Demand Models with Random Partitions

Adam N. Smith and Greg M. Allenby *

March 12, 2019

Abstract

Many economic models of consumer demand require researchers to partition sets of products or attributes prior to the analysis. These models are common in applied problems when the product space is large or spans multiple categories. While the partition is traditionally fixed a priori, we let the partition be a model parameter and propose a Bayesian method for inference. The challenge is that demand systems are commonly multivariate models that are not conditionally conjugate with respect to partition indices, precluding the use of Gibbs sampling. We solve this problem by constructing a new location-scale partition distribution that can generate random-walk Metropolis-Hastings proposals and also serve as a prior. Our method is illustrated in the context of a store-level category demand model where we find that allowing for partition uncertainty is important for preserving model flexibility, improving demand forecasts, and learning about the structure of demand.

Keywords: Bayesian inference, location-scale family, Pólya urn, Markov chain Monte Carlo, price elasticity.

*Adam N. Smith, UCL School of Management, University College London, a.smith@ucl.ac.uk; Greg M. Allenby, Fisher College of Business, Ohio State University, allenby.1@osu.edu.

1 Introduction

At the core of any empirical demand analysis is the measurement of how consumers substitute between goods in response to changes in price, promotion, or other product features. Demand parameters are then commonly used as inputs to a variety of managerial tasks such as setting optimal prices and promotion schedules, determining the size and scope of assortments, and arranging physical or online product displays. A practical challenge to the accurate and precise measurement of consumer preferences is that the space of relevant products is often large, spanning multiple product categories or high-dimensional sets of product attributes.

One approach to modeling demand in this context is to partition the set of products or attributes into a lower-dimensional set of groups prior to the analysis. In fact, most empirical work begins with this problem as researchers must choose sets of products or categories to include and exclude from their analysis. Many demand models are also parameterized in a way that formally conditions on this partitioning of goods or attributes. Examples include models of separable demand (Strotz, 1957; Gorman, 1959; Goldman and Uzawa, 1964), cross-category demand (Chib et al., 2002; Song and Chintagunta, 2006; Mehta, 2007; Thomassen et al., 2017), and nested logit demand (McFadden, 1978; Train, 2009). In each case, the partition defines rigid boundaries for the ways that consumers perceive products to compete. The advantage offered by partitioning demand is one of dimension reduction – both for the researcher wanting to reduce the number of model parameters and for the retailer or brand manager wanting reduce their decision/action space to a lower-dimensional set of product groups. However, doing so can also lead to unrealistic substitution patterns and demand forecasts unless the right grouping structure is chosen a priori.

In this paper, we let the partition be an unknown model parameter and propose a Bayesian method for inference. Formally, let $\pi_n = (g_1, \ldots, g_n)$ denote a partition of n products or attributes where g_i indicates the group to which item i belongs. Our aim is to make joint inference about π_n and a vector of other demand parameters $\boldsymbol{\theta}$ (e.g., sensitivity to price or advertising) through the posterior distribution $p(\boldsymbol{\theta}, \pi_n|\text{data})$. This inference problem

is challenging for three reasons. First, it requires the specification of a probability model $p(\pi_n)$ that is defined on the space of possible partitions $\mathcal{P}_n$. Constructing coherent and useful probability models on $\mathcal{P}_n$ is generally difficult because the space is high-dimensional, non-Euclidean, and discrete. The second challenge is that π_n is defined over a correlated space of products rather than a conditionally independent space of data points. Therefore, and in contrast to most clustering applications, partitioned demand models are unlikely to be conditionally conjugate with respect to the set of item-group indicator variables $g_1, \ldots, g_n$. The third challenge, which is specific to the problem of joint inference, is that the dimension of $\boldsymbol{\theta}$ may depend on π_n. The consequence of latter two challenges is that the traditional Gibbs-style posterior sampling routines which move incrementally through the posterior by updating g_i conditional on g_{-i} and $\boldsymbol{\theta}$ will no longer apply.

