Quantitative Marketing and Economics (2023) 21:95–146
https://doi.org/10.1007/s11129-022-09260-7

# Shrinkage priors for high-dimensional demand estimation

**Adam N. Smith**[1] · **Jim E. Griffin**[2]

Received: 24 April 2021 / Accepted: 1 November 2022 / Published online: 29 December 2022


## Abstract

Estimating demand for large assortments of differentiated goods requires the specification of a demand system that is sufficiently flexible. However, flexible models are highly parameterized so estimation requires appropriate forms of regularization to avoid overfitting. In this paper, we study the specification of Bayesian shrinkage priors for pairwise product substitution parameters. We use a log-linear demand system as a leading example. Log-linear models are parameterized by own and cross-price elasticities, and the total number of elasticities grows quadratically in the number of goods. Traditional regularized estimators shrink regression coefficients towards zero which can be at odds with many economic properties of price effects. We propose a hierarchical extension of the class of global-local priors commonly used in regression modeling to allow the direction and rate of shrinkage to depend on a product classification tree. We use both simulated data and retail scanner data to show that, in the absence of a strong signal in the data, estimates of price elasticities and demand predictions can be improved by imposing shrinkage to higher-level group elasticities rather than zero.

**Keywords** Hierarchical priors · Global-local priors · Non-sparse shrinkage · Horseshoe · Seemingly unrelated regression · Price elasticities

**JEL Classification** C11 · C13 · D12 · M31

We thank IRI for making the data available. All estimates and analysis in this paper based on data provided by IRI are by the authors and not by IRI. Accompanying R code is available at https://github.com/adam-n-smith/hierarchical-shrinkage.

✉ Adam N. Smith
a.smith@ucl.ac.uk

Jim E. Griffin
j.griffin@ucl.ac.uk

[1] UCL School of Management, University College London, London, UK

[2] Department of Statistical Science, University College London, London, UK

## 1 Introduction

Measuring price and promotional effects from store-level transaction data is a mainstay of economics and marketing research. However, many challenges arise in model specification and inference when the product space is large. For example, demand models for many goods should be flexible so as to not impose strong prior restrictions on cross-price effects, but flexible models contain many parameters so some form of dimension reduction or regularization is needed to avoid overfitting. While there is a large literature on regularized estimation, existing methods typically assume that the underlying parameter vector is sparse and so shrinkage points are fixed at zero. In this paper, we demonstrate the value of non-sparse, data-driven shrinkage in high-dimensional demand models.

Our primary contribution is to develop Bayesian shrinkage priors for pairwise product substitution parameters that may not be sparse but instead conform to a low-dimensional structure. We propose a hierarchical extension of the class of global-local priors (Polson & Scott, 2010) that is parameterized using information from a product classification tree. Typical retail scanner panel data sets, such as those provided by IRI and Nielsen, include product classification tables where goods are partitioned into broad categories (e.g., "Salty Snacks"), subcategories ("Potato Chips"), brands ("Lay's"), and ultimately UPCs ("Lay's Classic Potato Chips 20 oz."). We explicitly use this tree structure to parameterize: (i) prior means, so that substitution between two goods will depend on the nature of substitution between their respective subcategories and categories; (ii) prior variances, so that shrinkage can propagate down through the tree. We also consider different forms of shrinkage at each level of the tree—e.g., ridge (Hoerl & Kennard, 1970) vs. horseshoe (Carvalho et al., 2010)—and provide novel results on the shrinkage behavior of the induced marginal priors.

We apply the proposed hierarchical shrinkage priors to a log-linear demand system with hundreds of goods and thousands of cross-price elasticity parameters. Log-linear models remain widely used in elasticity-based pricing applications because of their simple yet flexible functional form. In practice, however, researchers typically only allow a small subset of "within category" cross-price effects to enter the demand equation for any good (e.g., Hausman et al., 1994; Montgomery, 1997; Hitsch et al., 2021; Semenova et al., 2021) or even omit cross effects altogether (e.g., DellaVigna & Gentzkow, 2019). We contribute to this literature by allowing a more complete, high-dimensional set of prices to enter each demand equation and then use the proposed shrinkage priors to do data-driven regularization on price elasticity parameters. We also propose a posterior computational strategy that partially mitigates the curse of dimensionality arising from estimating log-linear demand systems at scale.

We use both simulated and actual retail scanner data to highlight the value of hierarchical non-sparse shrinkage. In our simulation experiments, we show that the proposed priors are especially useful in settings where the true elasticity matrix is either: (i) dense, meaning all pairs of goods have a non-zero cross-price elasticity; (ii) dense but becomes sparse after subtracting off the prior mean, meaning all pairs of *product groups* have a non-zero cross-price elasticity and many product-level elasticities are exactly equal to the group-level elasticity; or (iii) group-wise sparse,

meaning many group-level cross elasticities are zero and this sparsity is inherited by product-level elasticities. In our empirical application, we use store-level data to estimate a high-dimensional log-linear demand system with 275 products that span 28 subcategories and nine categories. We show improvements in demand predictions and estimated elasticities from imposing non-sparse hierarchical shrinkage. Relative to the best-performing sparse prior, the best-performing hierarchical prior leads to a 4.5% improvement in predictions across all products, a 5.7% improvement for products with extrapolated price levels in the holdout sample, and a 6.8% improvement for products with limited price variation. In addition, we find that imposing hierarchical shrinkage leads to larger and more dispersed cross-price elasticities as well as more negative and precisely estimated own-price elasticities. We believe that being able to produce more economically reasonable elasticity estimates by simply imposing shrinkage to higher-level category elasticities is a strength of our approach. Finally, we produce competitive maps from each estimated $275 \times 275$ price elasticity matrix and show that hierarchical shrinkage priors lead to a more interpretable analysis of product competition and market structure.

### 1.1 Related literature

Methodologically, our work relates to the literature on Bayesian shrinkage priors for high-dimensional regression models (Polson & Scott, 2012a; Bhadra et al., 2019). This literature has largely focused on the problem of sparse signal recovery in which the parameter vector is assumed to be zero except for a few possibly large components. Our goal is to illustrate how many existing shrinkage technologies can still be used for detecting deviations from some non-zero mean structure. Because our prior is hierarchical in nature, it also relates to the hierarchical penalties of Bien et al. (2013) and hierarchical priors of Griffin and Brown (2017) developed to control shrinkage in linear regression models with interaction terms. These penalties/priors control the rate of shrinkage of regression effects at different levels of the hierarchy, allowing for higher-order interactions to be present only when the main effects are present. In contrast, our priors control both the *rate* and *direction* of shrinkage and have applications beyond regression models with many interactions.

We also contribute to a rich literature on shrinkage estimation of price and promotional elasticities. For example, Blattberg and George (1991), Montgomery (1997), and Boatwright et al. (1999) estimate Bayesian hierarchical models that pool information across stores to improve elasticity estimates. Rather than exploiting a hierarchical structure over stores, we impose a hierarchical structure over products in order to estimate high-dimensional regression-based demand systems. Moreover, Montgomery and Rossi (1999), Fessler and Kasy (2019), and Smith et al. (2019) demonstrate the benefits of constructing priors that shrink towards economic theory. Relative to Smith et al. (2019), who also consider group-wise restrictions on price elasticity parameters in a log-linear demand system, we explicitly focus on a high-dimensional setting in which there are more prices and cross-price effects than there are observations. For example, while Smith et al. (2019) estimate demand for up to 32 goods spanning eight subcategories and one category, we estimate demand for 275 goods spanning 28 subcategories and nine categories. Smith et al. (2019) also impose

group-wise equality restrictions on elasticity parameters, which is a very dogmatic form of shrinkage. Ensuring shrinkage points are correctly specified is a first-order concern and motivates their approach of estimating the grouping structure from the data. While this offers the ability to uncover boundaries in substitution, it comes at a severe cost of scalability and would not be feasible in our high-dimensional setting. Although the mean structure of our hierarchical prior is not explicitly derived from microeconomic theory, the idea of group-wise shrinkage is similar in spirit to the dimension reduction offered by the theory of separable preferences and multi-stage budgeting (Goldman & Uzawa, 1964; Gorman, 1971; Deaton & Muellbauer, 1980; Hausman et al., 1994).

Lastly, we contribute to a growing literature on large-scale demand estimation. One strand of literature has focused on modeling consumer choice among a high-dimensional set of substitutes from one product category. A variety of large-scale discrete choice models have been proposed and made tractable through regularization (Bajari et al., 2015), random projection (Chiong & Shum, 2018), and the use of auxiliary data on consumer consideration sets and search (Amano et al., 2019; Morozov, 2020). A second strand has focused on modeling demand for wider assortments spanning multiple categories and has exploited the flexibility of deep learning (Gabel & Timoshenko, 2021) and representation learning methods such as word embeddings (Ruiz et al., 2020) and matrix factorization (Chen et al., 2020; Donnelly et al., 2021).[1] There is also a well-established literature on economic models of multi-category demand (e.g., Manchanda et al., 1999; Song & Chintagunta, 2006; Mehta, 2007; Thomassen et al., 2017; Ershov et al., 2021). However, with the exception of Ershov et al. (2021), microfounded multi-category models become intractable at scale and have only been estimated on data with relatively small choice sets spanning a few categories.

There are a few ways in which the methods in this paper differ from, and potentially complement, existing large-scale methods. The first difference is in the data required. Our estimation framework—comprising a log-linear demand system with hierarchical shrinkage priors on price elasticity parameters—uses aggregate store-level data while most papers discussed above use household-level purchase basket data. Although granular purchase data sets are becoming more ubiquitous, many marketing researchers continue to rely on store-level data to estimate price and promotional effects (Hitsch et al., 2021). Our framework thus adds to the toolkit for anyone wanting to forecast demand and estimate large cross-price elasticity matrices with aggregate data.

The second difference is in the assumptions about functional forms and the consumer choice process. Log-linear models do not require explicit assumptions about the functional form of consumer utility and can instead be viewed as a first-order approximation to some Marshallian demand system (Diewert, 1974; Pollak & Wales,

[1]There is a closely related literature on modeling purchase data alone (Jacobs et al., 2016; Kumar et al., 2020; Jacobs et al., 2021) which uses similar methods but abstracts away from demand estimation in the sense that prices do not enter the modeling framework.

1992).[2] Log-linear models are also directly parameterized by price elasticities, which is convenient when elasticities are focal objects of interest. That said, log-linear models do not allow for inferences about preference heterogeneity, are not guaranteed to satisfy global regularly conditions, and are less scalable—at least when modeled in a joint, seemingly unrelated regression system—than many existing machine learning methods.

Finally, we note that the novel contribution of this paper is not in the specification of demand, but instead in the approach to regularize high-dimensional demand parameters within a given functional form. We see at least two ways in which this paper can complement existing methods. First, estimating substitution patterns requires sufficient variation in key marketing variables such as price, which can be challenging in many retail settings. One common solution in empirical work is to exclude goods from the analysis whose prices either do not vary or perfectly covary with other prices (e.g., Donnelly et al., 2021). In contrast, our hierarchical shrinkage approach pools information across goods to produce reasonable estimates of cross-price elasticities for all products, even those with little to no price variation. Second, while we use log-linear demand as a leading case, we believe that hierarchical shrinkage priors can facilitate the estimation of many other demand systems at scale. Examples include models based on quadratic utility functions (Wales & Woodland, 1983; Thomassen et al., 2017) and discrete choice over product bundles (Gentzkow, 2007; Song & Chintagunta, 2006; Ershov et al., 2021), both of which contain sets of pairwise substitution parameters that grow quadratically in the number of goods.

The remainder of this paper is organized as follows. Section 2 defines the log-linear demand system and outlines existing approaches for imposing sparse shrinkage, including global-local priors. Section 3 develops hierarchical global-local priors. Section 4 discusses posterior computation strategies. Section 5 presents results from a set of simulation experiments and Section 6 presents results of an empirical application to store-level scanner data. Section 7 concludes.

## 2 Background: Regularizing high-dimensional demand

In this paper we propose a non-sparse, hierarchical shrinkage approach to improve the estimation of high-dimensional demand models. Our priors apply to demand models characterized by a set of pairwise product substitution parameters whose dimension grows non-linearly in the number of goods. Throughout the paper, we use log-linear demand as a leading example. Before introducing our shrinkage framework, we first outline the specification of a log-linear demand system and standard existing approaches for regularization.

[2]Many existing large-scale methods still rely on assumptions of discrete choice. For example, Donnelly et al. (2021) and Gabel and Timoshenko (2021) model purchases as a two-stage process: first category incidence and then discrete choice among products within each of the chosen categories. Ruiz et al. (2020) propose a behavioral model of sequential discrete choice: a consumer decides which item to add to their basket (or to stop shopping) conditional on previously chosen items.

### 2.1 Demand specification

Consider an assortment of products indexed by $i = 1, \ldots, p$. For each product, we observe its unit sales $q_{it}$ and price $p_{it}$ across weeks $t = 1, \ldots, n$. In a log-linear demand system, the log of unit sales of product $i$ at time $t$ is regressed on its own log price, the log prices of all other products, and a vector of controls which can include product intercepts, seasonal trends, and promotional activity:

$$\log q_{it} = \beta_{ii} \log p_{it} + \sum_{j \neq i} \beta_{ij} \log p_{jt} + c_{it}' \phi_i + \varepsilon_{it}. \tag{1}$$

The $p \times 1$ error vector $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{pt})$ is assumed to follow a $\mathrm{N}(0, \Sigma)$ distribution where the covariance matrix $\Sigma$ captures any unobserved, contemporaneous correlations between goods.

Log-linear demand systems are attractive for a few reasons. First, the price coefficients represent own and cross-price elasticities, which are often focal economic objects in the analysis of market structure and pricing/promotion schedules. Second, the functional form is simple and admits tractable estimators of all model parameters. Third, the model is flexible in that there are no restrictions imposed on $\beta_{ij}$'s so we can accommodate both substitutes ($\beta_{ij} > 0$) and complements ($\beta_{ij} < 0$). Log-linear models are also flexible in the sense that they possess enough parameters to serve as a first-order approximation to a valid Marshallian demand system (Diewert, 1974; Pollak & Wales, 1992). However, because there are $p^2$ elasticities in a system with $p$ goods, regularization is needed at scale.

### 2.2 Sparse shrinkage and global-local priors

We now outline a standard sparse approach to regularizing parameters in the log-linear demand system in Eq. 1 which will both lay the groundwork for, and also serve as a benchmark against, the hierarchical priors we construct in the following section. We take a Bayesian approach to estimation and so regularization will be imparted by way of the prior. There are now many prior specifications that can be used for elasticity parameters $\beta_{ij}$, each having different tail behavior and thus inducing different forms of shrinkage (see, e.g., Bhadra et al., 2019). We specifically focus on global-local priors (Polson & Scott, 2010), which are scale mixtures of normals:

$$\begin{aligned} \beta_{ij} | \lambda_{ij}^2, \tau^2 &\sim \mathrm{N}(0, \tau^2 \lambda_{ij}^2), \\ \lambda_{ij}^2 &\sim G. \end{aligned} \tag{2}$$

Here $\tau^2$ is the "global" variance that controls the overall amount of shrinkage across all elasticity parameters $\beta_{ij}$ while $\lambda_{ij}^2$ is a "local" variance that allows for component-wise deviations from the shrinkage imposed by $\tau^2$. The local variances are distributed

according to a mixing distribution $G$, which is often assumed to be absolutely continuous and will thus admit an associated density $g(\cdot)$.[3]

One of the reasons that global-local priors are attractive is that, given an appropriate choice of mixing density, they can exhibit two useful properties for sparse signal recovery: (i) a spike of mass at zero to encourage sparsity for small noisy observations; and (ii) heavy tails to leave large signals unshrunk (Polson & Scott, 2010). Together, (i) and (ii) induce shrinkage behavior that mimics the hard selection rules of Bayesian two-group, spike-and-slab priors (Mitchell & Beauchamp, 1988; George & McCulloch, 1993; Ishwaran & Rao, 2005) while admitting much more efficient posterior computation strategies.

The global-local structure is also very general and nests many common Bayesian shrinkage priors. For example, ridge regression (Hoerl & Kennard, 1970) arises when $\lambda_{ij}^2 = 1$, the Bayesian lasso (Park & Casella, 2008; Hans, 2009) arises when $\lambda_{ij}^2$ follows an exponential distribution, and the horseshoe (Carvalho et al., 2010) arises when $\lambda_{ij}$ follows a half-Cauchy distribution. The priors can be compared using their shrinkage profiles, which measure the amount by which the posterior mean shrinks the least squares estimate of a regression coefficient to zero, and is related to the shape of the mixing density $g(\cdot)$. For example, the tails of an exponential mixing density are lighter than the polynomial tails of the half-Cauchy, suggesting that the Bayesian lasso may tend to over-shrink large regression coefficients and under-shrink small ones relative to the horseshoe (Polson & Scott, 2010; Datta & Ghosh, 2013; Bhadra et al., 2016).

Although the tail behavior and associated shrinkage properties of global-local priors have now been studied extensively, existing work typically considers the problem of sparse signal recovery and so the prior means in Eq. 2 are assumed to be fixed at zero. However, economic theory offers many reasons for why sparse shrinkage may not be appropriate for price elasticity parameters. First, own-price effects should be negative by the law of demand and cross-price effects in a Marshallian demand system need not be zero even if true substitution effects are zero. Second, the property of Cournot aggregation (or "adding-up") implies that demand must become more elastic as the set of available substitutes increases. If cross-effects are arbitrarily shrunk towards zero, then the magnitude of the own effects must fall, which would lead to estimates of own-price effects that are biased downward in magnitude. Third, line pricing can produce highly correlated prices among substitutable goods, which would lead to heavy shrinkage on the associated cross-price effects in estimation. However, fixing the shrinkage points to zero would suggest that varieties within the product line are unrelated even though the very nature of price correlation reflects a high degree of substitution. Together, these examples suggest that care must be taken when regularizing price effect parameters, which motivates our development of non-sparse shrinkage priors.

[3]Note that some of the earliest examples of Bayesian shrinkage arise under the model in Eq. 2 with $G$ being discrete. For example, spike-and-slab priors (Mitchell & Beauchamp, 1988; George & McCulloch, 1993) arise under a two-point mixture distribution $G(\lambda^2) = w\delta_{\lambda^2=1} + (1-w)\delta_{\lambda^2=0}$.

## 3 Hierarchical global-local priors

In this section, we develop hierarchical global-local priors with two goals in mind. The first is to allow own and cross-price elasticities to be shrunk towards higher-level group elasticities (rather than zero), where the hierarchy of groups is defined using a pre-specified product classification tree. The second is to allow shrinkage to propagate down through the tree. To this end, we extend the standard global-local specification by adding a hierarchical mean structure to connect shrinkage points across levels in the tree, as well as a hierarchical product structure on the local variances to connect the degree of shrinkage across levels in the tree.

### 3.1 Notation

We first define a minimum of notation. We assume that the researcher has access to an $L$-level product classification tree where levels are indexed by $\ell = 0, 1, \ldots, L-1$ and are ordered such that $\ell = 0$ is the lowest level corresponding to products, $\ell = 1$ corresponds to most granular definition of product groups, and $\ell = L-1$ corresponds to the highest level and most coarse definition of product groups. Further define $n_\ell$ to be the number of nodes (groups) on level $\ell$ where $n_0 > n_1 > \cdots > n_{L-1}$.

The indexing notation we introduce is slightly more complicated than the usual indexing for graphical models because, in our case, the target parameters are always defined with respect to two nodes (e.g., the price elasticity between two products), not one. So instead of considering the parent of node $i$, we must consider the parent of $(i, j)$, which itself will be a pair of nodes. We now define the indexing function that will be used throughout.

**Definition 1** Let $(i, j)$ denote a pair of nodes at the lowest level (level 0) of the classification tree. Then let $\pi(ij|\ell)$ denote the parent index function which produces the level-$\ell$ parent nodes of $(i, j)$.

To build intuition, consider the following two products: (i) Lay's potato chips, which belongs to the "Potato Chip" subcategory (level 1) and "Salty Snack" category (level 2); and (ii) Heineken beer, which belongs to the "Imported Beer" subcategory (level 1) and "Beer" category (level 2). Then the parents of (Lay's, Heineken) on levels 1 and 2 are as follows.

$$\begin{aligned}\pi(\text{Lay's, Heineken}|1) &= (\text{Potato Chips, Imported Beer})\\ \pi(\text{Lay's, Heineken}|2) &= (\text{Salty Snacks, Beer})\end{aligned}$$

Given a product-level elasticity $\beta_{ij}$, we let $\theta_{\pi(ij|m)}$ represent the relationship between the level-$m$ ancestors of $i$ and $j$. An example of this lineage of parameters is shown in Fig. 1. The darkest square in the left-hand-side grid denotes $\beta_{ij}$. The level-1 parent of $(i, j)$ is $\pi(ij|1) = (4, 2)$ and the level-2 parent of $(i, j)$ is $\pi(ij|2) = (2, 1)$. The idea is to then direct shrinkage of $\beta_{ij}$ towards $\theta_{\pi(ij|1)}$ (grid in the middle), which is in turn shrunk towards $\theta_{\pi(ij|2)}$ (grid on the right).

