Shrinkage Priors for High-Dimensional Demand Estimation

Adam N. Smith and Jim E. Griffin[*]

May 17, 2020

Preliminary and Incomplete

Abstract

Estimating demand for wide assortments of differentiated goods requires the specification of a demand system that is sufficiently flexible. However, flexible models contain many parameters and will require regularization in high dimensions. For example, log-linear models suffer from a curse of dimensionality as the number of price elasticity parameters grows quadratically in the number of goods. In this paper, we study the specification of Bayesian shrinkage priors for price elasticity parameters within a log-linear demand system. Traditional regularized estimators assume fixed shrinkage points set to zero which can be at odds with many economic properties of cross-price effects. We propose a hierarchical extension of the class of global-local priors 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 cross-price elasticities and demand predictions can be improved by imposing shrinkage to higher-level group effects rather than zero.

Keywords: global-local priors, non-sparse shrinkage, horseshoe, hierarchical prior, seemingly unrelated regression, price elasticities

[*] Smith: UCL School of Management, University College London, Email: a.smith@ucl.ac.uk; Griffin: Department of Statistical Science, University College London, Email: j.griffin@ucl.ac.uk. We would like to 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.

1 Introduction

Measuring price and promotional effects from consumer transaction data is a mainstay of economics and marketing research. However, doing so in the presence of many goods presents significant challenges. Flexible regression-based demand models suffer from a curse of dimensionality as the number of parameters to be estimated grows quadratically with the number of goods. Logit-based demand models attempt to solve this problem by projecting demand onto the space of product characteristics, but flexibility is limited by functional form assumptions on utility and parametric assumptions on unobservables. Random coefficient logit models, which enrich logit substitution patterns by allowing for heterogeneity in tastes, face scalability constraints because the likelihood requires integrating over the space of unobservable product-specific characteristics. Moreover, the assumption of discrete choice is a defining characteristic of this class of models and complicates the analysis of complements and the demand for variety. Thus, regression-based systems remain attractive in large-scale settings because of their flexible functional form and simple likelihood.

Within the class of regression-based demand systems, there are differing views on how to incorporate cross-price effects. One approach is to exclude cross-price effects altogether and only regress the demand for each good on its own price (e.g., Bajari et al., 2014; DellaVigna and Gentzkow, 2019). While this simplifies the model structure, there is a concern of omitted variable bias if the prices of related goods have appreciable effects on the demand of the focal good. A second approach is to include "relevant" cross effects, where the set of relevant goods is usually defined by some a priori product categorization (e.g., Semenova et al., 2017; Hitsch et al., 2019; Strulov-Shlain, 2019). The key assumption here is that the product categorization accurately reflects the boundaries of consumer substitution patterns, which may not always hold empirically (Smith et al., 2019). A third approach is to include a high-dimensional set of cross effects and apply dimension-reduction techniques. However, there is a question of whether "off-the-shelf" regularized estimators, which assume sparsity in the underlying elasticity vector, are appropriate for demand estimation.

Economic theory offers at least three reasons why shrinking price effects to zero may not be optimal. The first comes from the Slutsky equation which decomposes a total price effect into a substitution effect and an income/budgetary effect. Even if the set of substitution effects is sparse, the price effects measurable from a Marshallian demand system can be dense in the presence of nonzero budget effects. The second reason comes from the property of Cournot aggregation (or "adding-up"), which imposes a restriction on the share-weighted average of own and cross elasticities. One implication is that demand must become more elastic as the set of available substitutes increases. Therefore, if cross-effects are arbitrarily shrunk towards zero, then the magnitude of the own effects must fall, holding all else constant, which would in turn lead to overestimated (i.e., too inelastic) own price effects. A third reason comes from economic separability restrictions, which are consistent with lower-dimensional representations of price elasticity matrices but do not require these matrices to be sparse (Goldman and Uzawa, 1964).

The goal of this paper is to construct shrinkage priors for demand elasticity parameters that are possibly dense (rather than sparse), but are assumed to conform to a lower-dimensional structure. To allow for structured shrinkage, we propose a hierarchical extension of the class of global-local priors (Polson and Scott, 2010) where the direction and rate of shrinkage are governed by a product classification tree. Typical scanner panel data sets (such as those provided by IRI and Nielsen) come with product hierarchies in which goods are first partitioned into broad categories (e.g., salty snacks) and then subsequently classified into subcategories (e.g., chips and cookies), brands, etc. We explicitly use this hierarchical grouping structure to direct shrinkage so that the price effect between two goods will depend on the price effect between their respective subcategories, which in turns depends on the price effect between their respective categories, and so on. The benefit is that the elasticity-level shrinkage points are not arbitrarily fixed at zero, but are instead learned from the data and shaped by prior knowledge of the boundaries in substitution patterns.

We apply our hierarchical priors to price elasticity parameters of a log-linear demand system with many goods. Log-linear models remain widely used in elasticity-based pricing applications (e.g., Montgomery, 1997; Hitsch et al., 2019; DellaVigna and Gentzkow, 2019), but can be challenging to estimate in high dimensional ($p > n$) settings. In addition to requiring regularization, there are significant costs associated with computing posterior moments in each iteration of a Gibbs sampler. To overcome this challenge, we implement a version of the fast sampler developed by Bhattacharya et al. (2016) for high-dimensional regression models with normal scale-mixture priors. We then use both simulated and actual retail scanner data to highlight the value of non-sparse shrinkage. In our empirical application, we estimate price elasticities for 245 products across 10 product categories and show potential improvements in demand predictions and estimated elasticities from imposing non-sparse hierarchical shrinkage.

Our work relates to two broad streams: on the application side, it fits into the work on large-scale demand estimation; on the methodological side, it contributes to the work on Bayesian regularization for high-dimensional linear models. The demand estimation literature has, over the course of its long history, proposed a variety of dimension reduction techniques to tackle the problem of estimating demand models with "many" prices. For models defined in the product space, a common approach is to impose microeconomic restrictions such as separability (e.g., Barten, 1964; Byron, 1970; Pudney, 1981) which induce low-dimensional representations of the matrix of cross-price effects. However, it is widely known that separability restrictions are too strong in many contexts and can lead to unrealistic estimated substitution patterns (Blackorby et al., 1978; Deaton and Muellbauer, 1980). The empirical shortcomings of many theory-based restrictions can in part be explained by misspecified shrinkage points and overly dogmatic, fixed rates of shrinkage. Montgomery and Rossi (1999) and Fessler and Kasy (2019) demonstrate benefits of constructing shrinkage estimators that are "centered" around microeconomic restrictions but can adapt based on signals in the data. Although the mean structure of our hierarchical prior is not derived from microeconomic theory, the idea of group-level, hierarchical shrinkage is similar in spirit to the dimension reduction offered by utility trees (Strotz, 1957), separability (Gorman, 1959; Goldman and Uzawa, 1964), and multi-stage budgeting (Gorman, 1971; Deaton and Muellbauer, 1980; Hausman et al., 1994; Hausman, 1996).

Discrete choice models solve the problem of many prices by projecting demand onto the space of product characteristics. However, there are still concerns of scalability when the set of choices or product characteristics is large, and a recent literature has emerged to address these issues. Gillen et al. (2014) and Bajari et al. (2014) apply lasso-type estimators to reduce the dimension of large product characteristics vectors entering consumer utility. Chiong and Shum (2018) use random projections to first reduce the dimension of the high-dimensional choice data and then estimate a discrete choice model on the "projected-down" data. Amano et al. (2019) leverage data on consumer search to infer consideration sets, which dramatically reduces the dimension of the choice set for each individual. Scaling up random utility models has many advantages given their microfoundations (Train, 2003), especially for counterfactual welfare analysis. However, the assumption of discrete choice, the specification of observed and unobserved product characteristics, and the parametric assumptions on idiosyncratic errors can raise practical concerns when the goal is to flexibly estimate price effects for large assortments.

Methodologically, our work relates to the literature on Bayesian shrinkage priors and high-dimensional regression models. In particular, we propose a hierarchical extension to the class of global-local priors (Polson and Scott, 2010). Global-local priors are normal scale-mixture priors in which the variance of each regression coefficient is expressed as the product of a local variance unique to each parameter a global variance common across parameters. This structure is attractive because, given appropriate choices of mixing densities, it can mimic the selection behavior of "two-group" point mass mixture priors (e.g., Mitchell and Beauchamp, 1988; George and McCulloch, 1993) while admitting much simpler posterior sampling strategies (Polson and Scott, 2012). The class of global-local priors is very broad and nests many well-known shrinkage estimators such as the ridge (Hoerl and Kennard, 1970) and Bayesian lasso (Park and Casella, 2008; Hans, 2009). Detailed reviews of this literature and the many types of global-local priors that now exist can be found in Polson and Scott (2010) and Bhadra et al. (2019).

Our contribution is to extend the global-local framework to allow for non-sparse, hierarchical shrinkage. The literature on global-local priors, and Bayesian regularization more generally, has 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 prior of Griffin and Brown (2017) developed to control shrinkage in models with interaction terms. Their prior controls 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. Our prior, which has a similar structure, controls not only the *rate* of shrinkage (through prior variances) but also the *direction* of shrinkage (through prior means).