Our main contribution is to develop a random partition model that facilitates efficient posterior sampling for multivariate partitioned demand systems. To do this, we apply and extend recent work on covariate-dependent random partition models – specifically Park and Dunson (2010) and Müller et al. (2011) – to develop a new model called the location-scale partition (LSP) distribution. The LSP model is characterized by a location partition $\rho_n \in \mathcal{P}_n$ and a scale parameter $\tau > 0$. Partitions sampled from the LSP(ρ_n, τ) distribution will be close to ρ_n with proximity measured by τ. The key innovation is that the location-scale feature allows us to implement a random-walk Metropolis-Hastings (MH) algorithm in which candidate partitions π_n^* are sampled from an LSP distribution centered around the current state $\pi_n^{(r)}$ with step size v. We can then accept/reject the entire partition vector in one joint update rather than n incremental updates, as is traditionally done in Gibbs-style updating. We also show how the LSP distribution can be used as a prior and extended to incorporate information from other observable covariates.

The value of the LSP distribution is demonstrated empirically using store-level grocery retailer data from the salty snack product category. We consider an aggregate demand model in which the partition of products identifies isolated demand groups where the isolated

condition restricts the *cross-group* price elasticities to be zero. By doing inference on the partition itself, we are able see how product groups with high posterior probability compare to retailer subcategories. We find that the differential shrinkage imposed by our model improves both estimates of price elasticity parameters as well as demand forecasts.

The remainder of this paper is organized as follows. Section 2 reviews related literature. Section 3 outlines the development of the LSP distribution. Section 4 provides a general MCMC routine for sampling from a posterior distribution using LSP proposals. Section 5 presents the results of our empirical application. Section 6 discusses limitations and possible extensions of the current work.

2 Related Literature

2.1 Random Partition Models

Random partition models have a long history, much of which is due to the development of Bayesian nonparametric models and methods (see Müller et al. 2015 for a review). Consider a hierarchical representation of the typical Bayesian nonparametric model:

$$y_i|\phi_i \sim p(y_i|\phi_i),\ \phi_i|G \sim G,\ G \sim Q \tag{1}$$

where $p(y_i|\phi_i)$ is the likelihood for observations $i = 1, \ldots, n$ indexed by unit-level parameters ϕ_i, G is a discrete random probability measure serving as a nonparametric prior for ϕ_i, and Q is the directing measure serving as a prior on the space of random probability measures. The fact that G is discrete gives rise to a clustering of the ϕ_i's and therefore induces a probability model over $\mathcal{P}_n$. For example, consider choosing Q to be the Dirichlet process (DP) of Ferguson (1973) with scaling parameter $\alpha > 0$ and base distribution G_0. The induced partitioning of the ϕ_i's can be seen using the Pólya urn representation of Blackwell and MacQueen (1973).

$$\begin{aligned} \phi_i|\phi_{<i} &\sim w_0 G_0(\phi_i) + \sum_{k=1}^{K^{(i)}} w_k \delta_{\phi_k^*}(\phi_i) \\ w_0 &= \left(\frac{\alpha}{\alpha+i-1}\right);\ w_k = \left(\frac{n_k}{\alpha+i-1}\right) \end{aligned} \tag{2}$$

Here the items $\phi_1, \ldots, \phi_n$ are generated sequentially where each ϕ_i is a new draw from the base distribution G_0 with probability w_0 or exactly equal to one of $k = 1, \ldots, K^{(i)}$ unique previous values ϕ_k^* with probability w_k. The weights satisfy $w_0 + \sum_{k=1}^{K^{(i)}} w_k = 1$ and n_k denotes the number ϕ_i's assigned to group k. The discreteness of G ensures that ties among the ϕ_i's occur with positive probability, so the vector $(\phi_1, \ldots, \phi_n)$ can be used to create a partition $\pi_n = (g_1, \ldots, g_n)$ by letting $g_i = k$ if $\phi_i = \phi_k^*$. In some cases it becomes more convenient to write $\pi_n = \{G_1, \ldots, G_K\}$ where $G_k = \{j : g_j = k\}$. In either case, this mapping from ϕ_i to g_i induces a valid probability model defined over $\mathcal{P}_n$ (Müller et al., 2015) and is often referred to as the Ewens distribution (Ewens, 1972; Pitman, 1995).