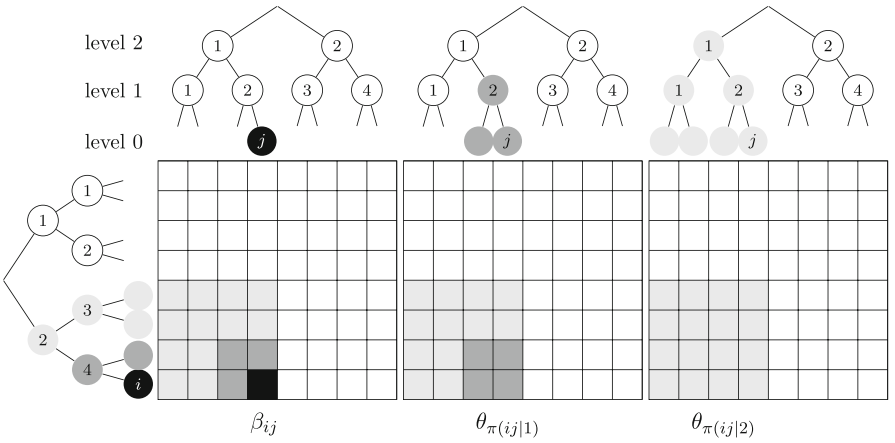



**Fig. 1** Visualization of hierarchical shrinkage using a three-level classification tree. The focal product pair is $(i, j)$ which has level-1 parent $\pi(ij|1) = (4, 2)$ and level-2 parent $\pi(ij|2) = (2, 1)$. The parameter $\beta_{ij}$ (darkest shaded square in the left grid) will be shrunk towards the level-1 parent $\theta_{\pi(ij|1)}$ (darkest shaded region of 4 squares in the middle grid), which will in turn be shrunk towards the level-2 parent $\theta_{\pi(ij|2)}$ (shaded region of 16 squares in the right grid)

### 3.2 Prior construction

Using the indexing notation above, we now introduce a sequence of prior distributions on the elasticity parameters of the log-linear demand model in Eq. 1. Starting at the lowest level of the tree, define the prior on the $(i, j)$ product pair as:

$$\beta_{ij} \sim \mathrm{N}\left(\theta_{\pi(ij|1)}, \tau_\beta^2 \Psi_{ij}\right) \tag{3}$$

which is a global-local prior with global variance $\tau_\beta^2$, local variance $\Psi_{ij}$, and prior mean $\theta_{\pi(ij|1)}$. Note that this notation holds for any $(i, j)$ pair including the own elasticities where $i = j$. However, in order to account for differences in the expected signs of own and cross-price elasticities, we will ultimately build up two separate hierarchical structures: one for the $\beta_{ii}$'s and one for the $\beta_{ij}$'s. The former will shrink the product-level own elasticities towards higher-level *own-price elasticities* whereas the latter will shrink product-level cross elasticities towards higher-level *cross-price elasticities*. For ease of exposition, we will focus our prior construction on the case where $i \neq j$ and the $\beta_{ij}$'s are cross elasticities.

**Hierarchy of means** We specify global-local priors on the higher-level group elasticities:

$$\theta_{\pi(ij|\ell)} \sim \mathrm{N}\left(\theta_{\pi(ij|\ell+1)}, \tau_\ell^2 \Psi_{\pi(ij|\ell)}\right), \qquad \ell = 1, \ldots, L-1, \tag{4}$$

where $\tau_\ell^2$ is the global variance across all level-$\ell$ elasticities $\theta_{\pi(ij|\ell)}$, $\Psi_{\pi(ij|\ell)}$ is a local variance, and $\theta_{\pi(ij|\ell+1)}$ is the parent cross-group elasticity to $\theta_{\pi(ij|\ell)}$. This hierarchical mean structure allows the direction of shrinkage to be governed by the classification tree. That is, in the absence of a strong signal in the data, elasticities will be shrunk towards higher-level elasticity parameters instead of zero.

**Hierarchy of variances** We also impose a hierarchical structure on the local variance parameters. At the product level, we have:

$$\Psi_{ij} = \lambda_{ij}^2 \prod_{k=\ell+1}^{L-1} \lambda_{\pi(ij|k)}^2, \quad \lambda_{ij}^2 \sim G_\beta, \tag{5}$$

and at higher levels in the tree, we have:

$$\Psi_{\pi(ij|\ell)} = \lambda_{\pi(ij|\ell)}^2 \prod_{k=\ell+1}^{L-1} \lambda_{\pi(ij|k)}^2, \quad \lambda_{\pi(ij|\ell)}^2 \sim G_\ell, \quad \ell = 1, \ldots, L-1. \tag{6}$$

Here $\lambda_{ij}^2$ and $\lambda_{\pi(ij|\ell)}^2$ are variances associated with $\beta_{ij}$ and $\theta_{\pi(ij|\ell)}$, respectively, and represent local variances absent any hierarchical structure connecting variances across levels. That is, without a hierarchy of variances we would simply have $\Psi_{ij} = \lambda_{ij}^2$ and $\Psi_{\pi(ij|\ell)} = \lambda_{\pi(ij|\ell)}^2$. With the product hierarchy of variances, the induced local variances $\Psi$ will be small whenever either $\lambda_{\pi(ij|\ell)}^2$ is small or any $\lambda_{\pi(ij|s)}^2$ is small for $s > \ell$ (i.e., higher levels in the tree), which allows shrinkage to propagate down through the tree. Taken together with the hierarchical mean structure, the hierarchal variance structure implies that, in the absence of a strong signal in the data, price elasticities will be strongly shrunk towards higher-level elasticities and these higher-level group elasticities will be strongly shrunk towards each other.

**Examples** A summary of model parameters is provided in Table 1. Below we provide examples of the complete hierarchical prior specification for trees different numbers of levels. A single-level tree corresponds to a prior for $\beta_{ij}$ that has a fixed mean $\bar{\theta}$ and does not encode any information about product groups. If the tree has at least two levels, then $\beta_{ij}$ will be shrunk towards its parent elasticity and shrinkage will be allowed to propagate. Note that the level at which shrinkage begins to propagate is

**Table 1** Summary of notation

| Parameter | Level | Description |
|---|---|---|
| Elasticities | | |
| $\beta_{ij}$ | $\ell = 0$ | Price elasticity of demand for product $i$ with respect to price of $j$ |
| $\theta_{\pi(ij|\ell)}$ | $\ell \geq 1$ | Level-$\ell$ ancestor elasticity of the $(i, j)$ product pair |
| Local Variances | | |
| $\lambda_{ij}^2$ | $\ell = 0$ | Local variance for $\beta_{ij}$ |
| $\lambda_{\pi(ij|\ell)}^2$ | $\ell \geq 1$ | Local variance for $\theta_{\pi(ij|\ell)}$ |
| $\Psi_{ij}$ | $\ell = 0$ | Product of all higher-level local variances including $\lambda_{ij}^2$ |
| $\Psi_{\pi(ij|\ell)}$ | $\ell \geq 1$ | Product of all higher-level local variances including $\lambda_{\pi(ij|\ell)}^2$ |
| Global Variances | | |
| $\tau_\beta^2$ | $\ell = 0$ | Global variance across all product elasticities $\beta_{ij}$ |
| $\tau_\ell^2$ | $\ell \geq 1$ | Global variance across all level-$\ell$ elasticities $\theta_{\pi(ij|\ell)}$ |

also a choice of the researcher. In the examples below, shrinkage starts to propagate at the top level ($\ell = L - 1$) of the tree.

(a) One-level prior

$$\beta_{ij} \sim \mathrm{N}\left(\bar{\theta}, \tau_\beta^2 \Psi_{ij}\right), \quad \Psi_{ij} = \lambda_{ij}^2$$

(b) Two-level prior

$$\begin{aligned} \beta_{ij} &\sim \mathrm{N}\left(\theta_{\pi(ij|1)}, \tau_\beta^2 \Psi_{ij}\right), & \Psi_{ij} &= \lambda_{ij}^2 \lambda_{\pi(ij|1)}^2 \\ \theta_{\pi(ij|1)} &\sim \mathrm{N}\left(\bar{\theta}, \tau_1^2 \Psi_{\pi(ij|1)}\right), & \Psi_{\pi(ij|1)} &= \lambda_{\pi(ij|1)}^2 \end{aligned}$$

(c) Three-level prior

$$\begin{aligned} \beta_{ij} &\sim \mathrm{N}\left(\theta_{\pi(ij|1)}, \tau_\beta^2 \Psi_{ij}\right), & \Psi_{ij} &= \lambda_{ij}^2 \lambda_{\pi(ij|1)}^2 \lambda_{\pi(ij|2)}^2 \\ \theta_{\pi(ij|1)} &\sim \mathrm{N}\left(\theta_{\pi(ij|2)}, \tau_1^2 \Psi_{\pi(ij|1)}\right), & \Psi_{\pi(ij|1)} &= \lambda_{\pi(ij|1)}^2 \lambda_{\pi(ij|2)}^2 \\ \theta_{\pi(ij|2)} &\sim \mathrm{N}\left(\bar{\theta}, \tau_2^2 \Psi_{\pi(ij|2)}\right), & \Psi_{\pi(ij|2)} &= \lambda_{\pi(ij|2)}^2 \end{aligned}$$

### 3.3 Choice of mixing densities

The hierarchical priors outlined above endow each product-level elasticity and group-level elasticity with a local variance $\lambda_{\pi(ij|\ell)}^2$ which is distributed according to a level-specific mixing distribution $G_\ell$. As discussed in Section 2.2, the tails of the associated densities $g_\ell(\cdot)$ will play a key role in shaping the shrinkage imposed by the prior. Although there are now many possible choices for $g_\ell(\cdot)$, we confine our attention to three forms of shrinkage: (i) ridge, where the mixing density is a degenerate distribution and local variances are fixed to one; (ii) lasso, with an exponential mixing density; and (iii) horseshoe, with a half-Cauchy mixing density. These three types of priors remain common choices in empirical work and have very different shrinkage profiles (Bhadra et al., 2019).

Our hierarchical global-local priors also require the specification of a mixing density for each level of the tree. We can therefore conceive of using different forms of shrinkage across different levels (e.g., ridge at the subcategory level and horseshoe at the product level). However, this also raises questions of whether properties of $g_\ell(\cdot)$ at higher levels in the tree affect shrinkage behavior at the product level, which we address in the following section.

### 3.4 Some theory on shrinkage properties

The typical strategy for characterizing shrinkage in global-local models is to examine the shape—specifically, the tails—of the marginal prior for regression coefficients $\beta_{ij}$. Because global-local priors are scale mixtures of normals, the heaviness of the tails of this marginal prior will be determined by the tails of the mixing density (Barndorff-Nielsen et al., 1982). However, in our setting this analysis is complicated by the fact that the marginal prior for $\beta_{ij}$ will depend on multiple mixing densities in the hierarchical global-local structure.

In this section, we provide three results to characterize shrinkage properties of hierarchical global-local priors. First, we show that the marginal prior for $\beta_{ij}$ still retains a scale mixtures of normals representation and so the mixing densities will continue to play a key role in shaping the shrinkage profile for $\beta_{ij}$. Second, we show that if the same heavy-tailed mixing density is specified at each level of the tree, then its heaviness will be preserved under the hierarchical product structure that we impose on local variances. Finally, we show that even if a combination of heavy and light-tailed mixing densities are specified across different levels, the heavy-tailed mixing densities will ultimately dominate and shape the product-level shrinkage profile.

We focus on two specific forms of shrinkage. The first is the shrinkage of $\beta_{ij}$ to $\theta_{\pi(ij|m)}$, which is the shrinkage of the product-level elasticity to its level-$m$ parent elasticity. This type of "vertical" shrinkage allows us to assess how quickly product-level elasticities can be pulled towards higher-level elasticities. Here we can write $\beta_{ij}$ as a function of its parent mean and error term:

$$\beta_{ij} = \theta_{\pi(ij|1)} + \xi_{ij}, \qquad \xi_{ij} \sim \mathrm{N}\left(0, \tau_\beta^2 \Psi_{ij}\right) \tag{7}$$

where, again, for any $\ell = 1, \ldots, L-1$:

$$\theta_{\pi(ij|\ell)} = \theta_{\pi(ij|\ell+1)} + \xi_{\pi(ij|\ell)}, \qquad \xi_{\pi(ij|\ell)} \sim \mathrm{N}\left(0, \tau_\ell^2 \Psi_{\pi(ij|\ell)}\right). \tag{8}$$

We can therefore write $\beta_{ij}$ as a function of its level-$m$ parent elasticity and a sum of errors across levels:

$$\beta_{ij} = \theta_{\pi(ij|m)} + \sum_{\ell=1}^{m-1} \xi_{\pi(ij|\ell)} + \xi_{ij}, \tag{9}$$

which yields the prior distribution:

$$\beta_{ij} - \theta_{\pi(ij|m)} \sim \mathrm{N}\left(0, \sum_{\ell=1}^{m-1} \tau_\ell^2 \Psi_{\pi(ij|\ell)} + \tau_\beta^2 \Psi_{ij}\right). \tag{10}$$

This prior is a scale mixture of normals where the scales are sums over each level's local variance $\Psi_{\pi(ij|\ell)}$, which is in turn a product of $\lambda^2_{\pi(ij|\ell)}, \ldots, \lambda^2_{\pi(ij|L-1)}$. Thus, the local variances will be a sum of products and if there is a value of $s$ (for $s \leq m$) for which $\lambda^2_{\pi(ij|s)}$ is "small" then $\beta_{ij}$ will tend to be very close to $\theta_{\pi(ij|m)}$.

We also consider the shrinkage of $\beta_{ij}$ to $\beta_{i'j'}$ (for $i \neq j, i' \neq j'$, and $i \neq i'$), which is the shrinkage between two product-level elasticities. This type of "horizontal" shrinkage allows us to assess the extent to which elasticities become more similar as they become closer in the tree. Formally define $m^\star = \min\{m : \pi(ij|m) = \pi(i'j'|m)\}$ to be the lowest level in the tree such that all four products $(i, j, i', j')$ share a common ancestor (i.e., belong to the same group at some level in the tree), where $m^\star = L$ if no common parent node exists for within the tree. Then we can write

$$\beta_{ij} - \beta_{i'j'} = \sum_{s=\ell}^{m^\star - 1} \left(\xi_{\pi(ij|s)} - \xi_{\pi(i'j'|s)}\right) \tag{11}$$

so the prior distribution of the difference is

$$\beta_{ij} - \beta_{i'j'} \sim \mathrm{N}\left(0, \sum_{s=\ell}^{m^\star - 1} \left(\Psi_{\pi(ij|s)} + \Psi_{\pi(i'j'|s)}\right)\tau_s^2\right). \tag{12}$$

Like the marginal distribution of the cross elasticities, the marginal prior distribution of the difference will be a scale mixtures of normals, where the scales are sums of products of $\lambda^2_{\pi(ij|s)}$ and $\lambda^2_{\pi(i'j'|s)}$. The number of terms in the sum is $m^\star - 1 - \ell$ and so the variance of differences will be tend to be larger if $m^\star$ is larger (i.e., the products are less similar). The form of the variances in Eq. 12 implies that if $\Psi_{\pi(ij|m)}$ is "small" on level $m$ then $\Psi_{\pi(ij|s)}$ and $\Psi_{\pi(i'j'|s)}$ will tend to be small for $s < m$ and so the variance of the difference will tend to be smaller further down the tree. This allows shrinkage to propagate down the tree with subsequent sub-categorizations of products tending to have similar cross-elasticities.

The results above show that the priors on $\beta_{ij}$ and the differences $(\beta_{ij} - \beta_{i'j'})$ can be expressed as normal scale mixtures and so, like in sparse signal detection settings, the shape of the marginal prior will again be determined by the mixing density (Barndorff-Nielsen et al., 1982). However, while there is only one mixing density in traditional regression priors, the marginal priors for $\beta_{ij}$ and $(\beta_{ij} - \beta_{i'j'})$ involve a "scaled sum of products" transformation over many mixing densities. It is therefore not clear whether the heaviness of the mixing density specified level $\ell$ is: (i) preserved under the scaled sum of products transformations; or (ii) tarnished by mixing densities with lighter tails at higher levels in the tree. We clarify both points in the following two propositions.

**Definition 2** The random variable $\zeta$ has an $L$-sum of scaled products distribution if it can be written as $\zeta = \sum_{\ell=1}^{L} \tau_\ell^2 \prod_{s=1}^{\ell} \lambda_s^2$ with fixed $\tau_1^2, \ldots, \tau_L^2$.

**Proposition 1** *Suppose $\zeta$ has an $L$-sum of scaled products distribution where $\lambda_\ell^2 \overset{iid}{\sim} G$ for $\ell = 1, \ldots, L$ and $G$ is a regularly varying distribution with index $\alpha$. Then $\zeta$ is regularly varying with index $\alpha$.*

*Proof* Because the $\lambda_s^2$'s are all independent and regularly varying with index $\alpha$, then $\Psi_\ell = \prod_{s=1}^{\ell} \lambda_s^2$ is also regularly varying with index $\alpha$ (Cline, 1987). Then the closure property of regularly varying functions guarantees that $\zeta = \sum_{\ell=1}^{L} \tau_\ell^2 \prod_{s=1}^{\ell} \lambda_s^2$ is also regularly varying with index $\alpha$. □

This first result shows that the heaviness of the mixing density at level $\ell$ is preserved under the scaled sum of products transformation. For example, if every $\lambda_s$ has a half-Cauchy prior then each $\lambda_s^2$ is an inverted-beta random variable with density $g(\lambda^2) \propto (\lambda^2)^{-1/2}(1 + \lambda^2)^{-1}$ (Polson & Scott, 2012b), which is regularly varying with index -3/2 and so $\lambda_s^2$ is regularly varying with index $1/2$ (Bingham et al., 1987). Then by Proposition 1, the sum of products would also have regularly varying tails, and the different forms of shrinkage in Eqs. 10 and 12 will all have tails of the same heaviness as a standard horseshoe prior.

**Proposition 2** *Suppose $\zeta$ has an $L$-sum of scaled products distribution with $\lambda_\ell^2 \sim G_\ell$ for $\ell = 1, \ldots, L$ and let $RV = \{\ell : G_\ell$ is regularly varying with index $\alpha\}$ be non-empty. If there exists an $\epsilon$ such that $G_s$ has a finite $\alpha + \epsilon$ moment for all $s \notin RV$ then $\zeta$ is regularly varying with index $\alpha$.*

*Proof* If at level $\ell$ at least one of $\lambda_1^2, \ldots, \lambda_\ell^2$ is regularly varying with index $\alpha$ then $\Psi_\ell = \prod_{s=1}^{\ell} \lambda_s^2$ is regularly varying with index $\alpha$. If none of $\lambda_1^2, \ldots, \lambda_\ell^2$ are regularly varying, then by assumption $\Psi_\ell$ has a finite $\alpha + \epsilon$ moment. Since, by assumption, at least one of $\lambda_1^2, \ldots, \lambda_L^2$ is regularly varying then at least one $\Psi_1, \ldots, \Psi_L$ is regularly varying while the others are guaranteed to have a finite $\alpha + \epsilon$ moment. Therefore, the closure properties of regularly varying random variables implies that $\zeta$ is also regularly varying with index $\alpha$ (Bingham et al., 1987). □

Proposition 2 shows that the sum of products has regular variation if at least one element is regularly-varying. This suggests that sparsity shrinkage of the cross-elasticities at the lowest level of the tree does not require sparsity shrinkage at all levels. For example, we know by Proposition 1 that if $\lambda_{ij}^2, \lambda_{\pi(ij|1)}^2, \ldots, \lambda_{\pi(ij|L-1)}^2$ are all regularly varying with index $\alpha$, then their product is also regularly varying with index $\alpha$. Now by Proposition 2, we know that even if $\lambda_{\pi(ij|s)}^2$ is not regularly varying for $s > \ell$ but has a finite $\alpha + \epsilon$ moment for some $\epsilon > 0$, then the product is again regularly varying. Examples of non-regularly varying distributions with a finite $\alpha + \epsilon$ moment are a degenerate distribution (i.e., ridge shrinkage with $\lambda^2 = 1$) and an exponential distribution (i.e., lasso shrinkage). Therefore, heavy tails at *any* level of the tree are all that is required to get sparsity shrinkage at for the product-level elasticities. We explore different combinations of shrinkage in the simulations and empirical applications below.