The remainder of this paper is organized as follows. Section 2 presents the general log-linear demand framework. Section 3 reviews existing approaches to regularization. Section 4 outlines the development of hierarchical global-local priors and posterior computation is discussed in Section 5. Results from simulation experiments are provided in Section 6. Section 7 presents results of an empirical application to store-level scanner data. Section 8 concludes.

2 Demand Specification

We specify a log-linear demand system in which the log of sales of product $i = 1, \ldots, p$ at time period $t = 1, \ldots, n$ are regressed on its own log price, the log prices of other products, and a set of product-specific controls, which can include product intercepts, seasonal trends, and promotional activity.

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

Let $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{pt})$ denote the $p \times 1$ vector of error terms at time t where $\varepsilon_t \sim \mathrm{N}(0, \Sigma)$. Any contemporaneous correlation between goods can be captured through Σ. The log-linear specification remains popular in practice because the price coefficients represent own and cross-price elasticities, which are often the focal objects of interest in the analysis of market structure and pricing/promotion schedules.

The p demand equations are tied together through the correlation structure in Σ, giving rise to a multivariate demand system. Although the vector of prices is the same for each good, the presence of product-specific control variables leads to a seemingly unrelated regression (SUR) specification.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} X & 0 & \cdots & 0 \\ 0 & X & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_p \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} \tag{2}$$

Here y_i is the $n \times 1$ vector of log sales for product i, X is the $n \times p$ matrix of log prices, β_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 ϕ_i. In vector form, we have

$$y = \mathrm{X}\beta + \mathrm{C}\phi + \varepsilon, \ \varepsilon \sim \mathrm{N}(0, \Sigma \otimes I_n) \tag{3}$$

where $\mathrm{X} = \mathrm{diag}(X, X, \ldots, X)$ and $\mathrm{C} = \mathrm{diag}(C_1, C_2, \ldots, C_p)$.

The demand system described above has long been used for the purposes of estimating price elasticities from store-level transaction data (e.g., Blattberg and George, 1991; Montgomery, 1997). However, in most applications the number of products is small relative to the number of observations. High-dimensional settings with $p > n$ create challenges to inference, as many common

SUR estimators such as the original feasible GLS estimator of Zellner (1962) perform poorly. An additional challenge for SUR models is the rank deficiency of the sample covariance matrix when $p > n$. In the following section, we briefly review existing approaches for regularization of regression coefficients. For covariance regularization in sparse SUR models, see the proposed methods of Bhadra and Mallick (2013), Tan et al. (2018), or Li et al. (2019a).

3 Existing Approaches to Regularization

The classical regularization approach to estimating a sparse coefficient vector β is based on the solution to the following optimization problem.

$$\min_{\beta} \sum_{t} (y_t^* - \mathrm{x}_t'\beta)^2 + \sum_{j} \mathrm{pen}_{\lambda}(\beta_j) \tag{4}$$

Here t indexes the observation space and j indexes the parameter space. We denote the response as y^* to allow for other covariates to be included in the model without being subject to regularization. For example, $y^* = y - \mathrm{C}\phi$ for the SUR model in (3). The objective function in (4) has two components: the first represents a measure of statistical fit while the second penalizes nonzero elements of β.

Many widely-used regularized estimators correspond to solutions of (4) with a particular choice of a penalty function. For example, ridge regression (Hoerl and Kennard, 1970) is defined by an ℓ_2 penalty $\mathrm{pen}_{\lambda}(\beta_j) = \lambda\beta_j^2$, the lasso (Tibshirani, 1996) is defined by an ℓ_1 penalty $\mathrm{pen}_{\lambda}(\beta_j) = \lambda|\beta_j|$, and the elastic net (Zou and Hastie, 2005) uses a convex combination $\mathrm{pen}_{\lambda}(\beta_j) = \lambda_1\beta_j^2 + \lambda_2|\beta_j|$. With penalties that exhibit a spike at zero, such as the ℓ_1, the estimator will produce sparse solutions in that elements of β will be exactly zero. This is in contrast to penalties that are smooth around zero, like the ℓ_2, which only induce shrinkage but not selection.

Regularization can also be cast in the Bayesian paradigm, which naturally admits shrinkage by way of the prior. If the measure of fit in (4) is taken to be the log likelihood and the penalty is the log prior, then the solution to the optimization problem above corresponds to the posterior mode. Thus, any regularized estimator corresponds to a Bayesian maximum a posteriori estimator with respect to *some* prior $p(\beta|\lambda)$. For example, the ridge estimator arises from a normal prior, the lasso estimator arises from a Laplace prior (Park and Casella, 2008; Hans, 2009), and the elastic net estimator arises from a mixture of normal and Laplace priors (Li and Lin, 2010).

Recently, much attention in the literature on Bayesian regularization has been given to the class of global-local scale mixture priors:

$$\begin{aligned} \beta_j|\lambda_j^2, \tau^2 &\sim \mathrm{N}(0, \lambda_j^2\tau^2) \\ \lambda_j &\sim p(\lambda_j) \end{aligned} \tag{5}$$

where λ_j^2 represents the local variance and τ^2 represents the global variance. This class of priors is very broad and nests a variety of shrinkage priors. For example, the normal prior of ridge regression arises with $\lambda_j^2 = 1$ and the Laplace prior of the lasso arises with an exponential mixing density. One benefit of the global-local framework is that it allows for comparison among priors based on the choice of mixing density. For example, Polson and Scott (2010) show that when $p(\lambda_j)$ has exponential (or lighter) tails, as in the case of the Bayesian lasso, then β_j will be subject to shrinkage by a non-vanishing amount. Ideally, however, the amount of shrinkage should vanish for large observations (or observations that sufficiently deviate from the prior mean) so that the true signal is preserved. Priors exhibiting this property are said to be *tail-robust*.

There is now a large literature on the development of tail-robust shrinkage priors. Among the most notable is the horseshoe prior (Carvalho et al., 2010), which features a half-Cauchy mixing density with polynomial tails.

$$\begin{aligned} \beta_j|\lambda_j^2, \tau^2 &\sim \mathrm{N}(0, \lambda_j^2\tau^2) \\ \lambda_j &\sim \mathrm{C}^+(0, 1) \end{aligned} \tag{6}$$

Here $\mathrm{C}^+(0, 1)$ denotes the standard half-Cauchy distribution with density $p(x) \propto (1 + x^2)^{-1}$. The parameterization of the half-Cauchy also precludes additional hyperparameter tuning. The superior performance of the horsehshoe prior has been documented by many (e.g., Polson and Scott, 2010; Datta and Ghosh, 2013; Bhadra et al., 2016) and has led to its widespread use in a variety of models and problem areas, including logistic regression (Piironen and Vehtari, 2017), dynamic linear models (Kowal et al., 2019), linear models with many instrumental variables (Hahn et al., 2018b), treatment effect estimation (Hahn et al., 2018a), and high-dimensional SUR models (Li et al., 2019b). A more comprehensive review of the horseshoe and its many variants can be found in Bhadra et al. (2019).

4 Hierarchical Global-Local Shrinkage Priors

In this section, we develop hierarchical global-local priors that extend previous work in two ways. First, the prior means are parameterized according to a product classification tree. Cross effects at each level of the tree depend on the cross effects at the previous level, so that shrinkage is governed by group-level effects and is not fixed to zero. Second, the prior variances at each level also depend on the variances at higher levels which allows the rate of shrinkage to propagate through the tree. To start, we remain agnostic towards the choice of mixing density $p(\lambda_j)$ to highlight the generality of the hierarchical structure. At the end of this section, we provide results on the shrinkage properties of the hierarchical prior for different choices of $p(\lambda_j)$.

4.1 Notation

In order to describe our hierarchical prior with an arbitrary number of levels, we first require a minimum of notation. Let $\mathcal{T}$ denote a product classification tree with levels $\ell = 0, 1, \ldots, L$ where level 0 is the root node and level L corresponds to the most granular definition of products chosen by the researcher (e.g., brands, UPCs). Given node i on level ℓ, let $\mathrm{p}_\ell^s(i)$ denote the level-s ancestor of i and let $\mathrm{c}_\ell^s(i)$ denote the set of level-s descendants of i. Traditionally, these two functions are sufficient for describing the parameters of a graphical model. However, because each node maps to a single product, the parameters in our model are defined on pairs of nodes (i, j) rather than individual nodes (e.g., the effect of a price change of j on the demand for i). This is illustrated in Figure 1, which shows an example of a three-level tree arranged around the grid of product-level price elasticities.