$$p(\pi_n) = \frac{\alpha^{K-1} \prod_{k=1}^{K} (n_k - 1)!}{(\alpha+1)\cdots(\alpha+n-1)} \tag{3}$$

There is a growing list of choices for $p(\pi_n)$ beyond that which is induced by the DP. One example is the class of species sampling models (SSMs) developed by Pitman (1995, 1996) which extends the DP by specifying the weights in (2) as nonnegative functions of the vector of cluster sizes $\boldsymbol{n} = (n_1, \ldots, n_{K^{(i)}})$.

$$\phi_i|\phi_{<i} \sim w_0(\boldsymbol{n}) G_0(\phi_i) + \sum_{k=1}^{K^{(i)}} w_k(\boldsymbol{n}) \delta_{\phi_k^*}(\phi_i) \tag{4}$$

This modified sampling scheme characterizes a species sampling sequence $\phi_1, \ldots, \phi_n$ and, like the Pólya urn scheme in (2), induces a model $p(\pi_n)$. In this case, the partitioning model takes the form $p(\pi_n) = p(n_1, \ldots, n_K)$ and depends on π_n only through the cluster sizes (Quintana, 2006). If the weights are chosen to be $w_k(\boldsymbol{n}) \propto n_k$ and $w_0(\boldsymbol{n}) \propto \alpha$ then the SSM reduces to the DP and the induced model $p(\pi_n)$ is the Ewens distribution in (3). Similarly, if $w_k(\boldsymbol{n}) \propto n_k - \delta$ and $w_0(\boldsymbol{n}) \propto \alpha + \delta K^{(i)}$ then the SSM reduces to the two-parameter

Poisson-Dirichlet process of Pitman and Yor (1997) and the induced model $p(\pi_n)$ is the Ewens-Pitman distribution.

Both the DP and the class of SSMs serve as nonparametric priors in (1) and induce exchangeable partition distributions in that $p(\pi_n)$ is invariant under permutations of the indices $\{1, \ldots, n\}$. Requiring $p(\pi_n)$ to be exchangeable matters in Bayesian nonparametric models because it guarantees that $p(\pi_n)$ can be rationalized by some underlying random measure Q. In some situations, however, insisting on exchangeability is not appropriate such as when the items being clustered have a natural ordering in time or space. Airoldi et al. (2014) relaxes this property and develops a family of nonexchangeable species sampling sequences in which the weights in (4) depend on realizations of latent variables instead of cluster sizes. Exchangeability is also relaxed in the Ewens-Pitman attraction (EPA) distribution of Dahl et al. (2017), where the weights in the species sampling sequence depend on pairwise distances between items. The EPA distribution closely resembles the distant-dependent Chinese restaurant process (ddCRP) of Blei and Frazier (2011), however the ddCRP defines a probability distribution over graphs instead of partitions and only indirectly defines a partitioning model $p(\pi_n)$. We revisit the comparison between these distributions in the next section.

The product partition models (PPMs) of Hartigan (1990) and Barry and Hartigan (1992) present another class of random partition models. Rather than define $p(\pi_n)$ by way of some underlying discrete random probability measure, a PPM defines $p(\pi_n)$ directly:

$$p(\pi_n) \propto \prod_{k=1}^{K} c(G_k) \tag{5}$$

where $c(G_k) \geq 0$ is a cohesion function measuring the similarity between the elements of G_k. Since $c(G_k)$ can be any nonnegative function, PPMs give rise to a very general class of partitioning models. Quintana and Iglesias (2003) show that PPMs nest the DP partitioning model as a special case if the cohesion function is chosen to be $c(G_k) = \alpha \times (|G_k| - 1)!$ where $|G_k| = n_k$ counts the number of items in group k. Moreover, Quintana (2006) shows that

if the cohesion functions depend on G_k only though the cluster sizes, then the PPM is exchangeable and a special case of the distribution induced by SSMs.