## 4 Posterior computation

Given the log-linear demand system defined in Section 2.1 and hierarchical priors outlined in Section 3, we now turn to our posterior sampling strategy. Note that the presence of product-specific control variables and general error structure in Eq. 1 leads to a seemingly unrelated regression (SUR) model:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} X & 0 & \cdots & 0 \\ 0 & X & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_p \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} \tag{13}$$

where $y_i$ is the $n \times 1$ vector of log sales for product $i$, $X$ is the $n \times p$ matrix of log prices, $\beta_i$ is the $p \times 1$ vector of own and cross-price elasticities associated with product $i$, and $C_i$ is a $n \times d$ matrix of control variables with coefficients $\phi_i$. In vector form, we have

$$y = \mathrm{X}\beta + \mathrm{C}\phi + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \Sigma \otimes I_n) \tag{14}$$

where $y = (y_1', y_2', \ldots, y_p')'$, $\mathrm{X} = \mathrm{diag}(X, X, \ldots, X)$, $\beta = (\beta_1', \beta_2', \ldots, \beta_p')'$, $\mathrm{C} = \mathrm{diag}(C_1, C_2, \ldots, C_p)$, $\phi = (\phi_1', \phi_2', \ldots, \phi_p')'$ and $\varepsilon = (\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_p')'$.

Sampling from the posterior of a SUR model first requires transforming the model in Eq. 14 into one with homogeneous unit errors. Let $U$ denote the upper triangular Cholesky root of $\Sigma$ and define $\tilde{y} = (U^{-1} \otimes I_n)y$, $\tilde{\mathrm{X}} = (U^{-1} \otimes I_n)\mathrm{X}$, and $\tilde{\mathrm{C}} = (U^{-1} \otimes I_n)\mathrm{C}$. Then the following transformed system represents a standard normal linear regression model.

$$\tilde{y} = \tilde{\mathrm{X}}\beta + \tilde{\mathrm{C}}\phi + \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \sim \mathrm{N}(0, I_{np}). \tag{15}$$

The full set of model parameters includes the elasticities $\beta$, the coefficients on controls $\phi$, the error variance $\Sigma$, and the set of hierarchical prior parameters $\Omega = \left(\{\theta_{\pi(ij|\ell)}\}, \{\lambda^2_{\pi(ij|\ell)}\}, \{\tau^2_\ell\}\right)$.

**Priors** The priors for $\beta$ and all hierarchical hyperparameters are given in Section 3.2. We write the $p^2 \times p^2$ prior covariance matrix for $\beta$ as $\Lambda_* = \tau^2_\beta \mathrm{diag}(\mathrm{vec}(\Lambda))$, where $\Lambda$ is a $p \times p$ matrix of local variances $\Psi_{ij}$ as defined in Eq. 5. Note that for a standard global-local prior, the $(i, j)$th element of $\Lambda$ would be $\lambda^2_{ij}$. We place $\mathrm{N}(\bar{\phi}, A^{-1}_\phi)$ priors on the control variable coefficients, which are conditionally conjugate to the normal likelihood given $\Sigma$. Inverse Wishart priors are commonly used for covariance matrices in Bayesian SUR models, however if $p > n$ then $\Sigma$ will be rank deficient. One approach would be to also regularize $\Sigma$ (Li et al., 2019; Li et al., 2021). We instead impose a diagonal restriction $\Sigma = \mathrm{diag}(\sigma^2_1, \ldots, \sigma^2_p)$ and place independent $\mathrm{IG}(a, b)$ priors on each $\sigma^2_j$.

**Full Conditionals** We construct a Gibbs sampler that cycles between the following full conditional distributions.

$$\Omega|\beta, \text{data} \tag{16}$$
$$\Sigma|\beta, \phi, \Omega, \text{data} \tag{17}$$
$$\beta, \phi|\Sigma, \beta, \Omega, \text{data} \tag{18}$$

The first full conditional represents the posterior of all global/local variances and higher-level group elasticities. Sampling from these distributions is computationally inexpensive. The elements of $\theta$ each have independent normal posteriors conditional on all variance parameters. Both local variances $\lambda^2$ and global variances $\tau^2$ can also be sampled independently, but the form of their respective posteriors will depend on the choice of prior. Under ridge shrinkage, each $\lambda^2 = 1$ so no posterior sampling is necessary. Under lasso shrinkage, each $\lambda^2$ follows an independent exponential distribution and so the full conditionals of $1/\lambda^2$ have independent inverse Gaussian distributions (Park & Casella, 2008). Under horseshoe shrinkage, each $\lambda$ follows an independent half-Cauchy distribution. We follow Makalic and Schmidt (2015) and represent the half-Cauchy as a scale mixture of inverse gammas, which is conjugate to the normal density so the target full conditional can be sampled from directly. Details are provided in Appendix A.1.

The second full conditional represents the posterior of the observational error covariance matrix. Assuming $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with independent $\text{IG}(a, b)$ priors yields the following posterior:

$$\sigma_j^2 | \beta, \phi, \Omega, \text{data} \sim \text{IG}\Big(a + n/2, b + (y_j - X\beta_j - C_j\phi_j)'(y_j - X\beta_j - C_j\phi_j)/2\Big) \quad (19)$$

where the $j$ subscript denotes all elements of the given vector or matrix associated with product $j$. Note that the case of $p > n$ calls for a judicious choice of $(a, b)$ given that diffuse priors will yield barely proper posteriors. If $n > p$ and $\Sigma$ is unrestricted, typical inverse Wishart priors can be used.

The last full conditional represents the joint posterior of the regression coefficients $(\beta, \phi)$. One approach to sampling from the joint posterior is to iterate between each full conditional. For example, the posterior of $\beta$ conditional on $\phi$ is:

$$\beta | \phi, \Sigma, \Omega, \text{data} \sim \text{N}\Big((\tilde{\text{X}}'\tilde{\text{X}} + \Lambda_*^{-1})^{-1}(\tilde{\text{X}}'\tilde{y}_\phi^* + \Lambda_*^{-1}\bar{\beta}(\theta)), (\tilde{\text{X}}'\tilde{\text{X}} + \Lambda_*^{-1})^{-1}\Big) \quad (20)$$

where $\tilde{y}_\phi^* = \tilde{y} - \tilde{\text{C}}\phi$. Similarly, the posterior of $\phi$ conditional on $\beta$ is:

$$\phi | \beta, \Sigma, \Omega, \text{data} \sim \text{N}\Big(\big(\tilde{\text{C}}'\tilde{\text{C}} + A_\phi\big)^{-1}\big(\tilde{\text{C}}'\tilde{y}_\beta^* + A_\phi\bar{\phi}\big), \big(\tilde{\text{C}}'\tilde{\text{C}} + A_\phi\big)^{-1}\Big) \quad (21)$$

where $\tilde{y}_\beta^* = \tilde{y} - \tilde{\text{X}}\beta$. However, note that these two Gibbs steps can be improved through blocking. For example, $\beta$ can be integrated out of the conditional posterior of $\phi$:

$$\phi | \Sigma, \Omega, \text{data} \sim \text{N}\Big(\big(\tilde{\text{C}}'\text{P}\tilde{\text{C}} + A_\phi\big)^{-1}\big(\tilde{\text{C}}'\text{P}\tilde{y} + A_\phi\bar{\phi}\big), \big(\tilde{\text{C}}'\text{P}\tilde{\text{C}} + A_\phi\big)^{-1}\Big) \quad (22)$$

where $\text{P} = I_{np} - \tilde{\text{X}}(\tilde{\text{X}}'\tilde{\text{X}} + \Lambda_*^{-1})^{-1}\tilde{\text{X}}'$ is an orthogonal projection matrix. Marginalizing over $\beta$ will yield improvements in convergence and mixing, and comes at virtually no additional cost since the inverse contained in the projection matrix must be computed to sample from Eq. 20. It should also be noted that the posterior precision matrix in Eq. 20 requires the inversion of a $p^2 \times p^2$ matrix which is computationally expensive when $p$ is large. We therefore present two strategies to facilitate scalability in the following subsections.

### 4.1 Diagonal restriction on $\Sigma$

As with all Bayesian regression models, a computational bottleneck arises in inverting the posterior precision matrix $(\tilde{\text{X}}'\tilde{\text{X}} + \Lambda_*^{-1})$. This especially true for Bayesian SUR models since the design matrix X contains $p$ stacked copies of the multivariate regression design matrix. If $\Sigma$ is unrestricted, then $\tilde{\text{X}}'\tilde{\text{X}}$ is a dense $p^2 \times p^2$ matrix and any sampler that directly inverts this matrix will be hopeless for large $p$. For example, even a sampler that calculates the inverse using Cholesky decompositions has complexity $\mathcal{O}(p^6)$. If instead $\Sigma$ is assumed to be diagonal then both $\tilde{\text{X}}'\tilde{\text{X}}$ and $(\tilde{\text{X}}'\tilde{\text{X}} + \Lambda_*^{-1})$ will have block diagonal structures, with each of the $p$ blocks containing an $p \times p$

matrix. Computing the inverse of $(\tilde{X}'\tilde{X} + \Lambda_*^{-1})$ then amounts to inverting each $p \times p$ block, which has computational complexity $\mathcal{O}(p^4)$ using Cholesky decompositions. While this is better than inverting $(\tilde{X}'\tilde{X} + \Lambda_*^{-1})$ directly, it can still be prohibitively expensive for large $p$.

### 4.2 Fast sampling normal scale mixtures

Bhattacharya et al. (2016) present an alternative approach for sampling from the posteriors of linear regression models with normal scale mixture priors. The idea is to use data augmentation and a series of linear transformations to avoid the inversion of $(\tilde{X}'\tilde{X} + \Lambda_*^{-1})$. Instead, their algorithm replaces the inversion of $\tilde{X}'\tilde{X}$ with the inversion of $\tilde{X}\tilde{X}'$. For a multiple regression model, this means the matrix being inverted is $n \times n$ instead of $p \times p$ and the proposed algorithm has complexity that is linear in $p$. In the context of our SUR model, the fast sampling algorithm has complexity $\mathcal{O}(n^2p^2)$ if $\Sigma$ is diagonal or $\mathcal{O}(n^2p^4)$ if $\Sigma$ is unrestricted.

Since the original algorithm was also developed for typical shrinkage priors centered at zero, we present a modified algorithm to allow for the nonzero mean structure, which we denote as $\bar{\beta}(\theta)$, in the proposed hierarchical priors:

1. Sample $u \sim \mathrm{N}(\bar{\beta}(\theta), \Lambda_*)$ and $\delta \sim \mathrm{N}(0, I_{np})$;
2. Set $v = \tilde{X}u + \delta$;
3. Compute $w = (\tilde{X}\Lambda_*\tilde{X}' + I_{np})^{-1}(\tilde{y} - v)$;
4. Set $\beta = u + \Lambda_*\tilde{X}'w$.

A constructive proof that $\beta$ retains the posterior in Eq. 20 is provided in Appendix A.2. Note that the computational gains come from the third step, which requires inverting the $np \times np$ matrix $(\tilde{X}\Lambda_*\tilde{X}' + I_{np})$ rather than the original $p^2 \times p^2$ precision matrix $(\tilde{X}'\tilde{X} + \Lambda_*^{-1})$. This also shows that the computational gains are largest when $p$ is much larger than $n$.

### 4.3 Scalability

To provide practical insights into the computational gains afforded by fast sampling algorithm above, we draw from the posterior of the elasticity vector $\beta$ using data generated with $n = 100$ and $p \in \{100, 200, 300, \ldots, 1000\}$. In addition to the fast sampler of Bhattacharya et al. (2016), we also provide results for a "standard" sampler that inverts the $p^2 \times p^2$ precision matrix $(\tilde{X}'\tilde{X} + \Lambda_*^{-1})$ via Cholesky decompositions (see, e.g., chapters 2.12 and 3.5 of Rossi et al., 2005). In both cases we assume $\Sigma$ is diagonal. The samplers are coded in Rcpp (Eddelbuettel & François, 2011) and run on a MacBook Pro laptop with 32GB of RAM and an Apple M1 Max processor. Figure 2 plots the computation time in log seconds against the number of products $p$. We find that the fast sampler offers significant computational savings: it is roughly two times faster when $p = 200$, 10 times faster when $p = 500$, and 30 times faster when $p = 1000$.

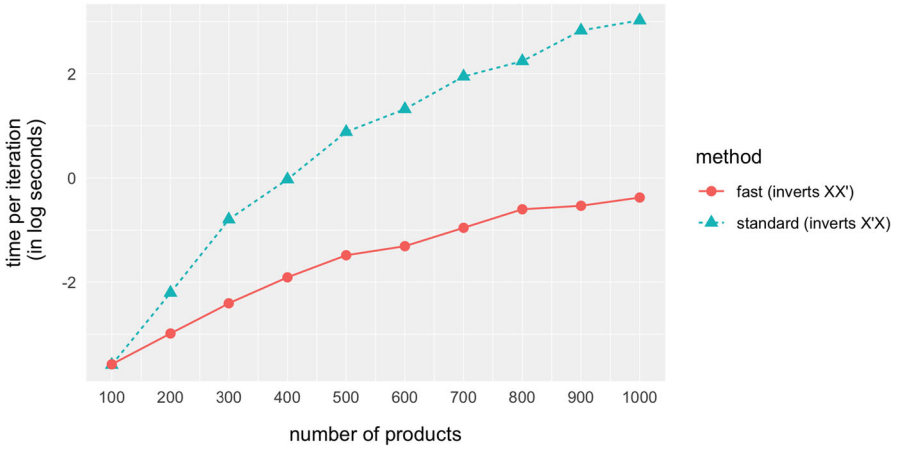



**Fig. 2** Computation time associated with taking one draw from the posterior of the product-level elasticities $\beta$ with fixed sample size $n = 100$ and varying numbers of products $p$

## 5 Simulation experiments

We first explore the performance of hierarchical global-local priors using simulated data. We generate data from price elasticity matrices that have different underlying structures in order to illustrate differences between sparse and non-sparse shrinkage, as well as differences in the heaviness of shrinkage used across levels in the tree. Each of these structures is described in detail below.

(I) **Hierarchical + Dense**: A dense vector of elasticities $\beta_{ij}$ is generated from a three-level hierarchical prior. The top-level coefficients $\theta_{\pi(ij|2)}$ are sampled from a uniform distribution over the interval {-3,-1}∪{1,3} and we fix $\lambda^2_{\pi(ij|\ell)} = 1$ and $\tau^2_\ell = 1$ for all $(i, j)$ and across all levels. The middle-level coefficients $\theta_{\pi(ij|1)}$ and product elasticities $\beta_{ij}$ are generated through the model outlined in Section 3.2. In this specification, all pairs of goods have a non-zero cross-price elasticity.

(II) **Hierarchical + Sparse After Transformation**: A dense vector of elasticities $\beta_{ij}$ is generated from a three-level hierarchical prior where 75% of the product-level local variances $\Psi_{ij}$ are set to zero so that the corresponding product-level elasticity $\beta_{ij}$ is exactly equal to its prior mean. Thus, all pairs of *product groups* have a non-zero cross-price elasticity and many product-level elasticities are exactly equal to the group-level elasticity. This creates a structure where $\beta$ appears dense, but is sparse after subtracting off the prior mean parameters $\theta_{\pi(ij|1)}$.

(III) **Hierarchical + Group-Wise Sparse**: A "group-wise" sparse vector of elasticities $\beta_{ij}$ is generated from a three-level hierarchical prior where 75% of the level-1 coefficients $\theta_{\pi(ij|1)}$ and local variances $\Psi_{\pi(ij|1)}$ are both set to zero. This allows the product-level elasticities to inherit their sparsity from higher

levels in the tree, which in turn produces blocks of cross elasticities that are either all dense or exactly equal to zero.

(IV) **Sparse**: A sparse vector of elasticities $\beta_{ij}$ is generated from a non-hierarchical prior with 95% of all elasticities set to zero.

For each specification above, we generate data from the regression model in Eq. 1 with dimensions ($n = 50$, $p = 100$) and ($n = 100$, $p = 300$). For specifications (I)–(III), we use a three-level classification tree with five equal-sized groups on the top level and 10 equal-sized groups in the middle level. Across all specifications, we simulate the own elasticities from a Unif(-5,-1) distribution, fix $\phi_i = 0$, define $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with $\sigma_j^2 = 1$, and generate elements of data matrices $X$ and $C_j$ from a N(0,1) distribution. We generate 25 data sets from each specification. Examples of the resulting elasticity matrices are shown in Fig. 3.

We take a total of nine models to the data. The first three models impose standard shrinkage with fixed shrinkage points set to zero. The second three models include a hierarchical prior on $\beta$ with ridge shrinkage for the upper-level parameters $\theta$. The final three models include a hierarchical prior on $\beta$ with horseshoe shrinkage at the upper level. Within each batch, we specify three types of mixing distributions on $\lambda_{ij}^2$ which in turn induce three types of shrinkage for $\beta$: ridge, lasso, and horseshoe. All models are fit using the MCMC sampling algorithms outlined in Section 4. In Appendix B.1, we plot the posterior means and 95% credible intervals for elasticity parameters in the ($\theta$-Ridge, $\beta$-Ridge) model to demonstrate that parameters can be accurately recovered.

Model fit statistics are reported in Table 2. The top panel reports results from the ($n = 50$, $p = 100$) data sets and the bottom panel reports results from the ($n = 100$, $p = 300$) data sets. The first four columns report the root mean squared error (RMSE) associated with the estimation of $\beta$ in each specification, averaged across the 25 data replicates. The results illustrate the relative strengths of each type of shrinkage prior. In specification (I) where $\beta$ is purely dense, the hierarchical prior with ridge shrinkage across all levels of the tree fits best. Not surprisingly, a ridge prior is most appropriate when the true parameter is dense. In specification (II) where $\beta$ appears dense but is sparse after subtracting off the prior mean, the hierarchical prior with horseshoe shrinkage for $\beta$ and ridge shrinkage for $\theta$ is

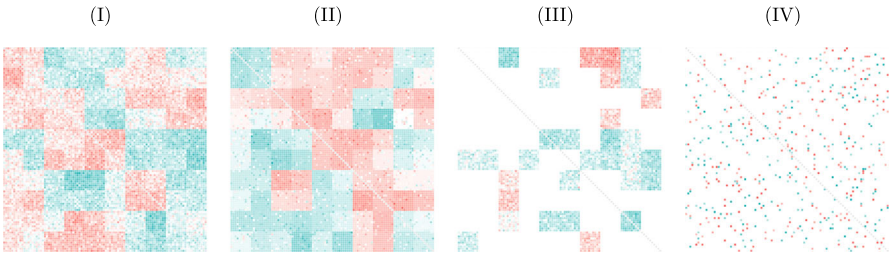



**Fig. 3** Examples of simulated 100 × 100 cross elasticity matrices from the following specifications: (I) hierarchical + dense; (II) hierarchical + sparse after transformation; (III) hierarchical + group-wise sparse; and (IV) sparse. Darker tiles indicate larger elasticities in magnitude. Color version: red tiles represent negative elasticities and green tiles represent positive elasticities

best. This shows how heavy horseshoe shrinkage can still be desirable for parameters that are sparse under some transformation of the data. In specification (III) where $\beta$ is "group-wise" sparse, a hierarchical prior with horseshoe shrinkage for $\theta$ and ridge shrinkage for $\beta$ is best. By forcing $\beta$ to inherit its heaviness from $\theta$, the prior for $\beta$ still exhibits heavy shrinkage—reaffirming some of the theory results in Section 3.4—and behaves like group-wise variable selection. In all three specifications, all hierarchical priors substantially outperform the standard shrinkage priors. In specification (III), unlike specifications (I) and (II), all $\theta$-Horseshoe priors outperform all $\theta$-Ridge priors. Finally, in specification (IV) where $\beta$ is sparse in the traditional sense, then the standard horseshoe prior with fixed shrinkage at zero is best. However, either of the hierarchical priors with horseshoe shrinkage for $\beta$ also offer competitive performance.

The last four columns of Table 2 report the share of signs that are correctly estimated, averaging across data replicates. Producing incorrectly signed elasticity estimates—e.g., negative cross elasticity estimates for goods that are obviously substitutes—is one of the long-standing empirical problems with unrestricted log-linear models (Allenby, 1989; Blattberg & George, 1991; Boatwright et al., 1999). We find that hierarchical priors are largely able to mitigate this issue. For example, while the standard shrinkage priors correctly estimate 73–78% of signs in specifications (I) and (II), hierarchical priors correctly estimate 92–98%.