Our hierarchical prior will require us to refer to the pair of level-s ancestors of (i, j). Rather than referring to this pair as $(p_\ell^s(i), p_\ell^s(j))$, we define an additional function to help keep our notation concise. Let $\mathrm{p}_\ell(i, j)$ be the set of all pairs of ancestors corresponding to i and j on levels higher than ℓ.

$$\mathrm{p}_\ell(i, j) = \Big\{ \big(\mathrm{p}_\ell^s(i), \mathrm{p}_\ell^s(j)\big) \text{ for all } s = 1, \ldots, \ell - 1 \Big\} \tag{7}$$

For example, in Figure 1 we have $\mathrm{p}_3(1, 12) = \{(1, 3), (1, 6)\}$ and $\mathrm{p}_2(1, 6) = \{(1, 3)\}$. Lastly, let

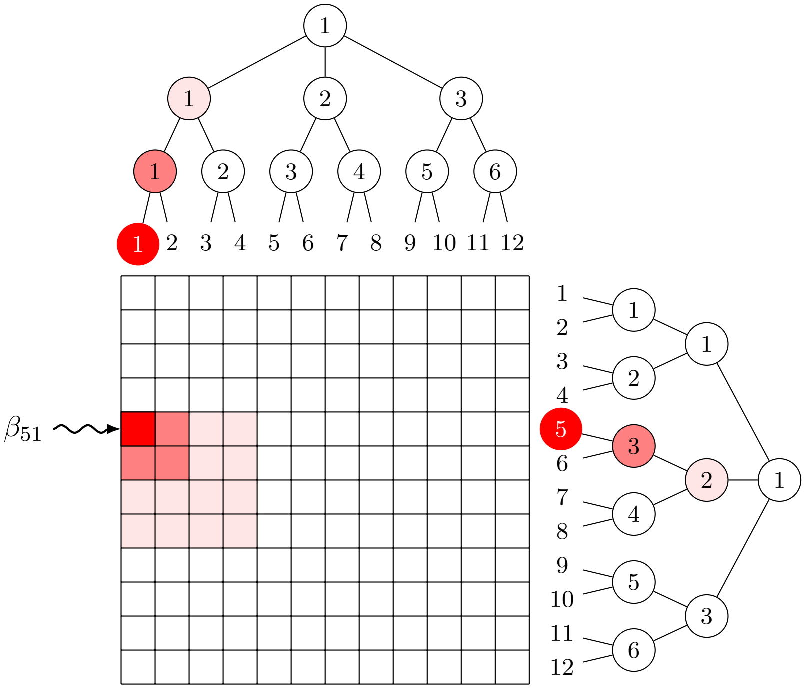


Figure 1: Visualization of hierarchical shrinkage using a three-level product classification tree.

$m_\ell(i, j)$ denote the highest level in the tree such that i and j from level ℓ share the same ancestor.

$$m_\ell(i, j) = \max \left\{ s \in \{1, \ldots, \ell - 1\} \text{ s.t. } p_\ell^{s-1}(i) = p_\ell^{s-1}(j) \right\} \tag{8}$$

For example, $m_3(1, 12) = 1$, $m_3(1, 4) = 2$, and $m_3(1, 2) = 2$.

Using this notation, we can now introduce parameters representing the relationship between nodes. For all interior nodes, we use θ's to represent group-level effects. Specifically, let $\theta_{\mathrm{p}_{ij}(\ell)}$ denote the vector of parameters assigned to each element of the set $\mathrm{p}_\ell(i, j)$ and let $\theta^{(m)}_{\mathrm{p}_\ell(i,j)}$ denote the mth element of $\theta_{\mathrm{p}_\ell(i,j)}$. Referring back to Figure 1 and the example above, we have $\theta_{\mathrm{p}_3(1,12)} = \{\theta^{(1)}_{13}, \theta^{(2)}_{16}\}$ and $\theta^{(1)}_{\mathrm{p}_3(1,12)} = \theta^{(1)}_{13}$.

4.2 Prior Construction

We specify the hierarchical prior starting from the top level down. Let $\theta^{(1)}_{kl}$ denote the relationship between groups k and l on the first level of the tree, which is assumed to have a normal prior without any local-shrinkage components.

$$\theta^{(1)}_{kl} \sim \mathrm{N}\left(\bar{\theta}, \bar{\tau}^2\right) \tag{9}$$

Then for any two groups on level $\ell \in \{2, \ldots, L-1\}$, define the prior on the group-level effects as

$$\theta_{kl}^{(\ell)} \sim \mathrm{N}\left(\theta_{\mathrm{p}_\ell(k,l)}^{(\ell-1)}, \lambda_{kl}^{2(\ell)} \Psi_{kl}^{(\ell-1)} \tau^{2(\ell)}\right) \tag{10}$$

where $\Psi_{kl}^{(1)} = 1$ and

$$\Psi_{kl}^{(\ell-1)} = \prod_{s=m_\ell(k,l)}^{\ell-1} \lambda_{\mathrm{p}_\ell(k,l)}^{2(s)}. \tag{11}$$

There are two key features of this specification. The first is that the prior mean is not fixed at zero but is instead equal to the relationship between the two parent groups of k and l. The second is that the prior variance equals the usual global-local structure – i.e., a local variance $\lambda_{kl}^{2(\ell)}$ multiplied by a level-specific global variance $\tau^{2(\ell)}$ – multiplied by the product of local variances from higher levels $\Psi_{kl}^{(\ell-1)}$. This causes the variance of $\theta_{kl}^{(\ell)}$ to be small (inducing more shrinkage towards the mean) whenever the variance of any higher-level effects are small, allowing rates of shrinkage to propagate through the tree.

At the last level of the tree we have product elasticities which follow the same structure as the group-level effects described above. Because these parameters enter the demand equation in (1), we let β_{ij} denote price elasticities.

$$\beta_{ij} \sim \mathrm{N}\left(\theta_{\mathrm{p}_L(i,j)}^{(L-1)}, \lambda_{ij}^{2(L)} \Psi_{ij}^{(L-1)} \tau^{2(L)}\right) \tag{12}$$

The hierarchical structure outlined above applies to all cross-price elasticity parameters (i.e., β_{ij} for $i \neq j$). Our prior specification is therefore not complete until we also define priors for the own-price elasticities, β_{ii}. In order to account for differences in the expected sign of own and cross elasticities, we let β_{ii} have a normal prior with mean and variance that is independent of the hierarchical shrinkage structure above: $\beta_{ii} \sim \mathrm{N}(\bar{\beta}_{ii}, \bar{\tau}_{ii}^2)$.

We bring together the priors on the own and cross-price elasticities and define $\bar{B}(\theta)$ and $\Lambda(\tau)$ to be $p \times p$ matrices of all prior means and variances, respectively.

$$\bar{B}(\theta)_{ij} = \begin{cases} \theta_{\mathrm{p}_L(i,j)}^{(L-1)} & \text{if } i \neq j \\ \bar{\beta}_{ii} & \text{if } i = j \end{cases} \tag{13}$$

$$\Lambda(\tau)_{ij} = \begin{cases} \lambda_{ij}^{2(L)} \Psi_{ij}^{(L-1)} \tau^{2(L)} & \text{if } i \neq j \\ \bar{\tau}_{ii}^2 & \text{if } i = j \end{cases} \tag{14}$$

Because these priors are defined for parameters of a SUR model, it will also be useful to define their vectorized counterparts. Let $\bar{\beta}(\theta) = \text{vec}(\bar{B}(\theta))$ be the p^2-dimensional vector of means and $\Lambda_* = \text{diag}(\text{vec}(\Lambda(\tau)))$ be the $p^2 \times p^2$ diagonal matrix of variances. The complete set of model parameters is summarized in Table 1.

Table 1: Summary of Model Parameters

		Parameters	
Tree Level	Effects	Means	Variances
Top ($\ell = 1$)	$\theta_{kl}^{(1)}$	$\bar{\theta}$	$\bar{\tau}^2$
Interior ($1 < \ell < L$)	$\theta_{kl}^{(\ell)}$	-	$\lambda_{kl}^{2(\ell)}$, $\tau^{2(\ell)}$
Bottom ($\ell = L$)	β_{ij}	-	$\lambda_{ij}^{2(L)}$, $\tau^{2(L)}$
	β_{ii}	$\bar{\beta}_{ii}$	$\bar{\tau}_{ii}^2$

4.3 Shrinkage Properties

In the global-local framework, the tails of the mixing density $p(\lambda_j)$ fully determine the shrinkage behavior of the induced prior. For example, the heavy tails of the half-Cauchy distribution are central to the success of the horseshoe approach and several papers have considered their effect on consistent estimation (van der Pas et al., 2014, 2016, 2017; Ghosh and Chakrabarti, 2017) and Bayesian risk properties (Datta and Ghosh, 2013). These results can also be directly linked to the property of "regular variation" of a distribution, which requires the tails to behave like a power law function (see Bhadra et al., 2016, for further details). In this section, we explore the tail behavior of hierarchical global-local priors and show that their heaviness is still determined by the tails of the chosen mixing distribution.