Extensions of PPMs have since been developed to account for different types of prior information on π_n. For example, Park and Dunson (2010) and Müller et al. (2011) modify the cohesion functions in (5) to allow for effects of covariates. Therefore, items that are closer in the covariate space will also have a higher probability of being grouped together a priori. This leads to a nonexchangeable PPM and is useful whenever the researcher wants to incorporate covariates into the prior model for π_n. The empirical application of Müller et al. (2011), for example, uses covariates like treatment dosage, age, and tumor size to better cluster and predict survival times of breast cancer patients in a clinical trial. In addition to covariate effects, there have also been extensions to PPMs that account for temporal dependence within clusters (Monteiro et al., 2011), correlations across clusters (Ferreira et al., 2014), and spatially dependent clusters (Page and Quintana, 2016).

In practice, the choice between partitioning priors is based on empirical context, data availability, and a tradeoff between flexibility and tractability. Empirical context can suggest whether the data or unit-level parameters should have any natural ordering, and can therefore offer guidance as to whether assumptions of exchangeability should hold. For example, the development of the ddCRP in Blei and Frazier (2011) is in part motivated by a language modeling application in which news articles published around the same time should tend to be more similar. When item-level covariates are available, researchers may also want to exploit them when specifying $p(\pi_n)$ akin to the covariate-dependent PPMs or distance-based models. As the partitioning model becomes more flexible or enriched with additional prior information, however, the challenge is to ensure that posterior sampling remains feasible. This is why the random partition model induced by the DP, even with its relative inflexibility, remains a popular choice, as it is a model for which posterior simulation methods are well developed (Müller et al., 2015).

In summary, this stream of literature has generated more flexible classes of partitioning

models, some of which we directly build on when constructing the LSP model. Beyond this methodological overlap, however, our work differs in three ways. First, the demand models we estimate are not Bayesian nonparametric models in the sense of (1). Instead, we consider parametric models of the form:

$$\boldsymbol{y}_t|\boldsymbol{\theta}, \pi_n \sim p(\boldsymbol{y}_t|\boldsymbol{\theta}, \pi_n), \;\; \boldsymbol{\theta}|\pi_n \sim p(\boldsymbol{\theta}_{\pi_n}), \;\; \pi_n \sim p(\pi_n) \tag{6}$$

where the response $\boldsymbol{y}_t = (y_{t1}, \ldots, y_{tn})$ is a vector of demand across n products at time t, $p(\boldsymbol{y}_t|\boldsymbol{\theta}, \pi_n)$ is the likelihood indexed by a vector of demand parameters $\boldsymbol{\theta}$ (whose dimension is independent of t) and the partition π_n, and both the conditional prior of $\boldsymbol{\theta}|\pi_n$ and marginal prior of π_n are specified parametrically. We specify $\boldsymbol{\theta}$ conditional on π_n to allow potential dependence between the two. The y_{ti}'s are assumed to be conditionally iid over time periods $t = 1, \ldots, T$ but not over products $i = 1, \ldots, n$. This reflects the fact that products may exhibit unobservable similarities which could manifest themselves through some correlation structure in the model likelihood.

The second way our work differs is based on how the clustering is imposed. In (1), the clustering arises through ties in the *unit-level* parameters $\phi_1, \ldots, \phi_n$ which are induced through the discrete random probability measure G. This implies a partitioning of the space of observational units. In contrast, the partitioning in (6) is not imposed on the space of observational units but rather the space of products. The clustering arises through π_n directly and there is no notion of a random measure G from which elements of $\boldsymbol{\theta}$ are drawn or clustered. For our purposes, the process of sampling from a random probability measure as in (2) only serves to define the LSP model and is entirely independent from the assumed distributions of $\boldsymbol{y}_t$ or $\boldsymbol{\theta}$.