So far we have assumed that the product classification tree fit to the data matches that of the data generating process. In reality, we do not know whether there is a true hierarchical structure governing cross elasticities, let alone what that structure looks like. In this paper we take the view that category definitions of the retailer or data provider can be used as a reasonable proxy for boundaries in substitution, though this may ultimately be an empirical question (see, e.g., Smith et al., 2019). In Appendix B we provide additional simulations that explore robustness to tree misspecification. We conduct three sets of simulations that vary in the degree of tree misspecification. We find that hierarchical shrinkage can still provide gains (and is no-worse than standard shrinkage) in the presence of misspecification, although the magnitude of these gains vanishes as the degree of misspecification grows. This suggests that hierarchical priors are fairly robust to misspecification of the tree but performance gains over standard sparse approaches depends on the quality of the tree.

## 6 Empirical application

### 6.1 Data

We apply log-linear demand models with standard and hierarchical shrinkage priors to IRI store-level transaction data (Bronnenberg et al., 2008). We use data from the largest grocery retail store in Pittsfield, Massachusetts for the two-year period spanning 2011–2012. We use the first 78 weeks as training data and the last 26 weeks as holdout data to evaluate model fit. Although there is the potential to add data from

**Table 2** Simulation results

| | Estimation RMSE | | | | Correct Signs | | | |
|---|---|---|---|---|---|---|---|---|
| Data: $n = 50$, $p = 100$ | (I) | (II) | (III) | (IV) | (I) | (II) | (III) | (IV) |
| Standard Shrinkage | | | | | | | | |
| $\beta$-Ridge | 2.541 | 2.375 | 1.219 | 0.492 | 0.77 | 0.78 | 0.87 | 1.00 |
| $\beta$-Lasso | 2.622 | 2.473 | 1.079 | 0.314 | 0.76 | 0.77 | 0.89 | 1.00 |
| $\beta$-Horseshoe | 2.879 | 2.745 | 0.969 | **0.098** | 0.73 | 0.74 | 0.88 | 1.00 |
| Hierarchical Shrinkage | | | | | | | | |
| $\theta$-Ridge, $\beta$-Ridge | **1.033** | 0.530 | 0.515 | 0.490 | 0.94 | 0.97 | 0.94 | 1.00 |
| $\theta$-Ridge, $\beta$-Lasso | 1.047 | 0.461 | 0.451 | 0.313 | 0.93 | 0.98 | 0.95 | 1.00 |
| $\theta$-Ridge, $\beta$-Horseshoe | 1.126 | **0.385** | 0.407 | 0.102 | 0.92 | 0.98 | 0.95 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Ridge | 1.044 | 0.526 | **0.190** | 0.472 | 0.93 | 0.98 | 0.98 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Lasso | 1.053 | 0.461 | 0.192 | 0.310 | 0.93 | 0.98 | 0.98 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Horseshoe | 1.130 | 0.386 | 0.270 | 0.100 | 0.92 | 0.98 | 0.97 | 1.00 |
| Data: $n = 100$, $p = 300$ | (I) | (II) | (III) | (IV) | (I) | (II) | (III) | (IV) |
| Standard Shrinkage | | | | | | | | |
| $\beta$-Ridge | 2.938 | 2.743 | 1.275 | 0.545 | 0.72 | 0.73 | 0.83 | 0.99 |
| $\beta$-Lasso | 2.972 | 2.790 | 1.232 | 0.427 | 0.72 | 0.73 | 0.84 | 1.00 |
| $\beta$-Horseshoe | 3.225 | 3.045 | 1.185 | **0.068** | 0.69 | 0.70 | 0.84 | 1.00 |
| Hierarchical Shrinkage | | | | | | | | |
| $\theta$-Ridge, $\beta$-Ridge | **1.162** | 0.585 | 0.530 | 0.544 | 0.94 | 0.98 | 0.93 | 0.99 |
| $\theta$-Ridge, $\beta$-Lasso | 1.170 | 0.547 | 0.488 | 0.427 | 0.93 | 0.98 | 0.94 | 1.00 |
| $\theta$-Ridge, $\beta$-Horseshoe | 1.253 | **0.503** | 0.471 | 0.069 | 0.93 | 0.99 | 0.94 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Ridge | 1.169 | 0.586 | **0.170** | 0.543 | 0.93 | 0.98 | 0.98 | 0.99 |
| $\theta$-Horseshoe, $\beta$-Lasso | 1.172 | 0.548 | 0.174 | 0.426 | 0.93 | 0.98 | 0.98 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Horseshoe | 1.255 | 0.508 | 0.233 | 0.068 | 0.93 | 0.99 | 0.97 | 1.00 |

Fit statistics are averaged across 25 simulated data sets. Columns denote the four different true parameter specifications: (I) hierarchical + dense; (II) hierarchical + sparse after transformation; (III) hierarchical + group-wise sparse; and (IV) sparse. Standard shrinkage priors have a fixed mean of zero and do not incorporate any hierarchical structure. Hierarchical shrinkage priors are based on a three-level tree

other chains and markets, we take the perspective of the retailer who wants to estimate store-level elasticities to allow for more granular customization of marketing instruments (Montgomery, 1997).

The scope of products included in a large-scale demand analysis usually take one of two forms: (i) narrow and deep, where products only come from one category or subcategory but are defined at a very granular level like the UPC; or (ii) wide and shallow, where products span many categories but are defined at a higher level of aggregation. Here we take the latter approach, in part because a UPC-level analysis often creates challenges for log-linear models due to the potentially high incidence of

zero quantities. Estimating demand for wider multi-category assortments also highlights the general flexibility of log-linear systems—e.g., the ability to accommodate a mix of substitutes and complements.

Our procedure for selecting products is as follows. We first choose nine product categories that cover a large portion of total expenditure. Within each category, we remove any subcategories that account for less than 2% of within category revenue, resulting in a total of 28 subcategories. We then aggregate UPCs to the category-subcategory-brand level using total volume (in ounces) and revenue-weighted prices (per ounce). We sort brands by their within-subcategory revenue and then keep top brands that together capture 85% of subcategory revenue and are also present in at least 95% of weeks in the data. This results in $p = 275$ total products, comprised of over 170,000 UPCs, that make up 81.5% of total revenue within the nine chosen categories. A list of the categories and subcategories used in our analysis is provided in Table 3.

### 6.2 Models

We estimate log-linear demand models of the form in Eq. 1 assuming a diagonal error variance matrix $\Sigma$. In addition to the high-dimensional vector of prices, we also include product intercepts, summer and holiday dummy variables, and feature and display promotion variables as controls. We consider three different types of shrinkage priors for the price elasticity parameters: standard (where the prior mean is fixed to zero), hierarchical with ridge shrinkage at the upper levels, and hierarchical with horseshoe shrinkage at the upper levels. At the product level we consider ridge, lasso, and horseshoe shrinkage. Together, this results in a total of nine models.

Across all models, we allow the own elasticities to have a different prior than the cross elasticities to account for differences in their expected sign and magnitude. Specifically, in the sparse models we allow the own elasticities to have a separate global variance parameter. In the hierarchical models, we specify separate hierarchical priors for the own and cross elasticities. Each is based on a three-level tree (products at $\ell = 0$, subcategories at $\ell = 1$, categories at $\ell = 2$) where shrinkage starts to propagate at the subcategory level. A complete list of prior and hyperparameter specifications is provided in Appendix C.1. All models are estimated using the MCMC algorithms outlined in Section 4 which are run for 100,000 iterations and thinned by keeping every 100th draw. The standard shrinkage models, hierarchical $\theta$-Ridge models, and hierarchical $\theta$-Horseshoe models take roughly two minutes, four minutes, and six minutes per 1,000 iterations, respectively. Mixing diagnostics are reported in Appendix C.

### 6.3 Results

#### 6.3.1 Predictive fit

Table 4 reports the out-of-sample RMSEs associated with three sets of demand predictions. The first two columns in Table 4 report the mean and standard deviation of

**Table 3** Summary of product categories

| Category | Subcategories | No. of Products | Share of Revenue |
|---|---|---|---|
| BEER/ALE | Domestic Beer/Ale | 62 | 16.1 |
| | Imported Beer/Ale | | |
| CARBONATED BEVERAGES | Low Calorie Soft Drinks | 30 | 19.1 |
| | Regular Soft Drinks | | |
| | Seltzer/Tonic/Club Soda | | |
| COFFEE | Ground Coffee | 27 | 8.5 |
| | Ground Decaffeinated Coffee | | |
| | Instant Coffee | | |
| | Single Cup Coffee | | |
| | Whole Coffee Beans | | |
| COLD CEREAL | Ready-to-Eat Cereal | 53 | 9.3 |
| FZ DINNERS/ENTREES | Fz Handheld Entrees | 36 | 7.8 |
| | Multi-Serve Fz Dinners | | |
| | Single-Serve Fz Dinners | | |
| FZ PIZZA | Fz Pizza | 7 | 5.0 |
| MILK | Rfg Almond Milk | 10 | 12.9 |
| | Rfg Flavored Milk/Eggnog/Buttermilk | | |
| | Rfg Skim/Lowfat Milk | | |
| | Rfg Soy Milk | | |
| | Rfg Whole Milk | | |
| SALTY SNACKS | Cheese Snacks | 35 | 13.4 |
| | Corn Snacks | | |
| | Other Salted Snacks | | |
| | Potato Chips | | |
| | Pretzels | | |
| | Ready to Eat Popcorn/Caramel Corn | | |
| | Tortilla/Tostada Chips | | |
| YOGURT | Rfg Yogurt | 15 | 7.9 |
| Total Count = 9 | 28 | 275 | 100% |

In our analysis, we aggregate over flavors and pack sizes and define the product at the brand level. In this product classification tree, the category is the top level ($\ell = 2$), the subcategory is the middle level ($\ell = 1$), and the product is the bottom level ($\ell = 0$). Here there are $9^2 = 81$ category elasticities $\theta_{\pi(ij|2)}$, $28^2 = 784$ subcategory elasticities $\theta_{\pi(ij|1)}$, and $275^2 = 75,625$ product elasticities $\beta_{ij}$. See Table 1 for a complete glossary of parameters associated with each level of the tree

**Table 4** Out-of-sample fit statistics

| | All Products | | Extrapolated Price Levels | | Limited Price Variation | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Standard Shrinkage | | | | | | |
| $\beta$-Ridge | 0.846 | (0.126) | 0.871 | (0.190) | 0.968 | (0.347) |
| $\beta$-Lasso | 0.847 | (0.117) | 0.865 | (0.181) | 0.960 | (0.331) |
| $\beta$-Horseshoe | 0.895 | (0.170) | 0.937 | (0.258) | 1.085 | (0.449) |
| Hierarchical Shrinkage | | | | | | |
| $\theta$-Ridge, $\beta$-Ridge | **0.808** | (0.104) | **0.816** | (0.148) | **0.895** | (0.276) |
| $\theta$-Ridge, $\beta$-Lasso | 0.814 | (0.113) | 0.824 | (0.170) | 0.909 | (0.314) |
| $\theta$-Ridge, $\beta$-Horseshoe | 0.899 | (0.131) | 0.889 | (0.194) | 1.004 | (0.347) |
| $\theta$-Horseshoe, $\beta$-Ridge | 0.842 | (0.119) | 0.825 | (0.168) | 0.919 | (0.307) |
| $\theta$-Horseshoe, $\beta$-Lasso | 0.823 | (0.117) | 0.821 | (0.169) | 0.908 | (0.308) |
| $\theta$-Horseshoe, $\beta$-Horseshoe | 0.993 | (0.159) | 0.845 | (0.162) | 0.902 | (0.305) |

Out-of-sample RMSEs are reported for: (i) all 275 products; (ii) the 152 products with extrapolated price levels in the test sample; (iii) the 69 products in the lower quartile of the distribution of price variation. Standard shrinkage priors have a fixed mean of zero and do not incorporate any hierarchical structure. Hierarchical shrinkage priors are based on a three-level tree with means estimated at each level

RMSEs across all 275 products. The middle two columns report RMSEs for a subset of 152 products with extrapolated price levels in the test sample. A product is included in this set if at least one price point in the test sample falls outside of its range of prices in the training sample. The final two columns report RMSEs for a subset of 69 products with limited price variation. A product $j$ is included in this set if $\text{SD}_j = \text{SD}(\log p_{j1}, \ldots, \log p_{jn})$ falls in the lower quartile of the distribution of $\text{SD}_1, \ldots, \text{SD}_{275}$. Together, these three types of predictions allow us to examine model performance across a range of forecasting scenarios that retailers may face: predicting demand for an entire assortment, predicting demand at counterfactual price levels, and predicting demand for "low signal" goods.

We find that the ($\theta$-Ridge, $\beta$-Ridge) hierarchical prior provides the best predictions across all three tasks. Relative to the best-performing standard shrinkage prior, the ($\theta$-Ridge, $\beta$-Ridge) prior generates a 4.5% improvement in predictions across all products, a 5.7% improvement for products with extrapolated price levels, and a 6.8% improvement for products with limited price variation. We also find that horseshoe shrinkage at the product level tends to perform worst, regardless of if and how the prior mean is parameterized. One possible explanation is that the true elasticities are actually dense, rather than sparse, in which case ridge shrinkage should perform well (as shown in the simulations in Section 5). A second possibility is that the true magnitude of the elasticities may not be large enough to warrant a heavy-tailed prior like the horseshoe. In reality, it may be a combination of both; we will revisit this question as we analyze estimates from each model below.

### 6.3.2 Product-level elasticities

Next, we compare the estimated product-level elasticities from each model. Table 5 reports the overall average and the 10th, 50th, and 90th percentiles of the distribution of posterior means of $\beta_{ii}$ and $\beta_{ij}$. We also report the share of own elasticities that are negative and the share of own elasticities that are significant at the 5% level. Complete distributions of price elasticity and promotion effect estimates are shown in Appendix C.3. Elasticity estimates are markedly different across prior specifications. Starting with own elasticities, we find that standard priors produce distributions of own elasticities where roughly 84% of estimates are negative, 21% of estimates are significant, and the average (median) own elasticity is around -1.2 (-0.9). In contrast, hierarchical priors produce distributions of own elasticity estimates where 93% are negative, 50% are significantly away from zero, and the average (median) own elasticity is around -1.5 (-1.4). We believe that being able to produce more economically reasonable and precise own elasticity estimates by simply shrinking to higher-level elasticities rather than zero is a strength of our approach.

The distribution of cross elasticity estimates also differs across prior specifications. We are estimating 75,350 cross elasticities using less than 100 weeks of training data, and so it is not surprising that the prior imposes heavy regularization. Because typical sparse priors shrink estimates towards zero, the associated distribution of estimates is highly concentrated around zero. Hierarchical priors produce distributions of estimates that are also centered around zero, but spread mass more widely in the (-0.1, 0.1) interval. The exact shape of the distribution depends on the type of shrinkage imposed at each level of the tree. Interestingly, the ($\theta$-Horseshoe, $\beta$-Horseshoe)

**Table 5** Summary of product-level price elasticity estimates

| | Own Elasticities $\beta_{ii}$ | | | | | | Cross Elasticities $\beta_{ij}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Neg | Sig | Mean | 10th | 50th | 90th | Mean | 10th | 50th | 90th |
| Standard Shrinkage | | | | | | | | | | |
| $\beta$-Ridge | 84.0 | 25.8 | -1.20 | -2.97 | -1.09 | 0.41 | -0.000 | -0.000 | -0.000 | 0.000 |
| $\beta$-Lasso | 83.6 | 21.5 | -1.13 | -2.95 | -0.86 | 0.28 | -0.000 | -0.000 | -0.000 | 0.000 |
| $\beta$-Horseshoe | 85.1 | 16.7 | -1.23 | -3.38 | -0.82 | 0.26 | -0.001 | -0.004 | -0.000 | 0.003 |
| Hierarchical Shrinkage | | | | | | | | | | |
| $\theta$-Ridge, $\beta$-Ridge | 93.5 | 41.5 | -1.58 | -3.01 | -1.53 | -0.28 | -0.004 | -0.023 | -0.004 | 0.012 |
| $\theta$-Ridge, $\beta$-Lasso | 94.2 | 42.9 | -1.55 | -2.87 | -1.47 | -0.27 | -0.004 | -0.024 | -0.004 | 0.016 |
| $\theta$-Ridge, $\beta$-Horseshoe | 89.5 | 49.5 | -1.42 | -2.64 | -1.40 | 0.09 | -0.001 | -0.121 | -0.002 | 0.114 |
| $\theta$-Horseshoe, $\beta$-Ridge | 97.5 | 54.9 | -1.63 | -2.88 | -1.48 | -0.59 | -0.004 | -0.020 | -0.001 | 0.013 |
| $\theta$-Horseshoe, $\beta$-Lasso | 97.1 | 61.1 | -1.57 | -2.77 | -1.46 | -0.64 | -0.004 | -0.019 | -0.003 | 0.010 |
| $\theta$-Horseshoe, $\beta$-Horseshoe | 87.6 | 49.5 | -1.30 | -2.56 | -1.14 | 0.05 | -0.008 | -0.086 | -0.004 | 0.071 |

Percentiles of the distribution of own and cross-price elasticity estimates are reported for each model. We also report the share of negative own elasticities ("Neg") and the share of own elasticities whose 95% credible intervals do not contain zero ("Sig"). Standard shrinkage priors have a fixed mean of zero and do not incorporate any hierarchical structure. Hierarchical shrinkage priors are based on a three-level tree

prior generates a more widely dispersed distribution of elasticities than the ($\theta$-Ridge, $\beta$-Ridge) prior. This is a consequence of hierarchical priors: endowing $\beta_{ij}$ with a potentially non-zero prior mean leads to a build-up of mass away from zero when shrinkage in strong. Differences in the shape of cross-elasticity estimates can also be seen in the histograms reported in Appendix C.3.

### 6.3.3 Higher-level elasticities

In addition to product-level elasticities, we can also examine higher-level group elasticities in any of the hierarchical priors. For the sake of brevity, we focus our discussion on estimates from the ($\theta$-Ridge, $\beta$-Ridge) prior. Figure 4 plots the distributions of higher-level elasticities at the category and subcategory level. The figure on the left plots own elasticities and the figure on the right plots cross elasticities, where the *within-group* elasticities (green lines) are separated from the *across-group* elasticities (red lines). Here, the within-group elasticities represent the diagonal elements of the category or subcategory-level elasticity matrix. For example, the BEER within-category elasticity represents the average cross-price elasticity of all products within the BEER category. Similarly, the Domestic Beer/Ale within-subcategory elasticity represents the average cross-price elasticity of all products within the Domestic Beer/Ale subcategory.

Figure 4 illustrates two economic properties of elasticities that are not imposed on the model but are a consequence of hierarchical shrinkage. First, own elasticities tend to be negative and are slightly more negative at the subcategory level than at the category level. The fact that product-level own elasticities are shrunk towards these negative values rather than zero is one reason why own elasticities tend to be more negative under hierarchical priors (as shown in Table 5). Second, cross elasticities *within* each category/subcategory are mostly positive and shifted to the right of the distribution of elasticities *across* categories/subcategories. This

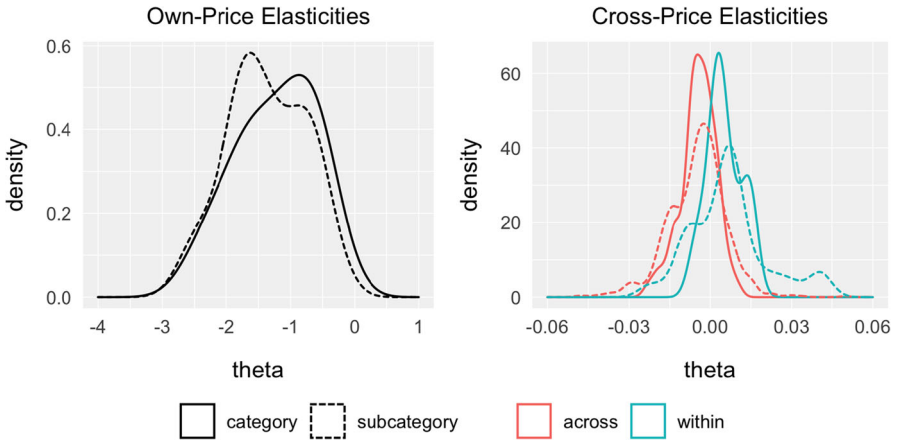



**Fig. 4** Distributions of higher-level elasticities from the ($\theta$-Ridge, $\beta$-Ridge) hierarchical prior

suggests that substitution and product competition tend to be strongest *within* each category/subcategory.