Focal Parameters

We first write the elasticities and higher-level mean parameters as a function of their respective means and idiosyncratic errors.

$$\beta_{ij} = \theta_{p_L(i,j)}^{(L-1)} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim \text{N}\left(0, \lambda_{ij}^{2(L)} \Psi_{ij}^{(L-1)} \tau^{2(L)}\right) \tag{15}$$

$$\theta_{kl}^{(\ell)} = \theta_{p_\ell(k,l)}^{(\ell-1)} + \epsilon_{kl}^{(\ell)}, \qquad \epsilon_{kl}^{(\ell)} \sim \text{N}\left(0, \lambda_{kl}^{2(\ell)} \Psi_{kl}^{(\ell-1)} \tau^{2(\ell)}\right) \tag{16}$$

Note that, for any $m > m_L(i,j)$, we can write β_{ij} as a function of its level-m parent mean and variance.

$$\beta_{ij} = \theta^{(m)}_{p_L(i,j)} + \epsilon_{ij} + \sum_{k=m+1}^{L-1} \epsilon^{(k)}_{p_L(i,j)}. \tag{17}$$

There are two forms of shrinkage that we are interested in.

1. The shrinkage of β_{ij} to β_{pq}, which is the shrinkage between two product-level elasticities. There are two cases to consider. The first is the case that the any of the four products (i,j,p,q) belong to the same group at any level of the tree. Formally, this can be stated as if there is an m for which $p^{(m)}_L(i,j) = p^{(m)}_L(k,l)$. In this case, then we can examine the shrinkage of the difference:

   $$\delta = \beta_{ij} - \beta_{pq} = \epsilon_{ij} - \epsilon_{pq} + \sum_{k=m^\star}^{L-1} \left(\epsilon^{(k)}_{p_L(i,j)} - \epsilon^{(k)}_{p_L(p,q)}\right) \tag{18}$$

   where $m^\star = \max\{m : p^{(m)}_L(i,j) = p^{(m)}_L(k,l)\}$. Clearly,

   $$\epsilon_{ij} - \epsilon_{pq} + \sum_{k=m^\star}^{L-1} \left(\epsilon^{(k)}_{p_L(i,j)} - \epsilon^{(k)}_{p_L(p,q)}\right) \sim \mathrm{N}(0,\eta) \tag{19}$$

   where

   $$\eta = \tau^{(L)}\left(\lambda^{2(L)}_{ij}\Psi^{(L-1)}_{ij} + \lambda^{2(L)}_{pq}\Psi^{(L-1)}_{pq}\right) + \sum_{k=m}^{L-1} \tau^{(k)}\left(\lambda^{2(k)}_{p_L(p,q)}\Psi^{(k)}_{p_L(p,q)} + \lambda^{2(k)}_{p_L(i,j)}\Psi^{(k)}_{p_L(i,j)}\right). \tag{20}$$

   The second case is if there is not an m for which $p^{(m)}_L(i,j) = p^{(m)}_L(k,l)$. Then we can write the difference as:

   $$\begin{aligned}\delta &= \beta_{ij} - \beta_{pq} \qquad (21)\\ &= \theta^{(m_L(i,j)+1)}_{p_L(i,j)} - \theta^{(m_L(p,q)+1)}_{p_L(p,q)} + \epsilon_{ij} - \epsilon_{pq} + \sum_{k=m_L(i,j)+2}^{L-1} \epsilon^{(k)}_{p_L(i,j)} - \sum_{k=m_L(p,q)+2}^{L-1} \epsilon^{(k)}_{p_L(p,q)}.\end{aligned}$$

2. The shrinkage of β_{ij} to $\theta^{(m)}_{p_L(i,j)}$, which is the shrinkage of a product-level elasticity to the cross elasticity of its parents at level m. This can be written as:

   $$\delta = \beta_{ij} - \theta^{(m-1)}_{p_L(i,j)} = \epsilon_{ij} + \sum_{k=m}^{L-1} \epsilon^{(k)}_{p_L(i,j)} \tag{22}$$

where

$$\epsilon_{ij} + \sum_{k=m}^{L-1} \epsilon^{(k)}_{p_L(i,j)} \sim \mathrm{N}\left(0, \lambda^{2(L)}_{ij}\Psi^{(L-1)}_{ij}\tau^{2(L)} + \sum_{k=m}^{L-1} \lambda^{2(k)}_{p_L(i,j)}\Psi^{(k)}_{p_L(i,j)}\tau^{2(k)}\right). \tag{23}$$

In both cases, the differences δ have priors which are scale mixtures of normals, where the scales are sums of products of the local variances.

Mixing Densities

Because the priors on δ can be expressed as normal scale mixtures, the shape of the marginal prior will again be determined by the mixing density (Barndorff-Nielsen et al., 1982). For example, in the case of the a hierarchical ridge, $\lambda^{2(\ell)}_{kl} = 1$ for each level $\ell = 2, \ldots, L$ so the marginal priors on δ will have the tails of a normal distribution. Perhaps a more interesting question is whether the heaviness of non-degenerate mixing densities are preserved under the "scaled sum of products" transformations of local-variances seen in (18), (21), and (22). The following proposition formalizes this statement.

Proposition 1. *Let G be a regularly varying distribution with density $g(\cdot)$. Define $\zeta = \sum_{i=1}^{q} \tau_i^2 \prod_{j=1}^{i} \lambda_j^2$ with $\lambda_j^2 \sim G$ and fixed $\tau_1^1, \ldots, \tau_q^2$. Then the probability density function of ζ can be written as:*

$$p(\zeta) \sim K(q)\, g(\zeta) \quad \textit{as} \quad \zeta \to 0 \quad \textit{or} \quad \zeta \to \infty$$

where $K(q)$ is the normalizing constant.

Proof. If G is regularly varying and $\lambda_1^2, \ldots, \lambda_i^2 \overset{i.i.d.}{\sim} G$, then $\Psi_i = \prod_{j=1}^{i} \lambda_j^2$ is also regularly varying (Cline, 1987). This implies that its density can be written as $p(\Psi_i) \sim \Psi_i^{\alpha} L(\Psi_i)$ where $\alpha \in \mathbb{R}$ and $L(\cdot)$ is slowly-varying (i.e., regularly varying with an index of 0). The closure property of regularly varying functions then guarantees that the scaled sum $\zeta = \tau_i^2 \sum_{i=1}^{q} \Psi_i$ is also slowly varying. □

This result shows that the heaviness of the chosen mixing density does indeed carry through. For example, consider the horseshoe prior with its regularly varying, half-Cauchy mixing density. If $\lambda \sim \mathrm{C}^+(0, 1)$, then λ^2 is an inverted-beta random variable with density $g(\lambda^2) \propto (\lambda^2)^{-1/2}(1 + \lambda^2)^{-1}$ (Polson and Scott, 2012), which is regularly varying with index -3/2 and so λ^2 is regularly varying with index 1/2 (Bingham et al., 1987). Then by Proposition 1, the sum of products of squared half-Cauchys would also have regularly varying tails, and the different forms of shrinkage in (18), (21), and (22) will all have tails of the same heaviness as a standard horseshoe prior. This is illustrated in Figure 2, which plots the marginal distribution of δ under a few different shrinkage priors: normal (ridge), Laplace (lasso), half-Cauchy (horseshoe), and sums of products of squared half-Cauchys (hierarchical horseshoe).

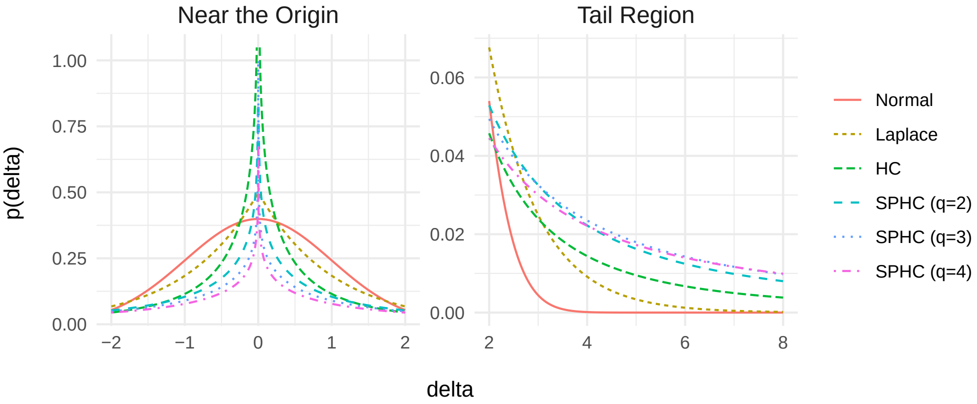


Figure 2: Marginal prior densities $p(\delta)$ are shown for different choices of shrinkage priors: Laplace, half-Cauchy (HC), and sums of products of half-Cauchys (SPHC) with $q \in \{2, 3, 4\}$.

5 Posterior Computation and Scalability

Sampling from the posterior of a SUR model first requires transforming the system in (3) into one with homogeneous unit errors. Let U denote the upper triangular Cholesky root of Σ 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 regression model.

$$\tilde{y} = \tilde{\mathrm{X}}\beta + \tilde{\mathrm{C}}\phi + \tilde{\varepsilon}, \qquad \tilde{\varepsilon} \sim \mathrm{N}(0, I_{np}) \tag{24}$$

The full set of model parameters includes the elasticities β, the coefficients on controls ϕ, the error variance Σ, and the set of hierarchical parameters $\Omega = \left(\{\theta_{kl}^{(\ell)}\}, \{\lambda_{kl}^{2(\ell)}\}, \{\tau^{2(\ell)}\}\right)$.