Lastly, many of the posterior sampling methods that exist for models of the form in (1) are not applicable to the model structure in (6). For example, the usual suite of Gibbs sampling routines associated with DP mixture models (Escobar and West, 1995; MacEachern and Müller, 1998; Neal, 2000) rely on closed-form expressions of $p(g_i = k|g_{-i}, \boldsymbol{\theta}, \text{else})$. However,

this expression is not well-defined in our case because of the potential dependence between $\boldsymbol{\theta}$ and π_n and because we partition the correlated rather than the conditionally independent dimension of the data. Therefore, while the literature mentioned above has mostly focused on the development of more flexible partitioning priors, we develop the LSP model mainly as a computational device to facilitate posterior sampling from the class of models in (6).

More general Metropolis-Hastings based posterior sampling routines for partitions have also been proposed, including Algorithms 5-7 in Neal (2000) and the split-merge algorithms of Green and Richardson (2001), Jain and Neal (2004), and Dahl (2003). Since these methods are based on MH updates, they can in principle be used to sample from the posteriors of partitioned demand models. However, these methods are still characterized by (group-wise) incremental moves in which the proposed partition can differ from the partition in the current state by one item (or group of items) at a time. In contrast, LSP proposals allow for more radical restructuring in which multiple items can change groups in each update (where the extent of the difference is controlled by a step size parameter). In a simulation study, we find that LSP proposals offer advantages in mixing over incremental updates when partitions with high posterior probability are separated by valleys of low posterior probability.

2.2 Partitioned Demand Models

One of the earliest examples of partitioning demand comes from the work on economic separability by Strotz (1957), Gorman (1959), and Goldman and Uzawa (1964) among others. Separability offers a set of conditions under which a consumer's utility function defined over a n-dimensional commodity bundle can be expressed with respect to a lower-dimensional set of product groups. The empirical advantage of assuming separable demand is a reduction in the number of the cross-price effect parameters. However, the consequence is that separability implies strong restrictions on cross-group substitution patterns when the partitioning of goods is fixed. While this limitation is well-known (Deaton and Muellbauer, 1980; Pudney, 1981), there has been little progress in the way of formal inference on π_n.

Another form of partitioning arises in the class of demand models with structured product covariance matrices. To illustrate, consider the model $\boldsymbol{y}_t = h(\boldsymbol{X}_t, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_t$ where $\boldsymbol{y}_t$ is an n-vector of demand at time t, $\boldsymbol{X}_t$ is $n \times p$ matrix of product covariates like price, $\boldsymbol{\theta}$ is a p-vector of demand parameters like price-effects, and $\text{Var}(\boldsymbol{\varepsilon}_t) = \Sigma$. This stream of work either parameterizes Σ or the prior covariance matrix on $\boldsymbol{\theta}$ to account for similarities among products due to shared attributes or correlated unobservables. Examples of adding structure to an error covariance matrix include nested logit models (McFadden, 1978; Train, 2009) and the source-of-volume probit model (Dotson et al., 2018). Ainslie and Rossi (1998) and Hansen et al. (2006) provide examples of adding structure to the prior covariance matrix to capture correlated preferences across categories.

In either case, our methodology applies whenever the covariance matrix is parameterized by a "hard constraint" (e.g., related or not). Our belief is that allowing for uncertainty in models with hard constraints like partitions will yield a level of flexibility similar to that of continuous parameterizations of item similarity. An additional benefit of modeling hard constraints is that they may more closely map to the decision space of managers who must think in discrete terms (e.g., deciding which products should be grouped together on a shelf).

Our approach can also be used to model grouping structures that are imposed by way of the conditional mean function. That is, instead of parameterizing a prior or error covariance matrix, one could directly impose restrictions on the functional form $h(\boldsymbol{X}_t, \boldsymbol{\theta})$ or on the model parameters themselves. Examples of this approach include adding restrictions to the functional form of utility (Song and Chintagunta, 2006; Mehta, 2007; Kim et al., 2017) as well as price-effect parameters (Montgomery and Rossi, 1999; Wedel and Zhang, 2004). Our aim is to provide an inference method for the class of partitioned demand models, regardless of whether the partition enters the model through a covariance matrix or restrictions to parameters and functional forms.