Table 6 provides further insight into the nature of category and subcategory-level substitution. For each category/subcategory, we report the associated largest (most positive) elasticity and smallest (most negative) elasticity. We find that for six of the nine categories, demand changes most in response to price changes within the same category (i.e., the *within* elasticities shown in Fig. 4). For example, the largest elasticity associated with the demand of BEER/ALE is the price of BEER/ALE. This is to be expected if the competition is stronger within categories than across categories. Many of the cross-category relationships also appear reasonable. For example, BEER/ALE and SALTY SNACKS, CARBONATED BEVERAGES and SALTY SNACKS, and MILK and CEREAL are all pairs of strong complements. Results at the subcategory level are similar. For 15 of the 28 subcategories, the largest elasticity corresponds to the either the within subcategory elasticity or the across elasticity within the category (e.g., Domestic vs. Imported Beer/Ale). Many of the strong cross-subcategory complements such as Imported Beer/Ale and Tortilla/Tostada Chips or Rfg Skim/Lowfat Milk and Ready-to-Eat Cereal appear to be inherited from their respective cross-category parent elasticities.

### 6.3.4 Shrinkage factors

In addition to the summary of elasticity parameters presented above, we can also summarize the variance parameters to learn about the strength of shrinkage imposed by the prior. To this end, we explore the posterior distribution of product-level shrinkage factors, which are common summary measures used to convey the strength of shrinkage in global-local models (Carvalho et al., 2010; Bhadra et al., 2019). Shrinkage factors measure the extent to which the posterior mean is pulled towards the prior. Formally, the posterior mean of the regression coefficient $\beta$ takes the form:

$$\mathbb{E}(\beta|\text{data}) = (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\tilde{\mathrm{X}}'\tilde{\mathrm{X}}\hat{\beta} + (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\Lambda_*^{-1}\bar{\beta}(\theta) \tag{23}$$

which is a weighted average of the maximum likelihood estimator $\hat{\beta}$ and the prior mean $\bar{\beta}(\theta)$. The weight on the prior mean, $(\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\Lambda_*^{-1}$, defines the matrix of shrinkage factors. For simplicity, we use the component-wise approximation:

$$\mathbb{E}(\beta_{ij}|\text{data}) \approx (1 - \kappa_{ij})\hat{\beta}_{ij} + \kappa_{ij}\bar{\beta}(\theta)_{ij} \tag{24}$$

where

$$\kappa_{ij} = \frac{1}{1 + ns_j^2\sigma_i^{-2}\tau_\beta^2\Psi_{ij}} \tag{25}$$

and $s_j^2 = \text{Var}(\log p_j)$. Notice that $ns_j^2/\sigma_i^2$ can be interpreted as a signal-to-noise ratio. When the signal dominates the noise, $\kappa_{ij} \to 0$ and the posterior mean of $\beta_{ij}$ converges to $\hat{\beta}_{ij}$; when the noise dominates, then $\kappa_{ij} \to 1$ and the posterior mean of $\beta_{ij}$ converges to the prior mean.

Table 7 reports summary statistics of global variances and product-level price elasticity shrinkage factors across model specifications. We first compute the pos-

**Table 6** Summaries of higher-level cross-price elasticities from the ($\theta$-Ridge, $\beta$-Ridge) Hierarchical Prior

| | | | Largest (Most Positive) Elasticity | | Smallest (Most Negative) Elasticity | |
|---|---|---|---|---|---|---|
| Category | 1 | BEER/ALE | BEER/ALE | 0.014 | SALTY SNACKS | -0.017 |
| | 2 | CARBONATED BEVERAGES | CARBONATED BEVERAGES | 0.015 | SALTY SNACKS | -0.020 |
| | 3 | COFFEE | COFFEE | 0.009 | BEER/ALE | -0.008 |
| | 4 | COLD CEREAL | CARBONATED BEVERAGES | 0.007 | MILK | -0.023 |
| | 5 | FZ DINNERS/ENTREES | FZ DINNERS/ENTREES | 0.004 | CARBONATED BEVERAGES | -0.019 |
| | 6 | FZ PIZZA | FZ PIZZA | 0.002 | COFFEE | -0.006 |
| | 7 | MILK | MILK | 0.005 | COLD CEREAL | -0.010 |
| | 8 | SALTY SNACKS | FZ DINNERS/ENTREES | 0.002 | MILK | -0.015 |
| | 9 | YOGURT | FZ PIZZA | 0.003 | SALTY SNACKS | -0.010 |
| Subcategory | 1 | Domestic Beer/Ale | Domestic Beer/Ale | 0.040 | Ground Decaffeinated Coffee | -0.046 |
| | 2 | Imported Beer/Ale | Domestic Beer/Ale | 0.050 | Tortilla/Tostada Chips | -0.041 |
| | 3 | Low Calorie Soft Drinks | Regular Soft Drinks | 0.034 | Tortilla/Tostada Chips | -0.033 |
| | 4 | Regular Soft Drinks | Regular Soft Drinks | 0.024 | Tortilla/Tostada Chips | -0.033 |
| | 5 | Seltzer/Tonic Water/Club Soda | Low Calorie Soft Drinks | 0.022 | Corn Snacks | -0.021 |
| | 6 | Ground Coffee | Ground Coffee | 0.030 | Domestic Beer/Ale | -0.016 |
| | 7 | Ground Decaffeinated Coffee | Single Serve Fz Dinners/Entrees | 0.013 | Domestic Beer/Ale | -0.020 |
| | 8 | Instant Coffee | Whole Coffee Beans | 0.008 | Rfg Yogurt | -0.019 |
| | 9 | Single Cup Coffee | Instant Coffee | 0.010 | Domestic Beer/Ale | -0.011 |
| | 10 | Whole Coffee Beans | Single Serve Fz Dinners/Entrees | 0.014 | Ready-to-Eat Cereal | -0.017 |
| | 11 | Ready-to-Eat Cereal | Regular Soft Drinks | 0.017 | Rfg Flavored Milk/Eggnog/Butterm | -0.049 |

**Table 6** (continued)

| | | Largest (Most Positive) Elasticity | | Smallest (Most Negative) Elasticity | |
|---|---|---|---|---|---|
| 12 | Fz Handheld Entrees | Domestic Beer/Ale | 0.009 | Regular Soft Drinks | -0.031 |
| 13 | Multi Serve Fz Dinners/Entrees | Single Serve Fz Dinners/Entrees | 0.023 | Low Calorie Soft Drinks | -0.028 |
| 14 | Single Serve Fz Dinners/Entrees | Single Serve Fz Dinners/Entrees | 0.041 | Regular Soft Drinks | -0.030 |
| 15 | Fz Pizza | Ready-to-Eat Popcorn/Caramel Cor | 0.006 | Single Serve Fz Dinners/Entrees | -0.018 |
| 16 | Rfg Almond Milk | Rfg Yogurt | 0.015 | Ready-to-Eat Cereal | -0.010 |
| 17 | Rfg Flavored Milk/Eggnog/Butterm | Rfg Flavored Milk/Eggnog/Butterm | 0.008 | Ready-to-Eat Cereal | -0.016 |
| 18 | Rfg Skim/Lowfat Milk | Rfg Whole Milk | 0.008 | Ready-to-Eat Cereal | -0.019 |
| 19 | Rfg Soy Milk | Rfg Flavored Milk/Eggnog/Butterm | 0.007 | Ready-to-Eat Cereal | -0.013 |
| 20 | Rfg Whole Milk | Rfg Whole Milk | 0.007 | Domestic Beer/Ale | -0.014 |
| 21 | Cheese Snacks | Domestic Beer/Ale | 0.002 | Rfg Almond Milk | -0.017 |
| 22 | Corn Snacks | Single Serve Fz Dinners/Entrees | 0.007 | Rfg Yogurt | -0.015 |
| 23 | Other Salted Snacks | Single Serve Fz Dinners/Entrees | 0.012 | Rfg Flavored Milk/Eggnog/Butterm | -0.019 |
| 24 | Potato Chips | Potato Chips | 0.010 | Rfg Flavored Milk/Eggnog/Butterm | -0.019 |
| 25 | Pretzels | Fz Handheld Entrees | 0.010 | Rfg Whole Milk | -0.015 |
| 26 | Ready-to-Eat Popcorn/Caramel Cor | Single Serve Fz Dinners/Entrees | 0.014 | Rfg Soy Milk | -0.016 |
| 27 | Tortilla/Tostada Chips | Single Serve Fz Dinners/Entrees | 0.008 | Rfg Whole Milk | -0.020 |
| 28 | Rfg Yogurt | Potato Chips | 0.011 | Ready-to-Eat Cereal | -0.022 |

The top section of the table reports posterior means of selected category elasticities $\theta_{\pi(ij|2)}$. The bottom section of the table reports posterior means of selected subcategory elasticities $\theta_{\pi(ij|1)}$. See Table 1 for a glossary of parameters associated with each level of the tree

**Table 7** Global variances and shrinkage factors

| | Own Elasticities | | | | Cross Elasticities | | | |
|---|---|---|---|---|---|---|---|---|
| | | $\kappa_{ii}$ | | | | $\kappa_{ij}$ | | |
| | $\tau^2_{\beta\text{own}}$ | Min | Mean | Max | $\tau^2_{\beta\text{cross}}$ | Min | Mean | Max |
| Standard Shrinkage | | | | | | | | |
| $\beta$-Ridge | 5.11 | 0.01 | 0.15 | 1.00 | 2.51E-06 | 1.00 | 1.00 | 1.00 |
| $\beta$-Lasso | 5.07 | 0.01 | 0.19 | 1.00 | 1.13E-05 | 1.00 | 1.00 | 1.00 |
| $\beta$-Horseshoe | 7.33 | 0.00 | 0.19 | 1.00 | 6.84E-06 | 0.01 | 1.00 | 1.00 |
| Hierarchical Shrinkage | | | | | | | | |
| $\theta$-Ridge, $\beta$-Ridge | 1.77 | 0.02 | 0.29 | 1.00 | 9.63E-05 | 1.00 | 1.00 | 1.00 |
| $\theta$-Ridge, $\beta$-Lasso | 1.82 | 0.02 | 0.35 | 1.00 | 3.15E-05 | 1.00 | 1.00 | 1.00 |
| $\theta$-Ridge, $\beta$-Horseshoe | 0.11 | 0.03 | 0.76 | 1.00 | 1.12E-05 | 0.01 | 1.00 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Ridge | 0.21 | 0.02 | 0.49 | 1.00 | 6.79E-06 | 0.03 | 0.99 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Lasso | 0.42 | 0.02 | 0.58 | 1.00 | 3.81E-05 | 0.02 | 1.00 | 1.00 |
| $\theta$-Horseshoe, $\beta$-Horseshoe | 0.08 | 0.01 | 0.76 | 1.00 | 1.51E-06 | 0.00 | 1.00 | 1.00 |

Posterior medians of $\kappa_{ii}$ and $\kappa_{ij}$ are calculated at the product level and then summarized for each model. There is maximum shrinkage when $\kappa = 1$ and no shrinkage when $\kappa = 0$

terior median of each $\kappa_{ij}$ and then report summary statistics across the distribution of estimates. The first finding is that there is a sizable difference in the strength of shrinkage between own and cross-price elasticities. We find that estimates of $\tau^2_{\beta\text{cross}}$ tend to be four to five orders of magnitude smaller than $\tau^2_{\beta\text{own}}$. Consequently, estimates of $\kappa_{ij}$ (for $i \neq j$) tend to be bunched at one while estimates of $\kappa_{ii}$ are more dispersed throughout the unit interval. One explanation for this difference in shrinkage is that retail scanner data tends to exhibit a stronger signal for estimating own elasticities than cross elasticities (Hitsch et al., 2021). Our estimation problem is also high-dimensional as we are estimating 75,350 cross elasticity parameters from 78 weeks of training data.

Across models, we find that hierarchical priors impose heavier shrinkage, on average, than standard global-local priors—especially for own elasticities. For example, the average shrinkage factor $\kappa_{ii}$ is 0.19 for the $\beta$-Horseshoe model but 0.76 for the ($\beta$-Ridge, $\beta$-Horseshoe) model. If the shrinkage points are misspecified (as appears to be the case for standard priors with a mean fixed at zero), then the prior variances will need to get larger to accommodate deviations from zero. Since the hierarchical priors center the product-level elasticities around more reasonable values, then the prior variance can get smaller and shrinkage will "kick in" for noisy estimates. For differences in shrinkage factors across product categories, see Appendix C.4 where we plot the empirical CDF of posterior medians of $\kappa_{ii}$ across both categories and models. We find appreciable variation in the strength of category-level shrinkage. While there is variation across models, we find that shrinkage tends to be heaviest in

categories like BEER/ALE and CARBONATED BEVERAGES, and lightest in FZ PIZZA and SALTY SNACKS.

### 6.4 Discussion

So far, we have provided evidence that hierarchical shrinkage priors can lead to more negative and precisely estimated own elasticities, larger and more dispersed cross elasticities, and improvements in out-of-sample demand predictions. We close this section with a discussion of implications for analysts and managers. We first discuss the nature of price variation in retail markets, and show how judicious prior choices can help mitigate challenges of estimating price elasticities with weakly informative data. We then discuss the benefits of imposing hierarchical shrinkage when producing competitive maps used for market structure analysis.

#### 6.4.1 Retail prices and prior regularization

In the era of big data it may be tempting to believe that prior specification is less of a concern as the data will simply overwhelm the prior with a sufficiently large sample size. We argue that this view is misguided in a demand estimation context because more data does not necessarily imply more variation in key demand shifters.[4] In many retail markets, for example, prices are notoriously "sticky" (Bils & Klenow, 2004) and exhibit limited variation over time—a feature that need not dissipate as more data are collected. With limited variation in prices, price coefficients will in turn be subject to heavy regularization. Analysts interested in using observational data to estimate price elasticities will almost always face a problem of weakly informative data, calling for more judicious prior choices.

We illustrate the interplay between weakly informative data and regularization in Fig. 5. Each scatter plot compares the own elasticities from a sparse prior (x-axis) to the own elasticities from a hierarchical prior (y-axis). For example, the top left corner compares estimates from the sparse $\beta$-Ridge prior to estimates from the hierarchical ($\theta$-Ridge, $\beta$-Ridge) prior. Each point is one of 275 products colored according to the strength of the signal provided its price vector. Our specific measure of signal strength for product $j$ is the standard deviation of log prices across weeks in the training data: $\text{SD}_j = \text{SD}(\log p_{j1}, \ldots, \log p_{jn})$. Red triangles represent products in the bottom quartile of this distribution (i.e., products with relatively limited price variation), green squares represent products in the top quartile of this distribution, and grey circles represent products in the second and third quartiles.

We find that most of green squares fall along the diagonal line, implying that the own elasticity estimates produced by sparse and hierarchical priors are similar for goods with relatively high price variation. In contrast, most of the red triangles fall below the diagonal line where estimates from the sparse prior estimates are greater

[4] We abstract away from the discussion of "clean" variation and price endogeneity. In our empirical application, we only rely on temporal variation rather than cross-market variation in prices which is the typical threat to exogeneity (Rossi, 2014).

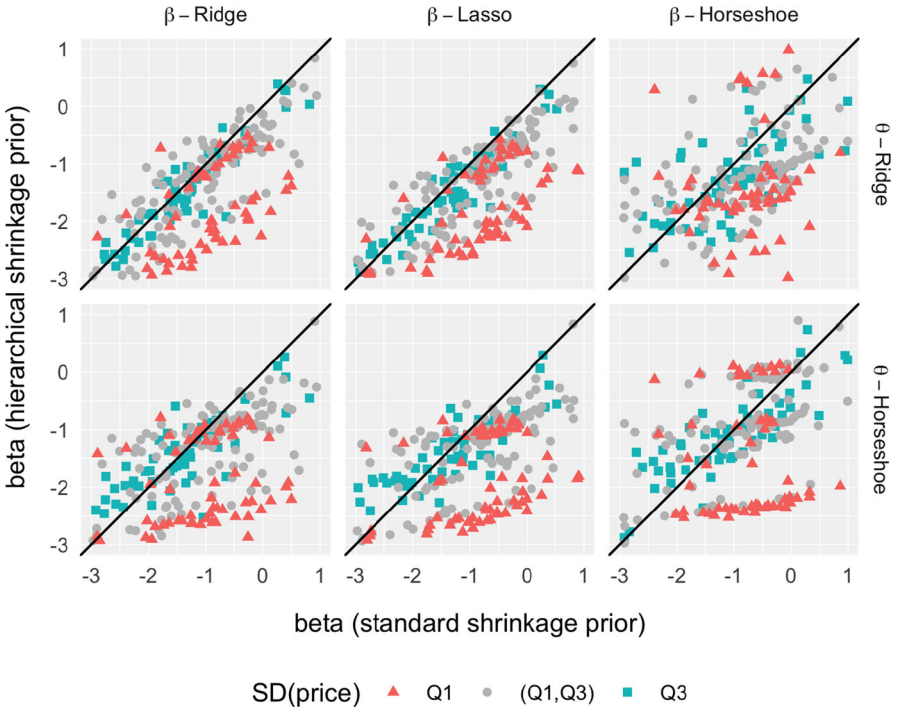



**Fig. 5** The effects of non-sparse shrinkage on own-price elasticities

than estimates from the hierarchical prior. This set of products illustrates the value of hierarchical shrinkage: in the absence of a strong signal in the data, we can produce more economically reasonable estimates by imposing shrinkage towards higher-level elasticities rather than zero.

### 6.4.2 Interpretable market structure

A second implication of hierarchical, non-sparse shrinkage is that it can allow for more interpretable market structure analysis. Marketing researchers have a long history of using competitive maps—i.e., visual representations of product or brand competition—for understanding customer preferences and guiding managerial decisions related to brand positioning (Elrod, 1988; Allenby, 1989; Rutz & Sonnier, 2011), product line management (Chintagunta, 1998), and assortment planning (Sinha et al., 2013). One challenge in producing competitive maps for large assortments is that it is hard to flexibly estimate demand at scale.[5] As we have shown,

[5]Visualizing market structure for large assortments remains an active area of research. In recent years, several papers have bypassed the challenge of estimating demand at scale and have instead developed visualization methods that either use purchase outcome data alone (France & Ghose, 2016; Gabel et al., 2019), or use auxiliary data such as data on consumer search (Kim et al., 2011; Ringel & Skiera, 2016) or online text reviews (Netzer et al., 2012).

log-linear demand models with appropriate forms of regularization can produce estimates of large elasticity matrices. Do different priors lead to appreciably different inferences about competition and market structure?

Figure 6 plots elasticity-based competitive maps across all nine prior specifications. Each map is made by using a two-dimensional t-SNE projection (van der

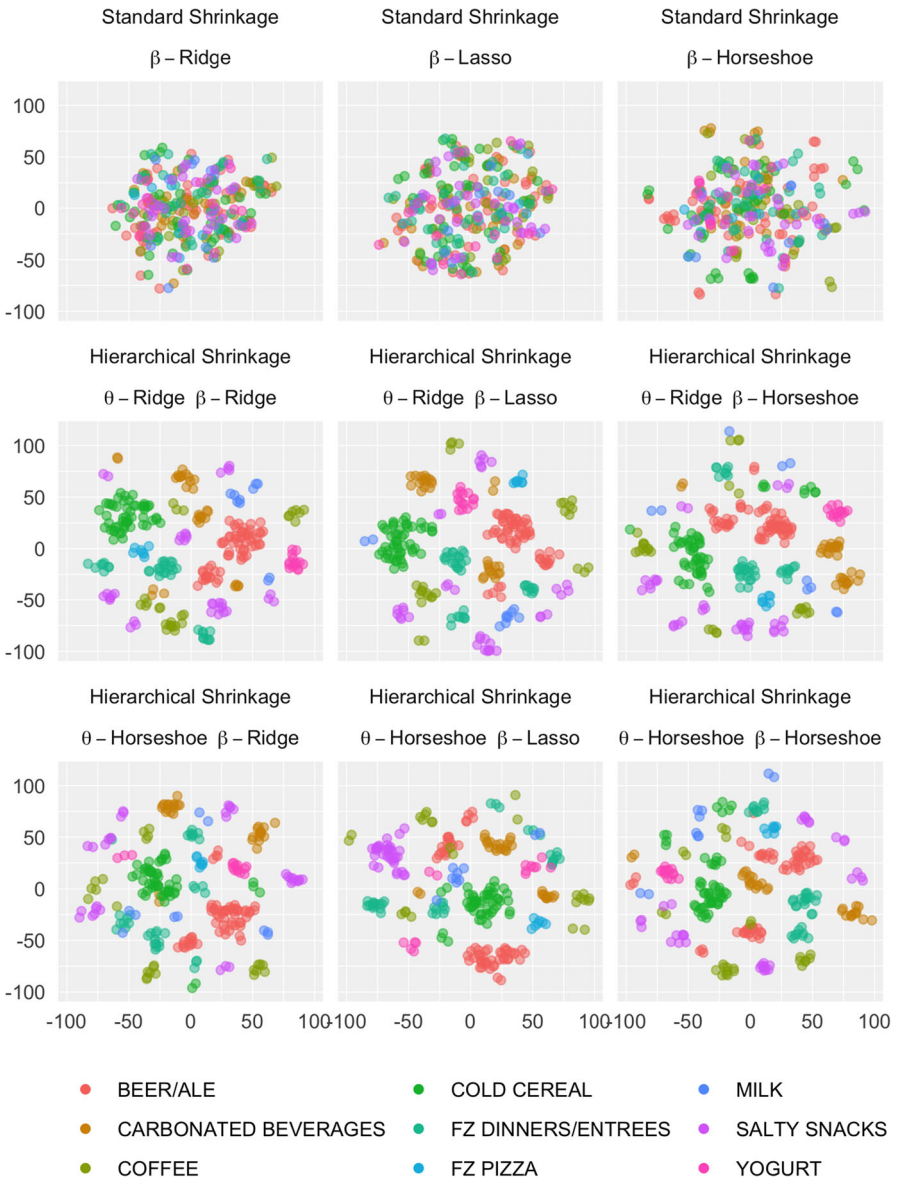



**Fig. 6** Competitive maps based on t-SNE projections of each model's estimated $275 \times 275$ price elasticity matrix. Each point represents a product and is colored by its category label

Maaten & Hinton, 2008) of the estimated price elasticity matrix. Each point on the map is one of 275 products and distances between products reflect degrees of substitution, with more related products appearing closer together. Points are colored based on category labels to help interpret competitive boundaries. We find substantial differences in the interpretability of competitive maps across sparse and hierarchical prior specifications. The maps produced by standard shrinkage priors do not, at first glance, reflect any deeper structure of demand. In contrast, the maps produced by hierarchical shrinkage priors feature a co-clustering of goods belonging to the same or similar categories. For example, in the ($\theta$-Ridge, $\beta$-Ridge) map, most beverages appear on the right-hand side and many beers and carbonated beverages are located near clusters of salty snacks, frozen dinners, and frozen pizza.