The priors for β and all hierarchical parameters are given in Section 4.2. We assume normal priors for the coefficients on control variables $\phi \sim \mathrm{N}(\bar{\phi}, A_{\phi}^{-1})$, which are conditionally conjugate to the normal likelihood given Σ. Inverse Wishart priors are commonly used for covariance matrices in Bayesian SUR models. However, in the high-dimensional case of $p > n$ then Σ will be rank deficient. One approach would be to also regularize Σ (see, e.g., Li et al., 2019a,b). However, we consider the simpler approach of imposing a diagonal restriction: $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$. In this case, independent inverse gamma priors can be used for each σ_j^2.

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

$$\Omega|\beta, \text{data} \tag{25}$$

$$\Sigma|\beta, \phi, \Omega, \text{data} \tag{26}$$

$$\beta, \phi|\Sigma, \beta, \Omega, \text{data} \tag{27}$$

The first full conditional represents the posterior of all global/local variances and higher-level means. Sampling from these distributions is computationally inexpensive. The elements of θ each have independent normal posteriors. The local and global variances can also be sampled independently, but their specific forms will depend on the choice of priors. For example, in the cause of the hierarchical horseshoe, the posteriors will consist of a normal likelihood from the mean parameters θ and a half-Cauchy prior on the variances. In this case, we take the approach of Makalic and Schmidt (2015) and represent the half-Cauchy priors as scale mixtures of inverse gammas, which are conjugate to the normal likelihood. Full details of the hierarchical horseshoe full conditional are given in Appendix A.

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

$$\sigma_j^2|\beta, \phi, \Omega, \text{data} \sim \text{IG}\Big(a + n/2, b + (y_j - X\beta_j - C_j\phi_j)'(y_j - X\beta_j - C_j\phi_j)/2\Big) \tag{28}$$

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 Σ is unrestricted, the typical conditionally conjugate inverse Wishart priors can be used.

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

$$\beta|\phi, \Sigma, \Omega, \text{data} \sim \text{N}\Big((\tilde{\text{X}}'\tilde{\text{X}} + \Lambda_*^{-1})^{-1}(\tilde{\text{X}}'\tilde{y}_\phi^* + \Lambda_*^{-1}\bar{\beta}(\theta)), (\tilde{\text{X}}'\tilde{\text{X}} + \Lambda_*^{-1})^{-1}\Big) \tag{29}$$

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

$$\phi|\beta, \Sigma, \Omega, \text{data} \sim \mathrm{N}\Big(\big(\tilde{\mathrm{C}}'\tilde{\mathrm{C}} + A_\phi\big)^{-1}\big(\tilde{\mathrm{C}}'\tilde{y}^*_\beta + A_\phi\bar{\phi}\big), \big(\tilde{\mathrm{C}}'\tilde{\mathrm{C}} + A_\phi\big)^{-1}\Big) \tag{30}$$

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

$$\phi|\Sigma, \Omega, \text{data} \sim \mathrm{N}\Big(\big(\tilde{\mathrm{C}}'\mathrm{P}\tilde{\mathrm{C}} + A_\phi\big)^{-1}\big(\tilde{\mathrm{C}}'\mathrm{P}\tilde{y} + A_\phi\bar{\phi}\big), \big(\tilde{\mathrm{C}}'\mathrm{P}\tilde{\mathrm{C}} + A_\phi\big)^{-1}\Big) \tag{31}$$

where $\mathrm{P} = I_{np} - \tilde{\mathrm{X}}(\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})^{-1}\tilde{\mathrm{X}}'$ is an orthogonal projection matrix. Marginalizing over β 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 (29). It should also be noted that the posterior precision matrix in (29) 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.

5.1 Diagonal restriction on Σ

As with all Bayesian regression models, a computational bottleneck arises in inverting the posterior precision matrix $(\tilde{\mathrm{X}}'\tilde{\mathrm{X}}+\Lambda_*^{-1})$. This especially true for Bayesian SUR models since the design matrix X contains stacked copies of the multivariate regression design matrix. If Σ is unrestricted, then $\tilde{\mathrm{X}}'\tilde{\mathrm{X}}$ 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 Σ is assumed to be diagonal then both $\tilde{\mathrm{X}}'\tilde{\mathrm{X}}$ and $(\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})$ will have block diagonal structures, with each of the p blocks containing an $p \times p$ matrix. Computing the inverse of $(\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})$ 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 $(\tilde{\mathrm{X}}'\tilde{\mathrm{X}} + \Lambda_*^{-1})$ directly, it can still be prohibitively expensive for large p.

5.2 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 $(\tilde{\mathrm{X}}'\tilde{\mathrm{X}}+\Lambda_*^{-1})$. 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 Σ is diagonal or $\mathcal{O}(n^2p^4)$ if Σ 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 in (12). A proof that β retains the correct posterior is given in Appendix B.

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 $\beta = u + \Lambda_*\tilde{X}'w$.

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.

To provide practical insights into the computational gains afforded by fast sampling algorithm above, we draw from the posterior of the elasticity vector β using data generated with $n = 50$ and $p \in \{50, 100, 250, 500, 1000\}$. 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 using Cholesky decompositions. In both cases we assume Σ is diagonal. The samplers are coded in Rcpp (Eddelbuettel and François, 2011) and run on a MacBook Pro laptop with a 2.3 GHz Dual-Core i5 processor. Figure 3 plots the computation time in log seconds against the number of products p. As expected from the complexity results above, the fast sampler's computational cost grows nonlinearly in p. For example, drawing β takes about 5 seconds per iteration with $p = 500$, but 35 seconds per iteration when $p = 1000$. However, the fast sampler provides drastic computational savings when p is large. In the case of $p = 1000$, the fast sampler is an order of magnitude faster than a standard sampler.

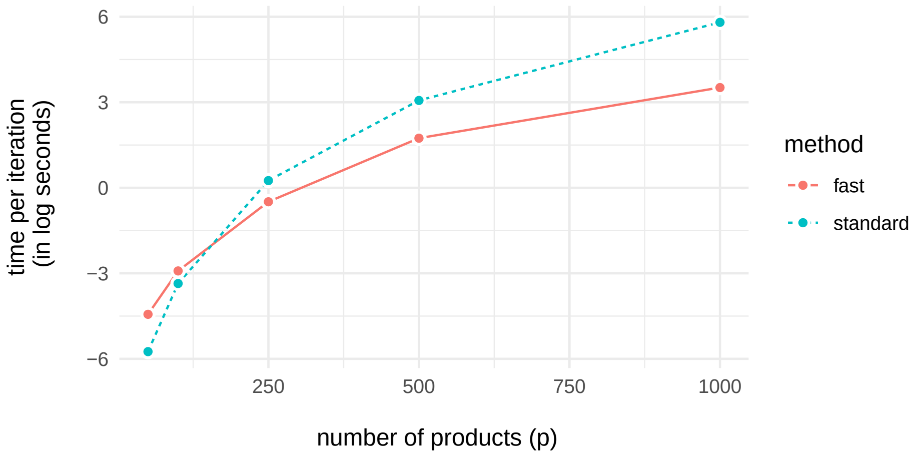


Figure 3: Computation time associated with one draw from the posterior of the product-level elasticities β across varying dimensions of the product space.

6 Simulation Results

We first explore the performance of hierarchical global-local priors using simulated data. We consider two data generating processes. The first is the hierarchical prior outlined in Section 4.2 that generates a dense elasticity vector β. For this model, we specify a three-level tree where the first level of the tree contains 5 groups and the second contains 10 groups. We draw $\tau^{(\ell)} \sim \text{IG}(2,2)$, $\lambda_{ij}^{(\ell)} \sim \text{IG}(2,2)$, and generate the group-level effects $\theta_{kl}^{(\ell)}$ from the hierarchy in (9) and (10). The second data generating process produces a sparse β vector, as is typically assumed in high-dimensional linear regression. Here we allow 5% of the elements in β to be non-zero and draw them from a $\text{N}(0,4)$ distribution. In both data generating processes, we draw $\sigma_j^2 \sim \text{Unif}(0,1)$ and set $\phi_j = 0$ for all $j = 1, \ldots, p$.

Hierarchical ridge and horseshoe priors are compared to normal (ridge), Laplace (lasso), and horseshoe priors. The priors associated with ridge and lasso estimators are expressed as global-local models with half-Cauchy priors on global components (τ). The priors on the local components λ_j give rise to different forms of shrinkage. Ridge regression corresponds to $\lambda_j^2 = 1$ and the lasso corresponds to $\lambda_j^2 \sim \text{Exp}(1/2)$ using the rate parameterization of the exponential distribution.