Finally, there is relatively little work which allows for partition uncertainty in applications to marketing and econometric models. One exception is Hui and Bradlow (2012), who

estimate partitions of contiguous areal units, such as states and retail store shopping zones, in the context of a multi-resolution spatial analysis. Although their application is outside of the scope of a traditional demand analysis, their work is similar in spirit, as they relax the assumption that the spatial configuration is known a priori and propose a method for estimation. However, one primary difference is that they use a simulating annealing algorithm to search for the partition with highest posterior probability. This approach will generate a point estimate of the partition parameter, but is unsuitable for inference. Our belief is that inference for partitioned demand models is desirable, especially when model output is used to inform policy decisions. For example, a maximum a posteriori estimate of π_n could still have very low posterior probability since the space of possible partitions is so large. Managerial actions that condition on a point estimate in the presence of great uncertainty will likely be suboptimal. Moreover, inference allows managers to improve demand forecasts by integrating over the posterior of π_n, akin to classic Bayesian model averaging.

3 The Location-Scale Partition Distribution

The LSP distribution is constructed using a variant of the Pólya urn scheme in (2). First define a location partition $\rho_n = (s_1, \ldots, s_n) \in \mathcal{P}_n$ and scale parameter $\tau > 0$. We then modify (2) so that the ϕ_i's are generated using the information in (ρ_n, τ).

$$\phi_i | \phi_{<i}, \rho_n, \tau \sim w_0(\rho_n, \tau) G_0(\phi_i) + \sum_{k=1}^{K^{(i)}} w_k(\rho_n, \tau) \delta_{\phi_k^*}(\phi_i) \tag{7}$$

Here $w_0(\cdot)$ and $w_k(\cdot)$ are positive similarity functions that satisfy $w_0(\cdot) + \sum_k w_k(\cdot) = 1$. Just as before, the partition π_n is formed by letting $g_i = k$ if $\phi_i = \phi_k^*$. The sequential nature of (7) imposes an order restriction on partitions in $\mathcal{P}_n$: $g_1 = 1$ and $g_i \in \{1, \ldots, K^{(i)} + 1\}$ where $K^{(i)} = \max\{g_j : j < i\}$. In contrast to (2), the composition of groups is now controlled by the information in ρ_n and τ. Holding τ fixed, we choose $w_0(\cdot)$ and $w_k(\cdot)$ so that items that are grouped together in ρ_n are also more likely to be grouped together in π_n.

3.1 Similarity Functions

Our choice of functional forms for $w_0(\cdot)$ and $w_k(\cdot)$ follows from Park and Dunson (2010) and Müller et al. (2011) who develop covariate-dependent PPMs. The idea is to let the item-group assignment probabilities be defined by an auxiliary probability model for the elements of ρ_n.

$$w_0(\rho_n, \tau) \equiv w_0(s_i, \tau) = \tilde{c}_i \int p(s_i|\boldsymbol{\xi}) f_0(\boldsymbol{\xi}|\tau) d\boldsymbol{\xi} \tag{8}$$

$$w_k(\rho_n, \tau) \equiv w_k(\{s_i, S_k\}, \tau) = \tilde{c}_i \int p(s_i|\boldsymbol{\xi}) f_k(\boldsymbol{\xi}|\tau, S_k) d\boldsymbol{\xi} \tag{9}$$

Here $S_k = \{s_j : g_j = k \text{ and } j < i\}$, $C^{(i)} = \max\{s_1, \ldots, s_{i-1}\}$, $\boldsymbol{\xi}$ is a $(C^{(i)} + 1)$-dimensional vector, and $\tilde{c}_i$ is a normalizing constant. In the context of Park and Dunson (2010) and Müller et al. (2011), we take elements of ρ_n to be "covariates" and then define similarity using a marginal probability model for s_i. Since each $s_i \in \{1, \ldots, C^{(i)} + 1\}$, we specify the following Dirichlet-categorical model:

$$p(s_i|\boldsymbol{\xi}) = \text{Cat}(\xi_1, \ldots, \xi_{C^{(i)}}, \xi_{C^{(i)}+1}) \tag{10}$$