In Appendix C5, we show a subcategory-level version of Fig. 6 for two selected models to show how hierarchical shrinkage leads to more coherent subcategory clustering behavior. We also provide zoomed-in plots for the SALTY SNACK category to examine the ways in which brands co-cluster. In the hierarchical shrinkage map, we find that category and subcategory labels tend have more influence over clustering outcomes than brand labels. For instance, we do not find that brands co-cluster across subcategories (as would be expected if umbrella branding or quality tiers induced strong cross-subcategory substitution, for example). Instead, we find that clusters tend to include all brands within the same subcategory. Relative to hierarchical shrinkage maps, standard shrinkage maps exhibit no discernible clustering behavior across brands or subcategories. Overall, we believe that improvements in map interpretability can lead to better inferences about product competition and more valuable insights for for decision-making.

## 7 Conclusion

This paper studies shrinkage priors for high-dimensional demand systems. We propose a hierarchical extension to the class of global-local priors where prior means and variances are parameterized by product classification trees that commonly accompany retail scanner data sets. The principal benefit is that the price elasticity between two goods will be shrunk towards higher-level category elasticities rather than zero. We also formally examine the shrinkage properties of hierarchical global-local priors and show that including a heavy tailed mixing distribution at any level of the tree is sufficient for imposing heavy shrinkage for product-level elasticities.

We apply our hierarchical priors to the elasticity parameters of a log-linear demand model in which store-level sales are regressed on a high-dimensional vector of prices as well as seasonal trends and other product controls. We propose a simple modified version of the fast sampling algorithm in Bhattacharya et al. (2016) to help alleviate the typical computational bottleneck that arises when inverting the posterior precision matrix. We then use both simulated data and retail scanner data to show the value of non-sparse shrinkage. Our simulation experiments shed light on the situations in which different types of shrinkage are most/least appropriate. Hierarchical priors can

offer significant gains when the true elasticity matrix is dense and inherits its structure from a product classification tree. Moreover, hierarchical priors are fairly robust to tree misspecification and are still competitive with standard sparsity priors when the parameters are actually sparse.

Our empirical application uses two years of data to estimate 75,625 price elasticities associated with an assortment of 275 goods that span 28 subcategories and nine categories. We find that the hierarchical prior with ridge shrinkage throughout the tree provides the best predictive performance. We also find that hierarchical priors provide much more reasonable estimates of own elasticities than the standard shrinkage priors which impose a fixed shrinkage point at zero. The hierarchical global-local structure also allows us to learn about within and cross-subcategory and category elasticities, which is useful for identifying boundaries of competition. Our results provide evidence that, in the absence of a strong signal in the data, hierarchical shrinkage can lead to improvements in different aspects of demand estimation relative to "off-the-shelf" regularization. More generally, we believe our work highlights the importance of judicious priors in high-dimensional estimation problems.

## Appendix A: Posterior computation with hierarchical global-local priors

### A.1: Full Conditionals

**Means** The higher-level elasticity parameters $\theta_{\pi(ij|\ell)}$ are assumed to follow a normal prior and therefore have a normal full conditional distribution: $\theta_{\pi(ij|\ell)}|\text{else} \sim \text{N}(\tilde{\theta}_{\pi(ij|\ell)}, V_{\pi(ij|\ell)})$, where

$$\tilde{\theta}_{\pi(ij|\ell)} = V_{\pi(ij|\ell)} \left( \sum_{a \in \mathcal{C}_i^\ell} \sum_{b \in \mathcal{C}_j^\ell} \frac{\theta_{\pi(ab|\ell-1)}}{\tau_{\ell-1}^2 \Psi_{\pi(ab|\ell-1)}} + \frac{\theta_{\pi(ij|\ell+1)}}{\tau_\ell^2 \Psi_{\pi(ij|\ell)}} \right),$$

$$V_{\pi(ij|\ell)} = \left( \sum_{a \in \mathcal{C}_i^\ell} \sum_{b \in \mathcal{C}_j^\ell} \frac{1}{\tau_{\ell-1}^2 \Psi_{\pi(ab|\ell-1)}} + \frac{1}{\tau_\ell^2 \Psi_{\pi(ij|\ell)}} \right)^{-1}.$$

Here $\mathcal{C}_i^\ell = \left\{ i' \in \{1, \ldots, n_\ell\} : \pi(i'j|\ell) = \pi(ij|\ell) \right\}$ and $\mathcal{C}_j^\ell = \left\{ j' \in \{1, \ldots, n_\ell\} : \pi(ij'|\ell) = \pi(ij|\ell) \right\}$ each represent sets of group indices that share a common ancestry with group $i$ or $j$ at level $\ell$.

**Local variances** First rewrite Eq. 6 as $\Psi_{\pi(ij|\ell)} = \lambda_{\pi(ij|\ell)}^2 \prod_{s=\ell+1}^{L-1} \lambda_{\pi(ij|s)}^2 = \lambda_{\pi(ij|\ell)}^2 \Psi_{\pi(ij|\ell+1)}$ and so at level $\ell$, all $\lambda_{\pi(ij|\ell)}^2$ terms need to be sampled from their

full conditional distribution given $\Psi_{\pi(ij|\ell+1)}$. Here we focus on the case where $\lambda_{\pi(ij|\ell)} \sim \mathrm{C}^+(0, 1)$, which corresponds to horseshoe shrinkage. We exploit the following scale mixture representation of the half-Cauchy prior (Makalic & Schmidt, 2015): if $\lambda^2|\zeta \sim \mathrm{IG}(1/2, 1/\zeta)$ and $\zeta \sim \mathrm{IG}(1/2, 1)$, then $\lambda \sim \mathrm{C}^+(0, 1)$.

$$\lambda^2_{\pi(ij|\ell)}|\zeta_{\pi(ij|\ell)}, \text{else} \sim \mathrm{IG}\left(\frac{1}{2} + \frac{1}{2}\sum_{s\leq\ell} n_s^2, \frac{1}{\zeta_{\pi(ij|\ell)}} + \frac{(\theta_{\pi(ij|\ell)} - \theta_{\pi(ij|\ell+1)})^2}{2\tau_\ell^2 \Psi_{\pi(ij|\ell+1)}} + \frac{1}{2}\sum_{s<\ell}\sum_{a\in\mathcal{C}_i^s}\sum_{b\in\mathcal{C}_j^s} \frac{(\theta_{\pi(ab|s)} - \theta_{\pi(ab|s+1)})^2}{\tau_s^2 \Psi_{\pi(ab|s+1)}}\right)$$

$$\zeta_{\pi(ij|\ell)}|\lambda^2_{\pi(ij|\ell)} \sim \mathrm{IG}\left(1, 1 + \frac{1}{\lambda^2_{\pi(ij|\ell)}}\right)$$

**Global variances** We place a half-Cauchy prior on $\tau$ and use the same scale mixture representation outlined above for the local variances.

$$\tau_\ell^2|\zeta_\ell, \text{else} \sim \mathrm{IG}\left(\frac{n_\ell^2 + 1}{2}, \frac{1}{\zeta_\ell} + \frac{1}{2}\sum_{a=1}^{n_\ell^2}\sum_{b=1}^{n_\ell^2} \frac{\left(\theta_{\pi(ab|\ell)} - \theta_{\pi(ab|\ell+1)}\right)^2}{\Psi_{\pi(ab|\ell)}}\right)$$

$$\zeta_\ell|\tau_\ell^2 \sim \mathrm{IG}\left(1, 1 + \frac{1}{\tau_\ell^2}\right)$$

### A.2: Fast sampling algorithm

We show that our version of the Bhattacharya et al. (2016) fast sampling algorithm, as outlined in Section 4.2, produces a draw of $\beta$ with the correct posterior mean and covariance matrix:

$$\begin{aligned}\mathbb{E}(\beta|\text{data}) &= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}(\tilde{\mathrm{X}}'\tilde{y} + \Lambda_*^{-1}\bar{\beta}(\theta)) \\ \mathrm{Cov}(\beta|\text{data}) &= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\end{aligned}$$

where $\bar{\beta}(\theta)$ denotes the prior mean of $\beta$ which depends on higher-level elasticity parameters $\theta$. Note that our proofs below rely on the Woodbury matrix identity: $(A +$

$$UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

$$\begin{aligned}
\mathbb{E}(\beta|\text{data}) &= \mathbb{E}(u) + \Lambda_*\tilde{\mathrm{X}}'\mathbb{E}(w) \\
&= \mathbb{E}(u) + \Lambda_*\tilde{\mathrm{X}}'\mathbb{E}\Big((\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}(\tilde{y} - v)\Big) \\
&= \mathbb{E}(u) + \Lambda_*\tilde{\mathrm{X}}'(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}(\tilde{y} - \mathbb{E}(v)) \\
&= \mathbb{E}(u) + \Lambda_*\tilde{\mathrm{X}}'(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}(\tilde{y} - \tilde{\mathrm{X}}\mathbb{E}(u) - \mathbb{E}(\delta)) \\
&= \Lambda_*\tilde{\mathrm{X}}'(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}\tilde{y} + \bar{\beta}(\theta) - \Lambda_*\tilde{\mathrm{X}}'(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}\tilde{\mathrm{X}}\bar{\beta}(\theta) \\
&= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\tilde{\mathrm{X}}'\tilde{y} \\
&\quad + \Big(I - \Lambda_*\tilde{\mathrm{X}}'(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}\tilde{\mathrm{X}}\Big)\bar{\beta}(\theta) \text{ by Woodbury} \\
&= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\tilde{\mathrm{X}}'\tilde{y} + \Big(I - \Lambda_*\tilde{\mathrm{X}}'(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}\tilde{\mathrm{X}}\Big)\Lambda_*\Lambda_*^{-1}\bar{\beta}(\theta) \\
&= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\tilde{\mathrm{X}}'\tilde{y} + \Big(\Lambda_* - \Lambda_*\tilde{\mathrm{X}}'(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}\tilde{\mathrm{X}}\Lambda_*\Big)\Lambda_*^{-1}\bar{\beta}(\theta) \\
&= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\tilde{\mathrm{X}}'\tilde{y} + (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})\Lambda_*^{-1}\bar{\beta}(\theta) \text{ by Woodbury} \\
&= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}(\tilde{\mathrm{X}}'\tilde{y} + \Lambda_*^{-1}\bar{\beta}(\theta)) \\
\text{Cov}(\beta|\text{data}) &= \text{Cov}(u + \Lambda_*'\tilde{\mathrm{X}}w) \\
&= \text{Cov}(u + \Lambda_*'\tilde{\mathrm{X}}\Sigma(\tilde{y} - v)) \\
&= \text{Cov}(u - \Lambda_*'\tilde{\mathrm{X}}\Sigma v) \\
&= \text{Var}(u) + \text{Var}(\Lambda_*'\tilde{\mathrm{X}}\Sigma v) - \text{Cov}(u, \Lambda_*'\tilde{\mathrm{X}}\Sigma v) - \text{Cov}(\Lambda_*'\tilde{\mathrm{X}}\Sigma v, u) \\
&= \text{Var}(u) + \Lambda_*'\tilde{\mathrm{X}}\Sigma\text{Var}(v)\Sigma\tilde{\mathrm{X}}'\Lambda_* \\
&\quad -\text{Cov}(u, v)\Sigma\tilde{\mathrm{X}}'\Lambda_* - \Lambda_*'\tilde{\mathrm{X}}\Sigma\text{Cov}(v, u) \\
&= \Lambda_* + \Lambda_*'\tilde{\mathrm{X}}\Sigma\Sigma^{-1}\Sigma\tilde{\mathrm{X}}'\Lambda_* - \Lambda_*'\tilde{\mathrm{X}}\Sigma\tilde{\mathrm{X}}'\Lambda_* - \Lambda_*'\tilde{\mathrm{X}}\Sigma\tilde{\mathrm{X}}'\Lambda_* \\
&= \Lambda_* - \Lambda_*'\tilde{\mathrm{X}}(\tilde{\mathrm{X}}\Lambda_*\tilde{\mathrm{X}}' + I)^{-1}\tilde{\mathrm{X}}'\Lambda_* \\
&= (\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1} \text{ by Woodbury}
\end{aligned}$$

## Appendix B: Additional simulation experiments

### B.1: Parameter recovery

The goal of this section is to show that the MCMC algorithms outlined in Section 4 can accurately recover model parameters at each level of the tree. We focus on one of the 25 data sets generated by specification (I) with $n = 50$ and $p = 100$ (for details, see the discussion in Section 5). The tree corresponding to this data set has 5 categories, 10 subcategories, and 100 products. There are then $5^2 = 25$ category elasticities $\theta_{\pi(ij|2)}$, $10^2 = 100$ subcategory elasticities $\theta_{\pi(ij|1)}$, and $100^2 = 10,000$ product elasticities $\beta_{ij}$.

For simplicity, we report estimates from two models: the ($\theta$-Ridge, $\beta$-Ridge) hierarchical prior and the standard ($\beta$-Ridge) prior. Figure 7 plots the posterior mean of each elasticity parameter against its true value. The vertical lines correspond to 95% credible intervals. The first three figures report estimates from the hierarchical shrinkage prior and the last figure reports estimates from the standard shrinkage prior. We find that all 95% credible intervals cover true values for the category and subcategory-level elasticities, and nearly all credible intervals cover true values for product-level elasticities.

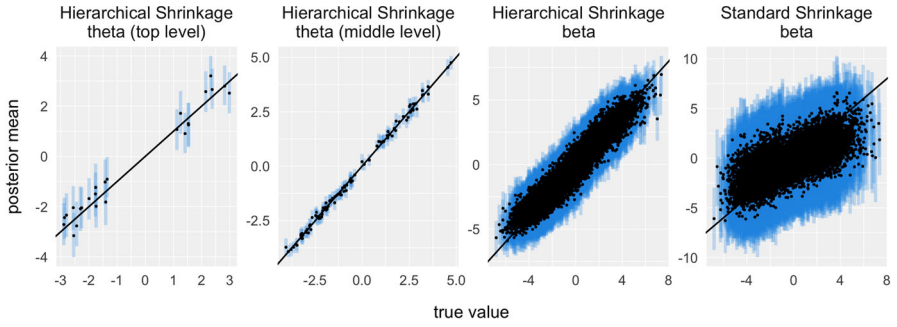



**Fig. 7** Posterior means and 95% credible intervals of elasticity parameters are plotted against true values. The first three figures report estimates from the hierarchical ($\theta$-Ridge, $\beta$-Ridge) prior and the last figure reports estimates from the standard ($\beta$-Ridge) model. The data are generated from specification (I) as described in Section 5 with $n = 50$ and $p = 100$

The benefits of hierarchical shrinkage can be seen by comparing the estimates of product-level elasticities $\beta_{ij}$ in the last two figures. With high-dimensional $p > n$ data, the design matrix will be rank deficient and the regularization imposed by the prior will bite. The last figure on the right shows that default prior assumptions can produce very noisy and biased estimates, especially for elasticities that are larger in magnitude. In comparison, our hierarchical priors shrink noisy estimates towards more reasonable values by pooling information across products. Overall, this further emphasizes the importances of judicious prior choices in high-dimensional settings.

### B.2: Tree misspecification

The simulation experiments in Section 5 assume that the tree used to construct the hierarchical shrinkage prior matches the "true" tree used to generate the data. When working with observational data, the existence or structure of a true tree is not known so there may be a concern of tree misspecification. In the context of scanner data, the reported product classification tree could be misspecified if it does not reflect consumers' true boundaries of substitution across product groups. For example, in the IRI scanner data set "Ready-to-Eat Cereal" is the only subcategory within the CEREAL category. This part of the tree may be too coarse if there is a more granular partitioning of cereal brands, flavors, and qualities that better aligns with consumer cross-product substitution.

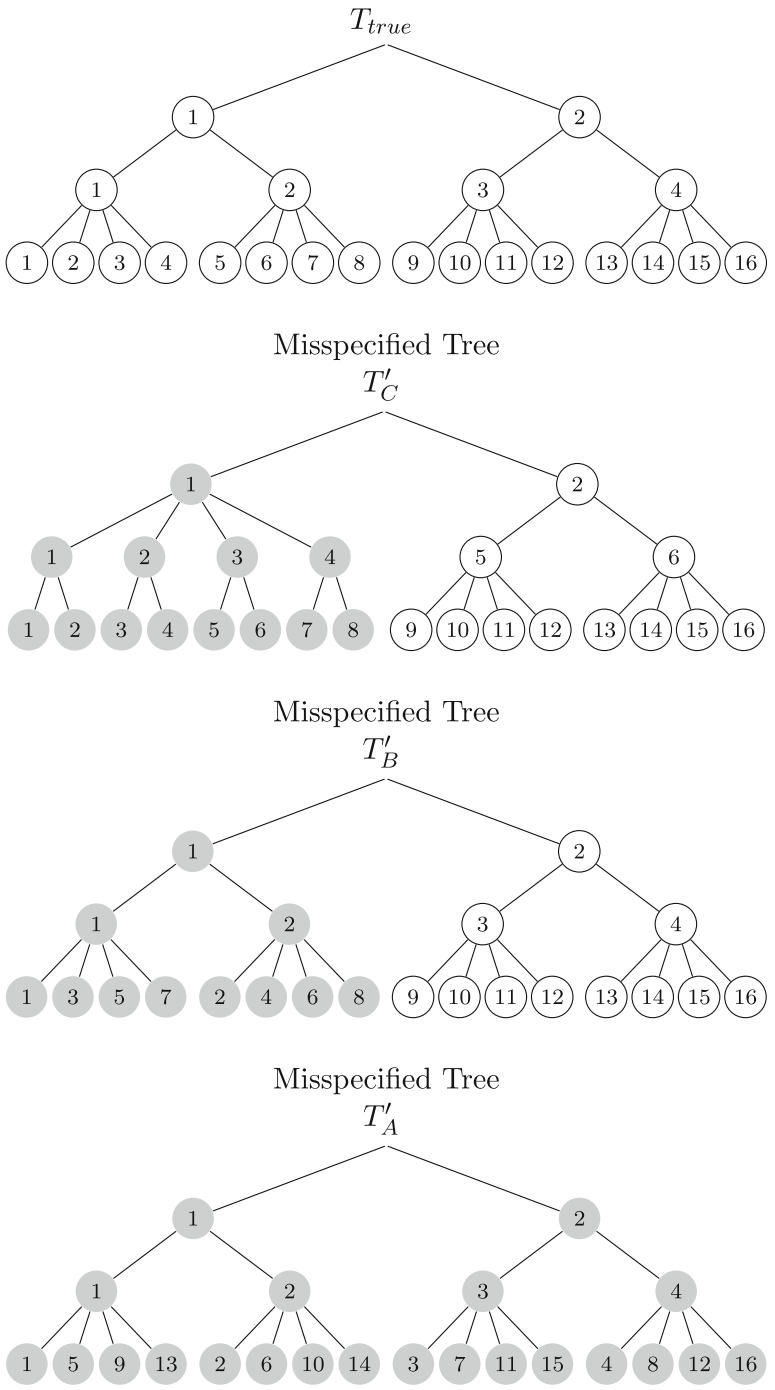



**Fig. 8** A minimal example with $p = 16$ highlighting the three forms of misspecification considered in our simulation experiments. Shaded circles indicate misspecification relative to $T_{true}$

In this section we use the same 25 ($n = 50$, $p = 100$) simulated data sets from specification (I) in Section 5 to explore consequences of tree misspecification. Let $T_{true}$ denote the tree used to simulate the data. $T_{true}$ has five equal-sized categories on the top level (each having two child subcategories), 10 equal-sized subcategories on the middle level (each having 10 child products), and 100 products on the bottom level. In addition to standard shrinkage priors and hierarchical priors based on $T_{true}$, we also fit hierarchical priors based on the following three misspecified trees.