Model fit statistics are shown in Table 2. We report the root mean squared error (RMSE) associated with out-of-sample prediction values $\hat{y}$ and estimated parameter values $\hat{\beta}$. The RMSE statistics are averaged over 25 simulated data sets with $n = 50$ and $p \in \{50, 100\}$. We find that when the elasticities are dense and generated from a hierarchical structure, the hierarchical horseshoe and ridge perform best. When the elasticities are sparse, the standard horseshoe has the lowest RMSE. However, the fit of the hierarchical horseshoe is still competitive, providing superior fit relative to the ridge and lasso.

Table 2: Model Comparison with Simulated Data

	True Model: Dense			
	Prediction RMSE		β RMSE	
Shrinkage	$p = 50$	$p = 100$	$p = 50$	$p = 100$
Ridge	8.564	30.191	1.148	3.000
Lasso	8.063	28.841	1.093	2.858
Horseshoe	7.982	29.955	1.084	2.963
Hierarchical Ridge	6.766	**23.824**	0.939	**1.860**
Hierarchical Horseshoe	**6.699**	25.698	**0.868**	2.354

	True Model: Sparse			
	Prediction RMSE		β RMSE	
Shrinkage	$p = 50$	$p = 100$	$p = 50$	$p = 100$
Ridge	2.126	3.390	0.280	0.324
Lasso	1.205	1.929	0.134	0.176
Horseshoe	**0.863**	**1.000**	**0.063**	**0.067**
Hierarchical Ridge	2.129	3.398	0.280	0.325
Hierarchical Horseshoe	0.938	1.430	0.081	0.119

Notes: Model fit statistics are averaged over 25 simulated data sets with $n = 50$. Root mean squared errors (RMSE) associated with out-of-sample predictions $\hat{y}$ and estimated parameters $\hat{\beta}$ are reported. The best models are shown in bold.

Our demand model and hierarchical horseshoe prior allow us to learn not only about product-level elasticities β but also about the higher-level group effects $\theta_{kl}^{(\ell)}$. So in addition to examining predictive performance, we also explore how well we these higher-level effects are recovered using simulated data. Data are simulated from the hierarchical model as described above with $n = 50$ and $p = 100$. The three-level tree contains 25 parameters at the top level and 100 parameters in the middle level. In Figure 4, we plot the posterior means and 95% credible intervals for $\theta_{kl}^{(\ell)}$ (from the hierarchical horseshoe) against their true values. We find that the majority of credible intervals cover the true values. Exceptions can be attributed to sampling error or heavy shrinkage.

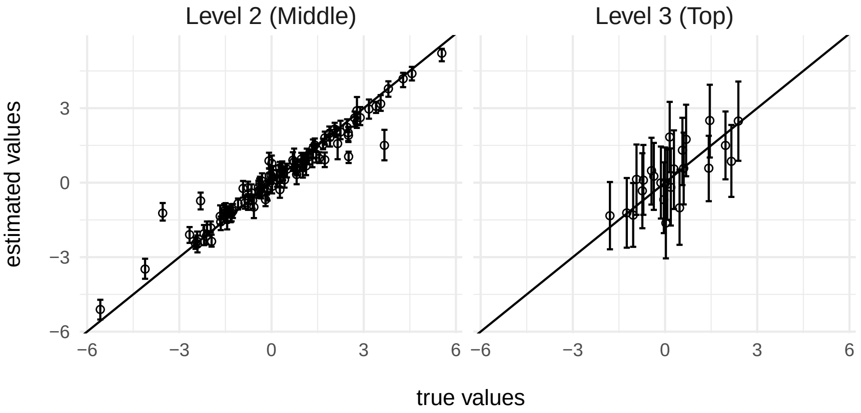


Figure 4: Posterior means and credible intervals for group-level effects $\theta_{kl}^{(\ell)}$ (from the hierarchical horseshoe prior) plotted against their true values.

7 Empirical Application

Our empirical analysis is based on store-level data from the IRI retail panel (Bronnenberg et al., 2008). We use data from one retail chain in Pittsfield, Massachusetts covering the 2011 calendar year. We use an approximate 75/25% split of the data for estimation and out-of-sample prediction, resulting in $n = 80$ weeks for the training data and 24 weeks for the test data. Although there is the potential to add data from other stores, chains, and markets, we take the perspective of the retailer who often want 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 – products only come from one category or subcategory, but are defined at a very granular level (e.g., UPC); (ii) wide and shallow – products span many categories, but are defined at a higher level of aggregation (e.g., brand). 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 and many perfectly collinear prices. Estimating demand for wider assortments also highlights the general flexibility of log-linear systems while calling into question many of the typical assumptions (e.g., discrete choice) underling logit-based models.

Our process for selecting products is as follows. We first choose 10 broad categories that cover a large portion of total expenditure. We then aggregate UPCs to the category-subcategory-brand level using total quantities and share-weighted prices normalized to the modal subcategory pack size. The number of products selected from each subcategory is proportional to its share of revenue, and the within subcategory selection of brands is also based on revenue shares. This selection process results in $p = 245$ products that are comprised of 2,437 unique UPCs and represent 82% of total revenue within the 10 chosen categories. A list of all categories and subcategories used in our analysis is provided in Table 3.

Table 3: Product Categories

Category	Subcategories	No. of Products
BEER	Domestic, Imported	31
CARBONATED BEVERAGES	Low Calorie Soft Drinks, Regular Soft Drinks, Seltzer / Tonic / Club Soda	50
COFFEE	Ground Coffee, Ground Decaffeinated Coffee, Instant Coffee, Instant Decaffeinated Coffee, Single Cup Coffee, Whole Coffee Beans	25
FROZEN DINNERS/ENTREES	Handheld Entrees, Multi-Serve Dinners, Single-Serve Dinners	23
FROZEN PIZZA	Pizza	7
MILK	Flavored Milk / Eggnog / Buttercream, Soy Milk, Skim / Lowfat Milk, Whole Milk	16
MUSTARD & KETCHUP	Ketchup, Mustard	8
SALTY SNACKS	Cheese Snacks, Corn Snacks, Other Salted Snacks, Potato Chips, Pretzels, Popcorn, Tortilla Chips	35
SOUP	Bouillon, Condensed Wet Soup, RTS Dry Soup, Ramen, Broth, RTS Wet Soup	27
YOGURT	Yogurt	23
Total Count = 10	36	245

We estimate log-linear SUR models of the form in (1) assuming a diagonal error variance matrix Σ. In addition to the high-dimensional vector of prices, we also include quarterly dummy variables and display and feature promotion incidence variables as product-specific controls. We place mildly informative inverse gamma priors on the error variances, diffuse normal priors on the control coefficients, and informative priors on the own-price elasticities that concentrate probability mass in the range (-5,-1). We consider five different shrinkage priors for the cross-price elasticities: ridge, lasso, horseshoe, hierarchical ridge, and hierarchical horseshoe. For each model, we use a half-Cauchy prior on the global variances which is widely accepted as an appropriate default choice (Polson and Scott, 2012). For the hierarchical models, we use a three-level tree. The top level corresponds to categories, the second level corresponds to subcategories, and the final level corresponds to products.

Table 4 reports the out-of-sample root mean squared error (RMSE) for all five models. We find that the hierarchical ridge model fits best, and provides a roughly 5 percentage point improvement relative to the ridge and lasso and a close to an 8.7 percentage point improvement relative to the standard horseshoe. The performance of the ridge in both standard and hierarchical cases provides evidence that the true underlying elasticity vectors are unlikely to be sparse.

Table 4: Model Fit Statistics

	Predictive RMSE	
Shrinkage Prior	Mean	SD
Ridge	0.820	(0.009)
Lasso	0.823	(0.009)
Horseshoe	0.851	(0.017)
Hierarchical Ridge	**0.764**	**(0.008)**
Hierarchical Horseshoe	0.901	(0.019)

Next, we compare the estimated own and cross-price elasticities generated from each model. Own elasticities are shown in the top panel of Figure 5 and cross elasticities are shown in the bottom panel. The estimated set of own elasticities are fairly similar across models, with average own elasticities of -1.70 (ridge), -1.72 (lasso), -1.79 (horseshoe), -1.64 (hierarchical ridge), and -1.98 (hierarchical horseshoe). The distribution of cross-price is markedly different across models. Both versions of the ridge yield estimates concentrated between -0.2 and 0.2, with the hierarchical version showing more of a mean shift away from zero. Estimates from the lasso exhibit a high concentration around zero with few larger estimates closer to -0.5 and 0.5. Finally, the distribution of estimates from both versions of the horseshoe shows a tall spike at zero with a few much larger values between -3 and 6. The distribution from the hierarchical horseshoe also has slightly broader shoulders around zero, suggesting that there is more appreciable mass placed on non-zero cross elasticities. In general, we find that the estimated distributions are consistent with the theoretical properties of each shrinkage prior discussed in Section 3.

