypers_icept[B-1];
 vector<lower=0>[2] hypers_price;
 vector<lower=0>[2] hypers_ftdsp;
}

transformed parameters{
 // individual-specific GPs: ---------------------------------------
 vector[P] beta_icept[N,B-1];
 vector[P] beta_price[N];
 vector[P] beta_ftdsp[N];

 // module to contain the covariance matrices:
 {
 // individual-level kernel matrices: ------------------------------
 matrix[P,P] K[B-1];
 matrix[P,P] L[B-1];
 matrix[P,P] K_price;
 matrix[P,P] L_price;
 matrix[P,P] K_ftdsp;
 matrix[P,P] L_ftdsp;

 // IN THIS SECTION: use the user defined functions to create covariance matrices,
 // then use the reparametrization of the normal distribution with the cholesky
 // decomposition of the kernel to form the mean function and function values

 // intercept kernels and function values:
 for(b in 1:(B-1)){
   K[b] = Kmat(P, hypers_icept[b,1], hypers_icept[b,2], ktype, 1e-8);
   L[b] = cholesky_decompose(K[b]);
   for(n in 1:N){
    beta_icept[n,b] = mu_icept[b] + L[b] * z_icept[n,b];
  }
 }
 K_price = Kmat(P, hypers_price[1], hypers_price[2], ktype, 1e-8);
 L_price = cholesky_decompose(K_price);
```

```
 for(n in 1:N){
  beta_price[n] = mu_price + L_price * z_price[n];
 }
 K_ftdsp = Kmat(P, hypers_ftdsp[1], hypers_ftdsp[2], ktype, 1e-8);
 L_ftdsp = cholesky_decompose(K_ftdsp);
 for(n in 1:N){
  beta_ftdsp[n] = mu_ftdsp + L_ftdsp * z_ftdsp[n];
  }
 }
}
model{
 // mean function:

 vector[P] err_icept[B-1];
 vector[P] err_price;
 vector[P] err_ftdsp;

 vector[P] nu_icept[B-1];
 vector[P] nu_price;
 vector[P] nu_ftdsp;

 for(b in 1:(B-1)){
  m_icept[b] ~ normal(0,10);
  phi_icept[b] ~ normal(0,2);
  theta_icept[b] ~ normal(0,2);
  tau_icept[b] ~ normal(0,1);

  nu_icept[b][1]=m_icept[b]+phi_icept[b]*m_icept[b];
  mu_icept[b][1] ~ normal(nu_icept[b][1], tau_icept[b]);
  err_icept[b][1]=mu_icept[b][1]-nu_icept[b][1];

  for(t in 2:P){
    nu_icept[b][t]= m_icept[b]+phi_icept[b]*mu_icept[b][t-1]+theta_icept[b]*err_icept[b][t-1];
    mu_icept[b][t] ~ normal(nu_icept[b][t], tau_icept[b]);
    err_icept[b][t] = mu_icept[b][t] - nu_icept[b][t];
  }
 }

 m_price ~ normal(0,10);
 phi_price ~ normal(0,2);
 theta_price ~ normal(0,2);
 tau_price ~ normal(0,1);
 nu_price[1]=m_price+phi_price*m_price;
 mu_price[1] ~ normal(nu_price[1], tau_price);
 err_price[1] = mu_price[1]-nu_price[1];

 for(t in 2:P){
  nu_price[t]=m_price+phi_price*mu_price[t-1]+theta_price*err_price[t-1];
  mu_price[t] ~ normal (nu_price[t], tau_price);
  err_price[t]=mu_price[t] - nu_price[t];
 }

 m_ftdsp ~ normal(0,10);
 phi_ftdsp ~ normal(0, 2);
 theta_ftdsp ~ normal(0,2);
 tau_ftdsp ~ normal(0,1);

 nu_ftdsp[1] = m_ftdsp+phi_ftdsp*m_ftdsp;
 mu_ftdsp[1] ~ normal(nu_ftdsp[1], tau_ftdsp);
 err_ftdsp[1] = mu_ftdsp[1]-nu_ftdsp[1];

 for(t in 2:P){
  nu_ftdsp[t]=m_ftdsp+phi_ftdsp*mu_ftdsp[t-1]+theta_ftdsp*err_ftdsp[t-1];
  mu_ftdsp[t] ~ normal (nu_ftdsp[t], tau_ftdsp);
```

```
 err_ftdsp[t]=mu_ftdsp[t] - nu_ftdsp[t];
}

// lower-level hyperparameters (function values)
for(b in 1:(B-1)){
 hypers_icept[b] ~ pc_prior(ktype, 5.0, 0.01, 1.0, 0.001);
}
hypers_price ~ pc_prior(ktype, 5.0, 0.01, 1.0, 0.001);
hypers_ftdsp ~ pc_prior(ktype, 5.0, 0.01, 1.0, 0.001);

// individual-specific functions (reparametrization form, don't save this)
for(i in 1:N){
 for(b in 1:(B-1)){
   z_icept[i,b] ~ normal(0,1);
 }
 z_price[i] ~ normal(0,1);
 z_ftdsp[i] ~ normal(0,1);
}

// likelihood for each choice occassion:
for(m in 1:M){
 vector[B] util;

 // compute utility for each good; first good has util = 0
 util[1] = beta_price[id[m]][pd[m]]*price[m,1] + beta_ftdsp[id[m]][pd[m]]*ftdsp[m,1];
 for(b in 2:B){
   util[b] = beta_icept[id[m],b-1][pd[m]] + beta_price[id[m]][pd[m]]*price[m,b] +
   beta_ftdsp[id[m]][pd[m]]*ftdsp[m,b];
 }
 y[m] ~ categorical_logit(util);
}
}
generated quantities{
 vector[M] log_lik;

 // compute observation-level log-likelihood for computing WAIC/LOO:
 for(m in 1:M){
    vector[B] util;

    util[1] = beta_price[id[m]][pd[m]]*price[m,1]+beta_ftdsp[id[m]][pd[m]]*ftdsp[m,1];
    for(b in 2:B){
      util[b] = beta_icept[id[m],b-1][pd[m]]+beta_price[id[m]][pd[m]]*price[m,b]+
      beta_ftdsp[id[m]][pd[m]]*ftdsp[m,b];
    }
    log_lik[m] = categorical_logit_lpmf(y[m] | util);
 }
}
```