A) In $T'_A$, the ordering of group nodes on the top and middle levels are kept the same, but the bottom-level nodes (the "products") are permuted. In $T_{true}$, products fill up one subcategory at a time—i.e., goods $1, \ldots, 10$ are assigned to subcategory 1, goods $11, \ldots, 20$ are assigned to subcategory 2, and so on. We generate $T'_A$ by assigning the first product to subcategory 1, the second product to subcategory 2, and so on until each subcategory has one product. This process repeats until all products have been assigned into a subcategory. The resulting tree shares no common structure with $T_{true}$.

B) In $T'_B$, half of the tree is misspecified (in the same way as $T'_A$) while the other half matches $T_{true}$.

C) In $T'_C$, half of the tree is misspecified as an overly granular version of $T_{true}$ (each group is split into two) while the other half matches $T_{true}$.

Examples of each form of misspecification are provided in Fig. 8 and simulation results are reported in Table 8. For each misspecified tree, we also report the distance between $T'$ and $T_{true}$ using the adjusted Rand index (ARI). Similarity is maximized when $ARI(T', T_{true}) = 1$. Our results indicate that hierarchical shrinkage can still provide improved estimates over standard shrinkage methods—and is never worse than standard shrinkage methods—even when the tree is misspecified. However, the magnitude of these gains vanishes as the degree of misspecification grows. If an analyst is concerned with the threat of misspecification, then different tree structures could be fit to the data and compared based on out-of-sample fit metrics.

**Table 8** Effects of tree misspecification

| | Hierarchical Shrinkage with $T'_A$ | Estimation RMSE | | | Correct Signs | | |
|---|---|---|---|---|---|---|---|
| A) | $ARI(T'_A, T_{true}) = -0.02$ | (I) | (II) | (III) | (I) | (II) | (III) |
| | $\theta$-Ridge, $\beta$-Ridge | **2.482** | **2.340** | 1.207 | 0.77 | 0.79 | 0.88 |
| | $\theta$-Ridge, $\beta$-Lasso | 2.552 | 2.429 | 1.066 | 0.77 | 0.77 | 0.89 |
| | $\theta$-Ridge, $\beta$-Horseshoe | 2.800 | 2.697 | **0.951** | 0.73 | 0.73 | 0.89 |
| | $\theta$-Horseshoe, $\beta$-Ridge | 2.533 | 2.399 | 1.220 | 0.76 | 0.77 | 0.86 |
| | $\theta$-Horseshoe, $\beta$-Lasso | 2.587 | 2.468 | 1.088 | 0.76 | 0.76 | 0.88 |
| | $\theta$-Horseshoe, $\beta$-Horseshoe | 2.815 | 2.713 | 0.962 | 0.72 | 0.72 | 0.88 |

**Table 8** (continued)

| | Hierarchical Shrinkage with $T'_B$ | Estimation RMSE | | | Correct Signs | | |
|---|---|---|---|---|---|---|---|
| B) | $ARI(T'_B, T_{true}) = 0.55$ | (I) | (II) | (III) | (I) | (II) | (III) |
| | $\theta$-Ridge, $\beta$-Ridge | 1.808 | 1.548 | 0.950 | 0.87 | 0.89 | 0.91 |
| | $\theta$-Ridge, $\beta$-Lasso | 1.820 | 1.538 | 0.866 | 0.86 | 0.89 | 0.92 |
| | $\theta$-Ridge, $\beta$-Horseshoe | 2.016 | 1.738 | 0.783 | 0.84 | 0.86 | 0.92 |
| | $\theta$-Horseshoe, $\beta$-Ridge | **1.704** | **1.340** | 0.711 | 0.87 | 0.91 | 0.93 |
| | $\theta$-Horseshoe, $\beta$-Lasso | 1.765 | 1.420 | 0.678 | 0.86 | 0.90 | 0.94 |
| | $\theta$-Horseshoe, $\beta$-Horseshoe | 1.958 | 1.645 | **0.668** | 0.84 | 0.87 | 0.93 |
| | Hierarchical Shrinkage with $T'_C$ | | | | | | |
| C) | $ARI(T'_C, T_{true}) = 0.63$ | (I) | (II) | (III) | (I) | (II) | (III) |
| | $\theta$-Ridge, $\beta$-Ridge | **1.047** | 0.537 | 0.522 | 0.93 | 0.97 | 0.94 |
| | $\theta$-Ridge, $\beta$-Lasso | 1.060 | 0.471 | 0.459 | 0.93 | 0.98 | 0.95 |
| | $\theta$-Ridge, $\beta$-Horseshoe | 1.138 | **0.394** | 0.408 | 0.92 | 0.98 | 0.95 |
| | $\theta$-Horseshoe, $\beta$-Ridge | 1.064 | 0.529 | **0.193** | 0.93 | 0.97 | 0.98 |
| | $\theta$-Horseshoe, $\beta$-Lasso | 1.073 | 0.469 | 0.194 | 0.93 | 0.98 | 0.98 |
| | $\theta$-Horseshoe, $\beta$-Horseshoe | 1.147 | 0.394 | 0.261 | 0.92 | 0.98 | 0.97 |
| | Hierarchical Shrinkage with $T_{true}$ | | | | | | |
| D) | $ARI(T_{true}, T_{true}) = 1$ | (I) | (II) | (III) | (I) | (II) | (III) |
| | $\theta$-Ridge, $\beta$-Ridge | **1.033** | 0.530 | 0.515 | 0.94 | 0.97 | 0.94 |
| | $\theta$-Ridge, $\beta$-Lasso | 1.047 | 0.461 | 0.451 | 0.93 | 0.98 | 0.95 |
| | $\theta$-Ridge, $\beta$-Horseshoe | 1.126 | **0.385** | 0.407 | 0.92 | 0.98 | 0.95 |
| | $\theta$-Horseshoe, $\beta$-Ridge | 1.044 | 0.526 | **0.190** | 0.93 | 0.98 | 0.98 |
| | $\theta$-Horseshoe, $\beta$-Lasso | 1.053 | 0.461 | 0.192 | 0.93 | 0.98 | 0.98 |
| | $\theta$-Horseshoe, $\beta$-Horseshoe | 1.130 | 0.386 | 0.270 | 0.92 | 0.98 | 0.97 |
| E) | Standard Shrinkage | (I) | (II) | (III) | (I) | (II) | (III) |
| | $\beta$-Ridge | **2.541** | **2.375** | 1.219 | 0.77 | 0.78 | 0.87 |
| | $\beta$-Lasso | 2.622 | 2.473 | 1.079 | 0.76 | 0.77 | 0.89 |
| | $\beta$-Horseshoe | 2.879 | 2.745 | **0.969** | 0.73 | 0.74 | 0.88 |

The table reports average fit statistics across 25 simulated data sets from specification (I) in Section 5 with $n = 50$ and $p = 100$. The first three panels show the estimation results from hierarchical priors with misspecified trees. In each case, we use the adjusted Rand index (ARI) to report the similarity between the misspecified tree $T'$ and $T_{true}$. The two benchmark cases from Section 5 (i.e., hierarchical priors based on $T_{true}$ and standard shrinkage priors) are shown in the last two panels

## Appendix C: Additional empirical details and results

### C.1: Summary of Priors and Hyperparameter Specifications

Standard shrinkage priors

- Cross elasticities

$$\beta_{ij} \sim \mathrm{N}\left(0, \lambda_{ij}^2 \tau_{\text{cross}}^2\right), \quad \tau_{\text{cross}} \sim \mathrm{C}^+(0, 1)$$

- Own elasticities

$$\beta_{ii} \sim \mathrm{N}\left(0, \lambda_{ii}^2 \tau_{\text{own}}^2\right), \quad \tau_{\text{own}} \sim \mathrm{C}^+(0, 1)$$

- Control coefficient vector (product intercepts, seasonal variables, promotional variables)

$$\phi_i \sim \mathrm{N}(0, 10^2 I_d)$$

- Observational error variance

$$\sigma_i^2 \sim \mathrm{IG}(5, 5)$$

Hierarchical shrinkage priors

- Cross elasticities

$$\begin{aligned}
\beta_{ij} &\sim \mathrm{N}\left(\theta_{\pi(ij|1)}, \Psi_{ij}\tau_{\beta\text{cross}}^2\right), & \tau_{\beta\text{cross}} &\sim \mathrm{C}^+(0, 1)\\
\theta_{\pi(ij|1)} &\sim \mathrm{N}\left(\theta_{\pi(ij|2)}, \Psi_{\pi(ij|1)}\tau_{1\text{cross}}^2\right), & \tau_{1\text{cross}} &\sim \mathrm{C}^+(0, 1)\\
\theta_{\pi(ij|2)} &\sim \mathrm{N}\left(\bar{\theta}_{\text{own}}, \Psi_{\pi(ij|2)}\tau_{2\text{cross}}^2\right), & \tau_{2\text{cross}} &\sim \mathrm{C}^+(0, 1)\\
\bar{\theta}_{\text{cross}} &\sim \mathrm{N}(0, 1)
\end{aligned}$$

- Own elasticities

$$\begin{aligned}
\beta_{ii} &\sim \mathrm{N}\left(\theta_{\pi(ii|1)}, \Psi_{ii}\tau_{\beta\text{own}}^2\right), & \tau_{\beta\text{own}} &\sim \mathrm{C}^+(0, 1)\\
\theta_{\pi(ii|1)} &\sim \mathrm{N}\left(\theta_{\pi(ii|2)}, \Psi_{\pi(ii|1)}\tau_{1\text{own}}^2\right), & \tau_{1\text{own}} &\sim \mathrm{C}^+(0, 1)\\
\theta_{\pi(ii|2)} &\sim \mathrm{N}\left(\bar{\theta}_{\text{cross}}, \Psi_{\pi(ii|2)}\tau_{2\text{own}}^2\right), & \tau_{2\text{own}} &\sim \mathrm{C}^+(0, 1)\\
\bar{\theta}_{\text{own}} &\sim \mathrm{N}(0, 1)
\end{aligned}$$

- Control coefficient vector (product intercepts, seasonal variables, promotional variables)

$$\phi_i \sim \mathrm{N}(0, 10^2 I_d)$$

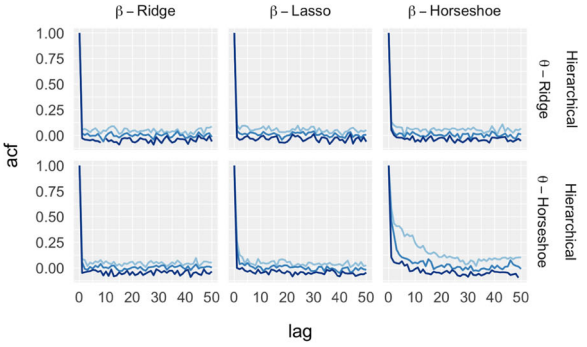
- Observational error variance

$$\sigma_i^2 \sim \mathrm{IG}(5, 5)$$

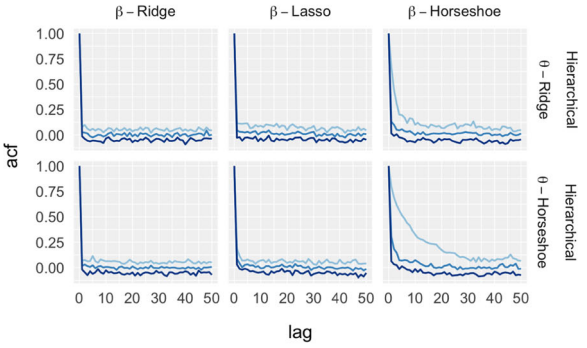
**C.2: Mixing diagnostics**

**Fig. 9** Percentiles of the autocorrelation function across own-price elasticity parameters

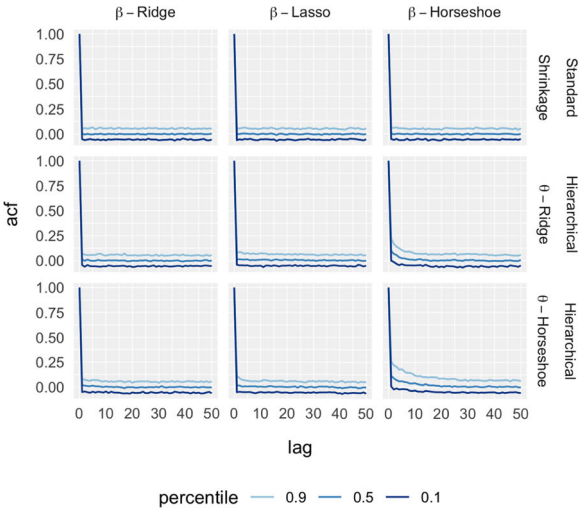
(a) Category-Level Own-Price Elasticities $\theta_{\pi(ii|2)}$




(b) Subcategory-Level Own-Price Elasticities $\theta_{\pi(ii|1)}$




(c) Product-Level Own-Price Elasticities $\beta_{ii}$



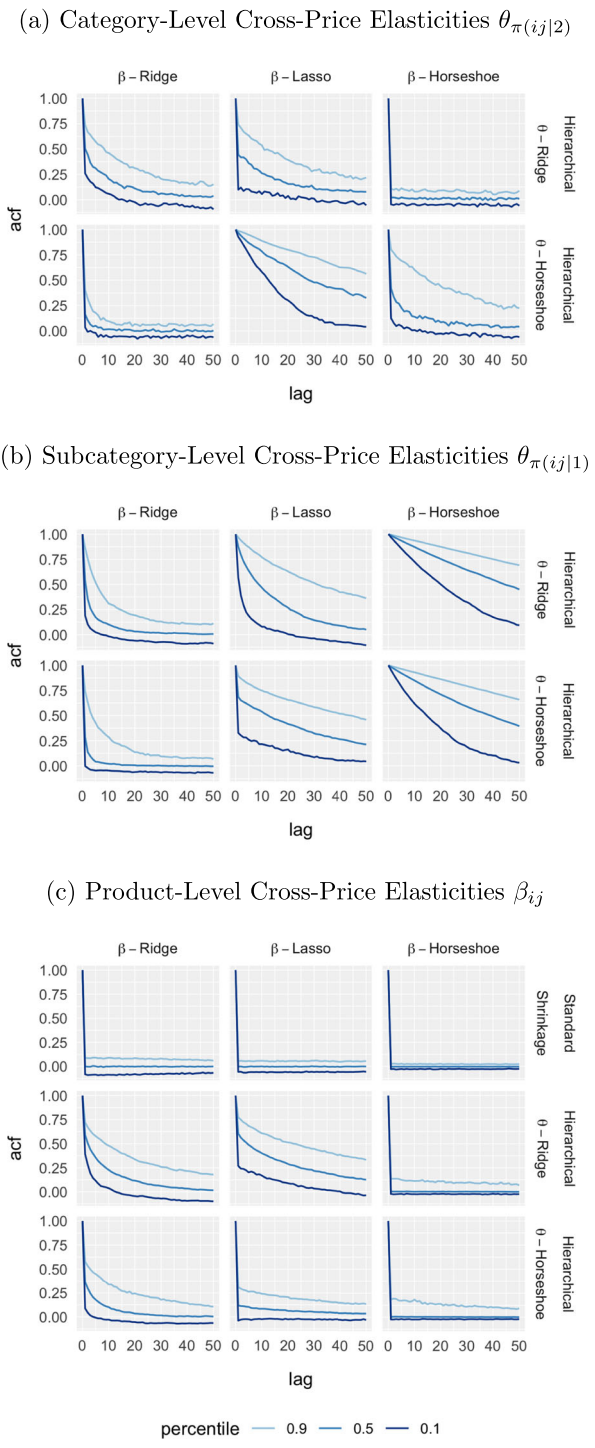



**Fig. 10** Percentiles of the autocorrelation function across cross-price elasticity parameters. In panel (c), we randomly sample 1,000 of the 75,625 total elasticity parameters

### C.3: Distributions of Price Elasticity and Promotional Effect Estimates

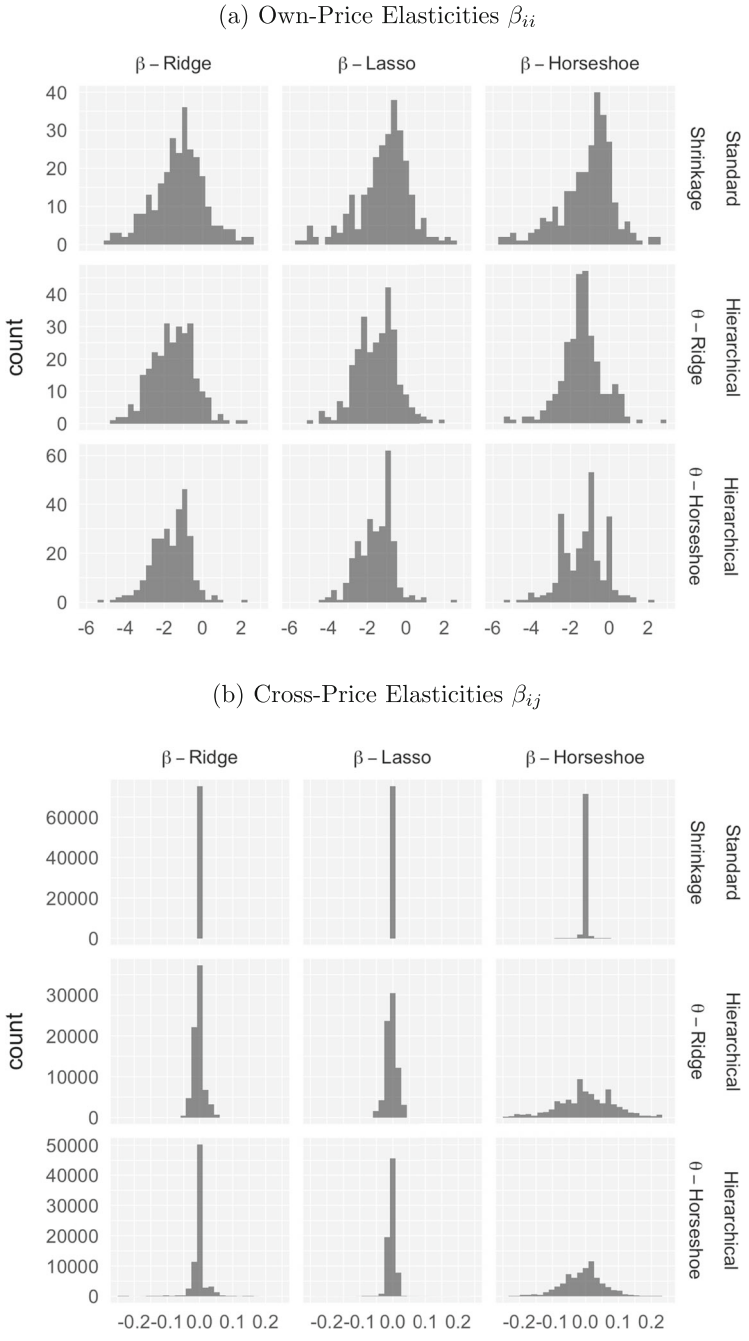
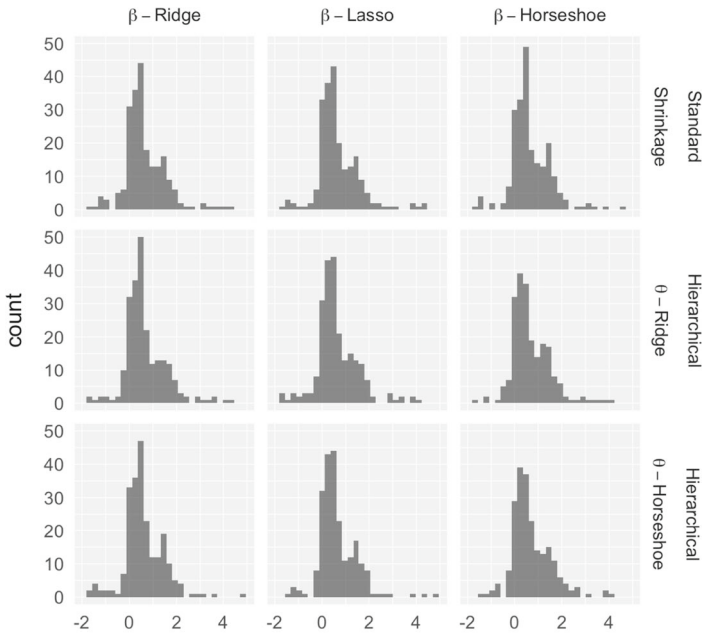