The hierarchical versions of the ridge and horseshoe allow us to further examine higher-level group effects. For the sake of brevity, we only report estimates from the hierarchical ridge since it

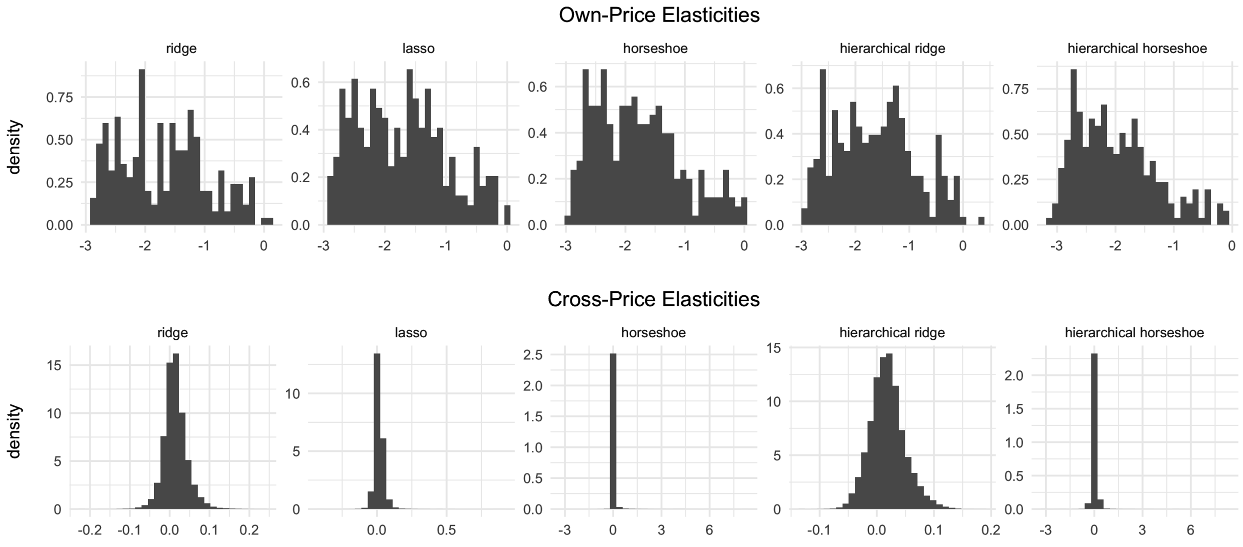
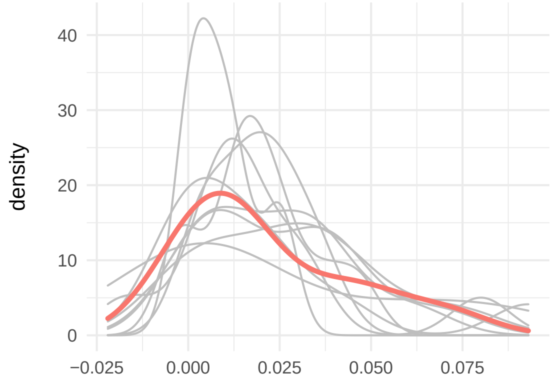


Figure 5: Distribution of own and cross-price elasticities.

had the best predictive performance. Figure 6 plots the distribution of category and subcategory-level effects. The top panel shows the category-level effects, and the left-hand figure plots the distribution of $\theta_{kl}^{(1)}$ for each of the 10 categories. On the right-hand side, we provide an example of the estimates for the BEER category. The estimates appear reasonable, with the largest effect being the "own" category effect which has a value of 0.07. Note that this is not indicative of an own elasticity, but instead represents the shrinkage point for all cross elasticities within BEER. The relatively large positive value suggests that competitive effects are greater within the category than across categories. However, many of the cross-category effects are not zero. The second and third-largest effects come from FROZEN PIZZA and CARBONATED BEVERAGE. Interestingly, the smallest effect (-0.003) is associated with SALTY SNACKS, which may be explained by the complementarity between beer and salty snacks.

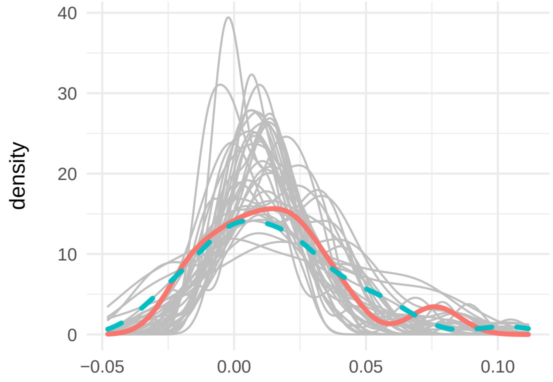
The bottom panel of Figure 6 shows the distribution of $\theta_{kl}^{(2)}$ for each of the 36 subcategories. The two subcategories within BEER are shown by the solid red line (Imported Beer) and dashed green line (Domestic Beer). The top three largest and smallest effects are reported for Imported Beer in the table on the right-hand side. Estimates again appear reasonable, with Imported and Domestic Beer categories appearing at the top. In addition, the smallest (negative) effects are exhibited by Potato Chips, Cheese Snacks, and Other Salted Snacks – all from the SALTY SNACKS category which is likely to have a complementarity relationship with BEER.

(a) Category-Level Effects $\theta_{kl}^{(1)}$



Category k	Category l	Mean	SD
BEER	BEER	0.070	(0.021)
	FROZEN PIZZA	0.048	(0.029)
	CARBONATED BEVERAGES	0.041	(0.016)
	MUSTARD & KETCHUP	0.020	(0.020)
	COFFEE	0.016	(0.012)
	FROZEN DINNERS/ENTREES	0.011	(0.016)
	SOUP	0.009	(0.013)
	MILK	0.005	(0.016)
	YOGURT	0.001	(0.024)
	SALTY SNACKS	-0.003	(0.011)

(b) Subcategory-Level Effects $\theta_{kl}^{(2)}$



Subcategory k	Subcategory l	Mean	SD
Imported Beer	Domestic Beer	0.079	(0.019)
	Imported Beer	0.077	(0.041)
	Seltzer / Tonic / Club Soda	0.072	(0.037)
	⋮	⋮	⋮
	Potato Chips	-0.019	(0.021)
	Cheese Snacks	-0.019	(0.028)
	Other Salted Snacks	-0.022	(0.025)

Figure 6: Distributions of category (top panel) and subcategory (bottom panel) level effects $\theta_{kl}^{(\ell)}$. In the top panel, the solid red line shows the distribution of $\theta_{kl}^{(1)}$ for the BEER category. In the bottom panel, the solid red line shows the distribution of $\theta_{kl}^{(2)}$ associated with Imported Beer and the dotted green line shows the distribution associated with Domestic Beer.

8 Discussion

This paper studies the specification of Bayesian shrinkage priors for high-dimensional demand systems. The flexibility of a demand system is important in large-scale settings to be able to accommodate a wide mix of substitutable, complementary, and unrelated goods. Because flexible models by definition have many parameters, they will often require regularization in high dimensions. However, regularized estimators typically assume a sparse underlying parameter vector, which can be at odds with many economic properties of cross-price effects. We therefore propose a hierarchical extension of the class of global-local priors to allow for structured, non-sparse shrinkage. In particular, our prior allows the elasticity between two goods to be shrunk towards higher-level group effects rather than zero. To define product groups, we use the retailer product classification trees that commonly accompany retail scanner data. We provide formal results regarding the shrinkage properties of these hierarchical priors and show that they ultimately behave like the original, non-hierarchical prior. For example, the hierarchical horseshoe will have the same heaviness (i.e., tail robustness) as the standard horseshoe.

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. Even under the assumption of a diagonal error variance matrix, a standard sampler will have complexity $\mathcal{O}(p^4)$. Our proposed sampler has complexity $\mathcal{O}(n^2p^2)$ which provides significant gains in $p > n$ settings.

We then use both simulated data and actual retail scanner data to show the importance of allowing for non-sparse shrinkage. In our simulation experiments, the hierarchical prior structure provides superior predictive fit when the true parameters are dense but conform to a low-dimensional structure. If the true parameters are indeed sparse, then typical sparsity-inducing priors perform best. In our empirical setting, we estimate demand for $p = 245$ goods that span 10 product categories. We find that a hierarchical ridge prior provides the best predictive performance, which supports the argument that “off-the-shelf” sparsity-inducing priors may not be ideal for price elasticities. The hierarchical nature of our proposed priors also allows us to learn about within and cross-subcategory/category effects, which is useful for identifying boundaries of competition.

There are many possible extensions of the current work. First, although we have focused on the log-linear demand system, we believe that hierarchical shrinkage priors have broader applicability. For example, flexible demand systems based on quadratic or translog utility (e.g., the Almost Ideal Demand System) contain high-dimensional parameter vectors representing product-level interactions. The same is true for discrete choice models for bundles which contain a $p \times p$ matrix of “demand synergy” parameters. When there are many products, it may be useful to employ priors that direct shrinkage based on some known hierarchical grouping structure. Our hierarchical priors may also be useful when researchers have prior beliefs about hierarchies of large vectors of product characteristics enter consumer utility.