**Fig. 11** Distributions of product-level price elasticity estimates

(a) Display Promotion Effects




(b) Feature Promotion Effects

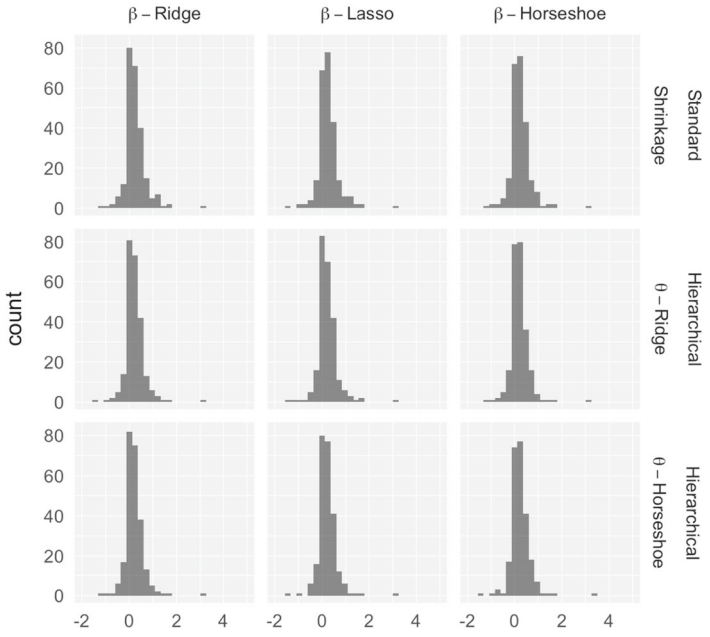



**Fig. 12** Distributions of product-level own-promotion effect estimates

### C.4: Shrinkage Factors by Category

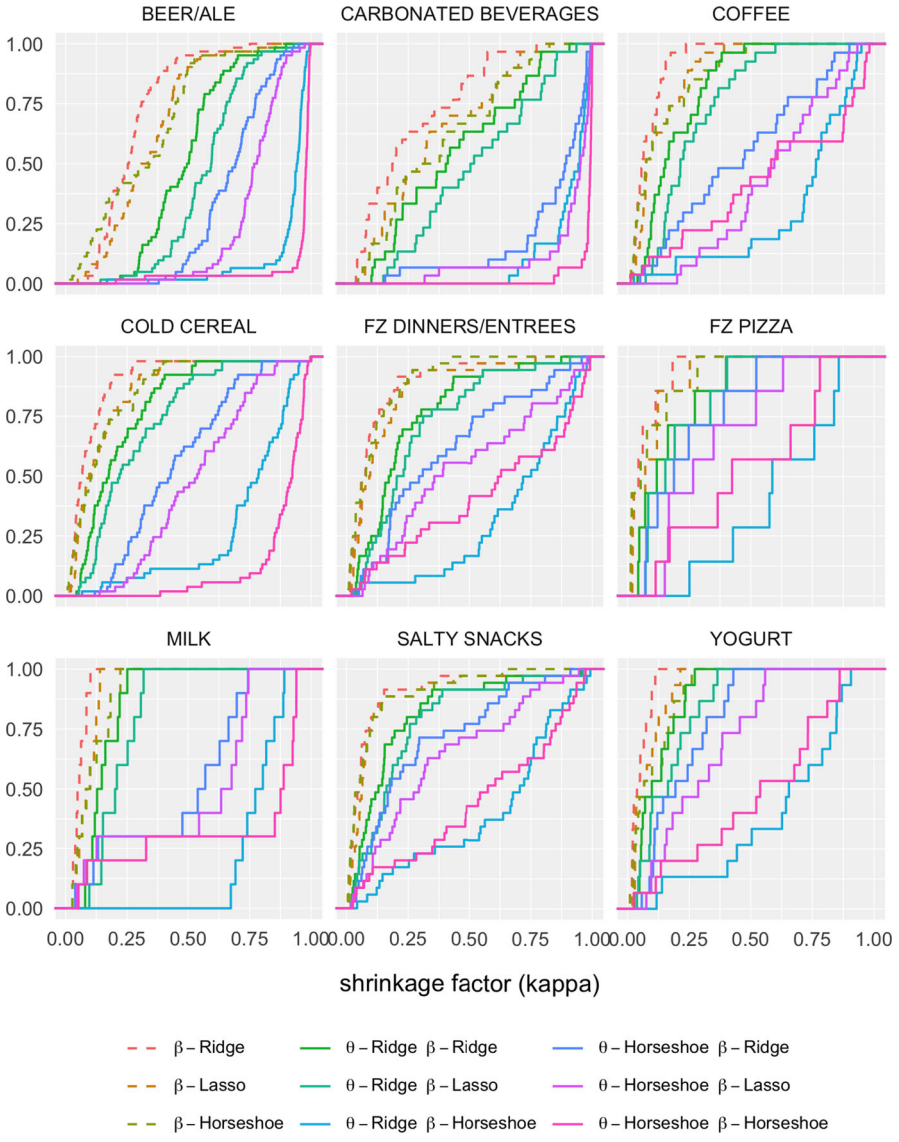



**Fig. 13** Empirical CDFs of own-price elasticity shrinkage factor estimates median($\kappa_{ii}$|data) plotted across models and product categories. There is maximum shrinkage when $\kappa_{ii} = 1$ and no shrinkage when $\kappa_{ii} = 0$

## C.5: Competitive maps by category

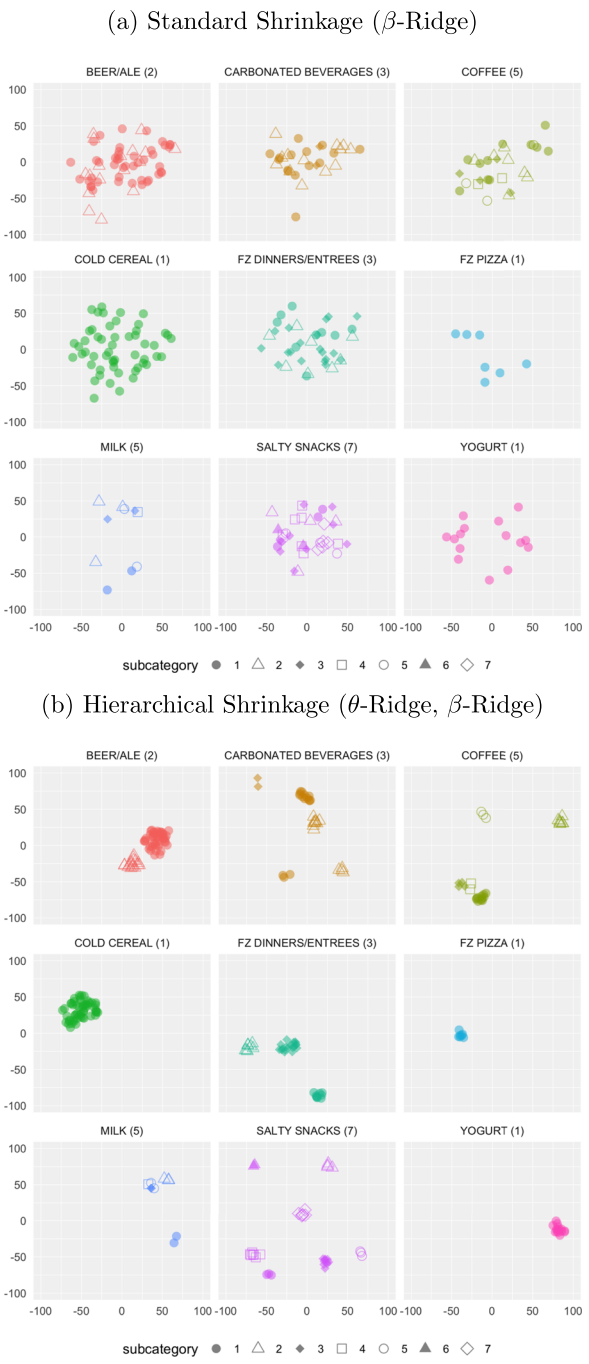
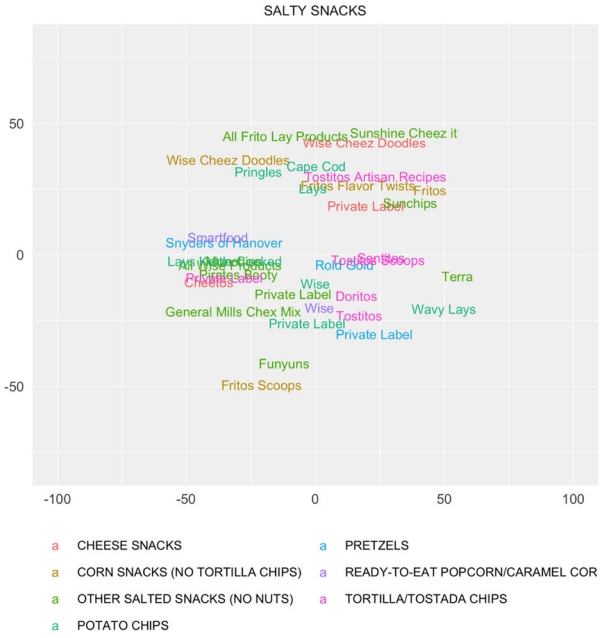



**Fig. 14** Competitive maps based on t-SNE projections of the $275 \times 275$ price elasticity matrix separated by category for two selected models. Points are assigned the same colors as in Fig. 6 and are also assigned separate shapes based on subcategory labels (see Table 3). The number of subcategories within a category is listed in parentheses next to the category name

(a) Standard Shrinkage ($\beta$-Ridge)




(b) Hierarchical Shrinkage ($\theta$-Ridge, $\beta$-Ridge)

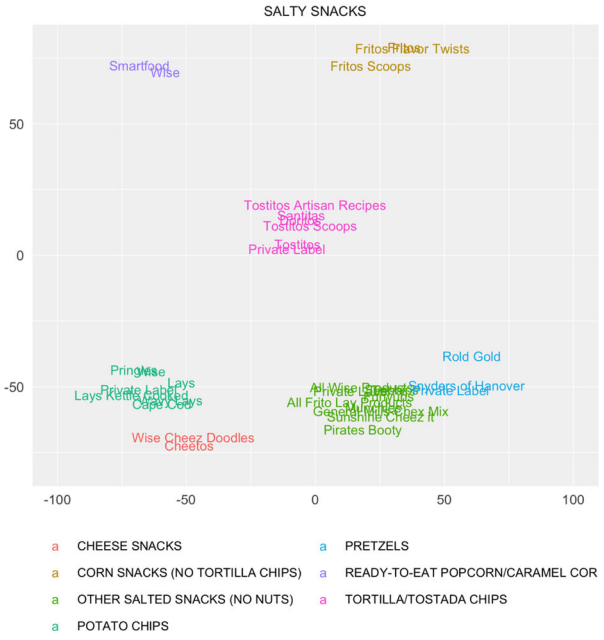



**Fig. 15** Competitive maps for the SALTY SNACK category from two selected models. This is a zoomed-in, product-level view of Fig. 6. Each point is a product (brand) and is colored by its subcategory label

## Declarations

**Conflict of Interests** The authors have no competing interests to declare that are relevant to the content of this article.



## References

Allenby, G. M. (1989). A unified approach to identifying, estimating and testing demand structures with aggregate scanner data. *Marketing Science*, *8*(3), 265–280.

Amano, T., Rhodes, A., & Seiler, S (2019). Large-scale demand estimation with search data. Working Paper.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Demand estimation with machine learning and model combination. NBER Working Paper No. 20955.

Barndorff-Nielsen, O. E., Kent, J. T., & Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review*, *50*, 145–159.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2016). Default bayesian analysis with global-local shrinkage priors. *Biometrika*, *103*(4), 955–969.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B.T. (2019). Lasso meets horseshoe: A survey. *Statistical Science*, *34*(3), 405–427.

Bhattacharya, A., Chakraborty, A., & Mallick, B.K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, *103*(4), 985–991.

Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, *41*(3), 1111–1141.

Bils, M., & Klenow, P. J. (2004). Some evidence on the importance of sticky prices. *Journal of Political Economy*, *112*(5), 947–985.

Bingham, N. H., Goldie, C. M., & Teugels, J.L. (1987). *Regular Variation*. Cambridge: Cambridge University Press.

Blattberg, R. C., & George, E. I. (1991). Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *Journal of the American Statistical Association*, *86*(414), 304–315.

Boatwright, P., McCulloch, R., & Rossi, P. (1999). Account-level modeling for trade promotion: an application of a constrained parameter hierarchical model. *Journal of the American Statistical Association*, *94*(448), 1063–1073.

Bronnenberg, B. J., Kruger, M. W., & Mela, C.F. (2008). Database paper—The IRI marketing data set. *Marketing Science*, *27*(4), 745–748.

Carvalho, C. M., Polson, N. G., & Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Chen, F., Liu, X., Proserpio, D., & Troncoso, I. (2020). Product2Vec: Understanding product-level competition using representation learning. Working Paper.

Chintagunta, P. K. (1998). Inertia and variety seeking in a model of brand-purchase timing. *Marketing Science*, *17*(3), 253–270.

Chiong, K. X., & Shum, M. (2018). Random projection estimation of discrete-choice models with large choice sets. *Management Science*, *65*(1), 256–271.

Cline, D. B. H. (1987). Convolutions of distributions with exponential and subexponential tails. *Journal of the Australian Mathematics Society: Series A*, *43*, 347–365.

Datta, J., & Ghosh, J. K. (2013). Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Analysis*, *8*(1), 111–132.
Deaton, A., & Muellbauer, J. (1980). *Economics and consumer behavior*. Cambridge: Cambridge University Press.
DellaVigna, S., & Gentzkow, M. (2019). Uniform pricing in US retail chains. *The Quarterly Journal of Economics*, *134*(4), 2011–2084.
Diewert, W. E. (1974). Applications of duality theory. In *Frontiers of Qumtitatiue Economics*. Amsterdam: North-Holland Publishing Company.
Donnelly, R., Ruiz, F. R., Blei, D., & Athey, S. (2021). Counterfactual inference for consumer choice across many product categories. *Quantitative Marketing and Economics*, *19*(3), 369–407.
Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18.
Elrod, T. (1988). Choice map: Inferring a product-market map from panel data. *Marketing Science*, *7*(1), 21–40.
Ershov, D., Laliberté, J.-W., Marcoux, M., & Orr, S. (2021). Estimating complementarity with large choice sets: An application to mergers. Working Paper.
Fessler, P., & Kasy, M. (2019). How to use economic theory to improve estimators: Shrinking toward theoretical restrictions. *Review of Economics and Statistics*, *101*(4), 681–698.
France, S. L., & Ghose, S. (2016). An analysis and visualization methodology for identifying and testing market structure. *Marketing Science*, *35*(1), 182–197.
Gabel, S., Guhl, D., & Klapper, D. (2019). P2V-MAP: Mapping market structures for large retail assortments. *Journal of Marketing Research*, *56*(4), 557–580.
Gabel, S., & Timoshenko, A. (2021). Product choice with large assortments: a scalable deep-learning model. *Management Science*, *68*(3), 1591–2376.
Gentzkow, M. (2007). Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review*, *97*(3), 713–744.
George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.
Goldman, S. M., & Uzawa, H. (1964). A note on separability in demand analysis. *Econometrica*, *32*(3), 387–398.
Gorman, W. (1971). Two stage budgeting. Unpublished Manuscript, London School of Economics.
Griffin, J., & Brown, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, *12*(1), 135–159.
Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845.
Hausman, J., Leonard, G., & Zona, J.D. (1994). Competitive analysis with differentiated products. *Annales d'Economie et de Statistique*, *34*, 159–180.
Hitsch, G. J., Hortaçsu, A., & Lin, X. (2021). Prices and promotions in US retail markets. *Quantitative Marketing and Economics*, *19*(3), 289–368.
Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.
Jacobs, B., Donkers, B., & Fok, D. (2016). Model-based purchase predictions for large assortments. *Marketing Science*, *35*(3), 389–404.
Jacobs, B., Fok, D., & Donkers, B. (2021). Understanding large-scale dynamic purchase behavior. *Marketing Science*, *40*(5), 844–870.
Kim, J. B., Albuquerque, P., & Bronnenberg, B.J. (2011). Mapping online consumer search. *Journal of Marketing Research*, *48*(1), 13–27.
Kumar, M., Eckles, D., & Aral, S (2020). Scalable bundling via dense product embeddings. Working Paper.
Li, Y., Craig, B. A., & Bhadra, A. (2019). The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, *28*(3), 747–757.
Li, Y., Datta, J., Craig, B. A., & Bhadra, A. (2021). Joint mean-covariance estimation via the horseshoe. *Journal of Multivariate Analysis*, *183*, 104716.
Makalic, E., & Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, *23*(1), 179–182.

Manchanda, P., Ansari, A., & Gupta, S. (1999). The "shopping basket": a model for multicategory purchase incidence decisions. *Marketing Science*, *18*(2), 95–114.
Mehta, N. (2007). Investigating consumers' purchase incidence and brand choice decisions across multiple product categories: a theoretical and empirical analysis. *Marketing Science*, *26*(2), 196–217.
Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.
Montgomery, A. L. (1997). Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science*, *16*(4), 315–337.
Montgomery, A. L., & Rossi, P. E. (1999). Estimating price elasticities with theory-based priors. *Journal of Marketing Research*, *36*(4), 413–423.
Morozov, I. (2020). Measuring benefits from new products in markets with information frictions. Working Paper.
Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521–543.
Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.
Pollak, R. A., & Wales, T. J. (1992). *Demand system specification and estimation*. Oxford: Oxford University Press, Inc.
Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, & M. West (Eds.) *Bayesian Statistics*, Vol. 9. Oxford: Oxford University Press.
Polson, N. G., & Scott, J. G. (2012a). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(2), 287–311.
Polson, N. G., & Scott, J. G. (2012b). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902.
Ringel, D. M., & Skiera, B. (2016). Visualizing asymmetric competition among more than 1,000 products using big search data. *Marketing Science*, *35*(3), 511–534.
Rossi, P. E. (2014). Even the rich can make themselves poor: a critical examination of IV methods in marketing applications. *Marketing Science*, *33*(5), 655–672.
Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. New York: Wiley.
Ruiz, F. J. R., Athey, S., & Blei, D.M. (2020). SHOPPER: A probabilistic model of consumer choice with substitutes and complements. *The Annals of Applied Statistics*, *14*(1), 1–27.
Rutz, O. J., & Sonnier, G. P. (2011). The evolution of internal market structure. *Marketing Science*, *30*(2), 274–289.
Semenova, V., Goldman, M., Chernozhukov, V., & Taddy, M. (2021). Estimation and inference about heterogeneous treatment effects in high-dimensional dynamic panels. Working Paper.
Sinha, A., Sahgal, A., & Mathur, S.K. (2013). Practice prize paper—Category optimizer: A dynamic-assortment, new-product-introduction, mix-optimization, and demand-planning system. *Marketing Science*, *32*(2), 221–228.
Smith, A. N., Rossi, P. E., & Allenby, G.M. (2019). Inference for product competition and separable demand. *Marketing Science*, *38*(4), 690–710.
Song, I., & Chintagunta, P. K. (2006). Measuring cross-category price effects with aggregate store data. *Management Science*, *52*(10), 1594–1609.
Thomassen, Ø., Smith, H., Seiler, S., & Schiraldi, P. (2017). Multi-category competition and market power: a model of supermarket pricing. *American Economic Review*, *107*(8), 2308–2351.
van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.
Wales, T. J., & Woodland, A. D. (1983). Estimation of consumer demand systems with binding non-negativity constraints. *Journal of Econometrics*, *21*(3), 263–285.