So far we have also ignored the potential endogeneity of prices which would arise if retailers set prices as a function of demand characteristics not included as right-hand side variables our demand system. Finding good, valid instruments is challenging in general (Rossi, 2014), and is only made more difficult by the large number of prices in our high-dimensional model. If the focus was solely on the estimation of own-price elasticities, then the problem could be reposed as a treatment effect estimation problem with prices of other goods serving as a high-dimensional set of controls. The issue of regularized-induced bias is well known (see Hahn et al. 2018a for a discussion within the Bayesian paradigm), and this bias may only be exacerbated if shrinkage points are misspecified. We therefore believe that a fruitful way forward would be to estimate the joint system of demand and pricing equations using non-sparse shrinkage priors such as the hierarchical ones we propose.

References

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

Bajari, P., Nekipelov, D., Ryan, S. P., and Yang, M. (2014). Demand estimation with machine learning and model combination. *Working Paper*.

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

Barten, A. P. (1964). Consumer demand functions under conditions of almost additive preferences. *Econometrica*, 32(1-2):1–38.

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

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

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2):447–457.

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

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

Blackorby, C., Primont, D., and Russell, P. R. (1978). *Duality, Separability and Functional Structure: Theory and Economic Applications*. Elsevier North-Holland.

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

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

Byron, R. (1970). A simple method for estimating demand systems under separable utility assumptions. *The Review of Economic Studies*, 37(2):261–274.

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

Chiong, K. X. and 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. and Ghosh, J. K. (2013). Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132.

Deaton, A. and Muellbauer, J. (1980). *Economics and Consumer Behavior*. Cambridge University Press.

DellaVigna, S. and Gentzkow, M. (2019). Uniform pricing in US retail chains. *The Quarterly Journal of Economics*, 134(4):2011–2084.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Fessler, P. and Kasy, M. (2019). How to use economic theory to improve estimators: Shrinking toward theoretical restrictions. *Review of Economics and Statistics*, 101(4):681–698.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Ghosh, P. and Chakrabarti, A. (2017). Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Analysis*, 12(4):1133–1161.

Gillen, B. J., Shum, M., and Moon, H. R. (2014). Demand Estimation with High-Dimensional Product Characteristics. In Ivan Jeliazkov, D. J. P., editor, *Bayesian Model Comparison (Advances in Econometrics)*, volume 34. Emerald Group Publishing Limited.

Goldman, S. M. and Uzawa, H. (1964). A note on separability in demand analysis. *Econometrica*, 32(3):387–398.

Gorman, W. (1959). Separable utility and aggregation. *Econometrica*, 27(3):469–481.

Gorman, W. (1971). Two stage budgeting. *Unpublished Manuscript, London School of Economics.*

Griffin, J. and Brown, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12(1):135–159.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018a). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

Hahn, P. R., He, J., and Lopes, H. (2018b). Bayesian factor model shrinkage for linear IV regression with many instruments. *Journal of Business & Economic Statistics*, 36:278–287.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hausman, J., Leonard, G., and Zona, J. D. (1994). Competitive analysis with differentiated products. *Annales d'Economie et de Statistique*, 34:159–180.

Hausman, J. A. (1996). Valuation of new goods under perfect and imperfect competition. In Timothy F. Bresnahan, R. J. G., editor, *The Economics of New Goods*, pages 207–248. Univeristy of Chicago Press.

Hitsch, G. J., Hortaçsu, A., and Lin, X. (2019). Prices and promotions in US retail markets: Evidence from big data. *Working Paper*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2019). Dynamic shrinkage processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:781–804.

Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170.

Li, Y., Craig, B. A., and Bhadra, A. (2019a). 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., and Bhadra, A. (2019b). Joint mean-covariance estimation via the horseshoe with an application in genomic data analysis. *Working Paper*.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Mitchell, T. J. and 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. and Rossi, P. E. (1999). Estimating price elasticities with theory-based priors. *Journal of Marketing Research*, 36(4):413–423.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics*, volume 9. Oxford University Press.

Polson, N. G. and Scott, J. G. (2012). On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis*, 7(4):887–902.

Pudney, S. E. (1981). An empirical method of approximating the separable structure of consumer preferences. *The Review of Economic Studies*, 48(4):561–577.

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.

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2017). Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels. *Working Paper*.

Smith, A. N., Rossi, P. E., and Allenby, G. M. (2019). Inference for product competition and separable demand. *Marketing Science*, 38(4):690–710.

Strotz, R. H. (1957). The empirical implications of a utility tree. *Econometrica*, 25(2):269–280.

Strulov-Shlain, A. (2019). More than a penny's worth: Left-digit bias and firm pricing. *Working Paper*.

Tan, L., Chiong, K. X., and Moon, H. R. (2018). Estimation of High-Dimensional Seemingly Unrelated Regression Models. *Working Paper*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Train, K. (2003). *Discrete Choice Methods with Simulation.* Cambridge University Press.

van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618.

van der Pas, S. L., Solomond, J.-B., and Schmidt-Heiber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics*, 10:976–1000.

van der Pas, S. L., Szabó, B., and van der Vaart, A. W. (2017). Uncertainty quantification for the horseshoe. *Bayesian Analysis*, 12:1221–1274.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

APPENDIX

A Hierarchical Horseshoe Full Conditionals

Means $\theta_{kl}^{(\ell)}$

$$\begin{aligned}
\theta_{kl}^{(\ell)}|\text{else} &\sim \mathrm{N}\left(\tilde{\theta}_{kl}^{(\ell)}, V_{kl}^{(\ell)}\right) \\
\tilde{\theta}_{kl}^{(\ell)} &= V_{kl}^{(\ell)}\left(\sum_{\substack{a\in \mathrm{c}_\ell^{\ell+1}(k)\\ b\in \mathrm{c}_\ell^{\ell+1}(l)}} \frac{\theta_{ab}^{(\ell+1)}}{\lambda_{ab}^{2(\ell+1)}\Psi_{ab}^{(\ell)}\tau^{2(\ell+1)}} + \frac{\theta_{kl}^{(\ell-1)}}{\lambda_{kl}^{2(\ell)}\Psi_{kl}^{(\ell-1)}\tau^{2(\ell)}}\right) \\
V_{kl}^{(\ell)} &= \left(\sum_{\substack{a\in \mathrm{c}_\ell^{\ell+1}(k)\\ b\in \mathrm{c}_\ell^{\ell+1}(l)}} \frac{1}{\lambda_{ab}^{2(\ell+1)}\Psi_{ab}^{(\ell)}\tau^{2(\ell+1)}} + \frac{1}{\lambda_{kl}^{2(\ell)}\Psi_{kl}^{(\ell-1)}\tau^{2(\ell)}}\right)^{-1}
\end{aligned}$$

Local Variances $\lambda_{kl}^{2(\ell)}$

$$\begin{aligned}
\lambda_{kl}^{2(\ell)}|\xi_{\lambda_{kl}^{(\ell)}}, \text{else} &\sim \mathrm{IG}\left(\frac{1}{2} + \frac{1}{2}\sum_{s\geq\ell} n_\theta^{(s)}, \frac{1}{\xi_{\lambda_{kl}^{(\ell)}}} + \frac{(\theta_{kl}^{(\ell)} - \theta_{\mathrm{p}_\ell(k,l)}^{(\ell-1)})^2}{2\Psi_{kl}^{(\ell-1)}\tau^{2(\ell)}} + \frac{1}{2}\sum_{s>\ell}\sum_{\substack{a\in \mathrm{c}_\ell^{s}(k)\\ b\in \mathrm{c}_\ell^{s}(l)}} \frac{(\theta_{ab}^{(s)} - \theta_{\mathrm{p}_\ell(a,b)}^{(s-1)})^2}{\Psi_{ab}^{(s-1)}\tau^{2(s)}}\right) \\
\xi_{\lambda_{kl}}|\lambda_{kl}^{2(\ell)} &\sim \mathrm{IG}\left(1, 1 + \frac{1}{\lambda_{kl}^{2(\ell)}}\right)
\end{aligned}$$

Global Variances $\tau^{2(\ell)}$

$$\begin{aligned}
\tau^{2(\ell)}|\xi_{\tau^{(\ell)}}, \text{else} &\sim \mathrm{IG}\left(\frac{n_\theta^{(\ell)}+1}{2}, \frac{1}{\xi_{\tau^{(\ell)}}} + \frac{1}{2}\sum_{\substack{a\in(1,\ldots,n_\ell)\\ b\in(1,\ldots,n_\ell)}} \frac{\left(\theta_{ab}^{(\ell)} - \theta_{\mathrm{p}_\ell(a,b)}^{(\ell-1)}\right)^2}{\lambda_{ab}^{2(\ell)}\Psi_{ab}^{(\ell-1)}}\right) \\
\xi_{\tau^{(\ell)}}|\tau^{2(\ell)} &\sim \mathrm{IG}\left(1, 1 + \frac{1}{\tau^{2(\ell)}}\right)
\end{aligned}$$

B Fast Sampler with Nonzero Prior Mean

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