$$f_0(\boldsymbol{\xi}|\tau) = \text{Dir}(\tau_1, \ldots, \tau_{C^{(i)}}, \tau_{C^{(i)}+1}) \tag{11}$$

$$f_k(\boldsymbol{\xi}|\tau, S_k) = \text{Dir}(\tau_1^*, \ldots, \tau_{C^{(i)}}^*, \tau_{C^{(i)}+1}^*) \tag{12}$$

where $\tau_c = \tau$, $\tau_c^* = \tau + n_{S_k}^c$, and $n_{S_k}^c$ counts the number of elements in S_k equal to c. We also let n_k denote the number of items in group k. The main advantage of specifying this conjugate family of models is that the similarity functions have closed-form expressions.

$$w_0(s_i, \tau) \propto \int \text{Cat}\big(\xi_1, \ldots, \xi_{C^{(i)}+1}\big) \text{Dir}\big(\tau_1, \ldots, \tau_{C^{(i)}+1}\big) d\boldsymbol{\xi} = \frac{\tau + 1(s_i = C^{(i)} + 1)}{\tau C^{(i)} + \tau + 1} \tag{13}$$

$$w_k(\{s_i, S_k\}, \tau) \propto \int \text{Cat}\big(\xi_1, \ldots, \xi_{C^{(i)}+1}\big) \text{Dir}\big(\tau_1^*, \ldots, \tau_{C^{(i)}+1}^*\big) d\boldsymbol{\xi} = \frac{\tau + n_{S_k}^{s_i}}{\tau C^{(i)} + \tau + n_k} \tag{14}$$

3.2 Properties of the LSP Distribution

We let LSP(ρ_n, τ) denote the probability distribution for π_n that is induced by (7) with similarity functions defined by (13) and (14). The sequential nature of the Pólya urn scheme in (7) provides a simple structure for computing the LSP probability mass function. The probability of observing a partition π_n from an LSP(ρ_n, τ) distribution can be factored into a sequence of conditional probabilities:

$$p(\pi_n|\rho_n, \tau) = \prod_{i=1}^{n} p(g_i|g_{<i}, \rho_n, \tau) \tag{15}$$

where $p(g_1) = 1$ and

$$p(g_i|g_{<i}, \rho_n, \tau) = \begin{cases} \tilde{c}_i \cdot \dfrac{\tau + n_{S_k}^{s_i}}{\tau C^{(i)} + \tau + n_k} & \text{if } i \text{ is assigned to group } k \\ \tilde{c}_i \cdot \dfrac{\tau + 1(s_i = C^{(i)} + 1)}{\tau C^{(i)} + \tau + 1} & \text{if } i \text{ starts a new group.} \end{cases} \tag{16}$$

Our parameterization of the similarity functions also leads to two noteworthy properties. First, $w_0(\cdot)$ and $w_k(\cdot)$ guarantee that the resulting LSP distribution behaves like a location-scale family. That is, as the scale parameter gets small, more mass is placed on the location partition. We refer to this property as location-scale consistency.

Property 1 (Location-Scale Consistency)**.** *If $\pi_n \sim$ LSP(ρ_n, τ), then for any number of items n and location partition $\rho_n \in \mathcal{P}_n$,*

$$\lim_{\tau \to 0} \Pr(\pi_n = \rho_n|\rho_n, \tau) = 1. \tag{17}$$

Proof. See Appendix A.

This property is illustrated in Figure 1. We generate $10,000$ samples from an LSP(ρ_n, τ) distribution with $n = 100$, ρ_n equal to the partition with five contiguous groups of twenty items each, and $\tau \in \{0.05, 0.5, 5\}$. For each value of τ, we then plot the associated $n \times n$ pairwise similarity matrix, which counts the proportion of times that two items are grouped

together in the given set of draws. As suggested by the location-scale consistency property, the LSP distribution shifts its mass towards ρ_n as τ gets small, but spreads its mass across $\mathcal{P}_n$ as τ gets large.

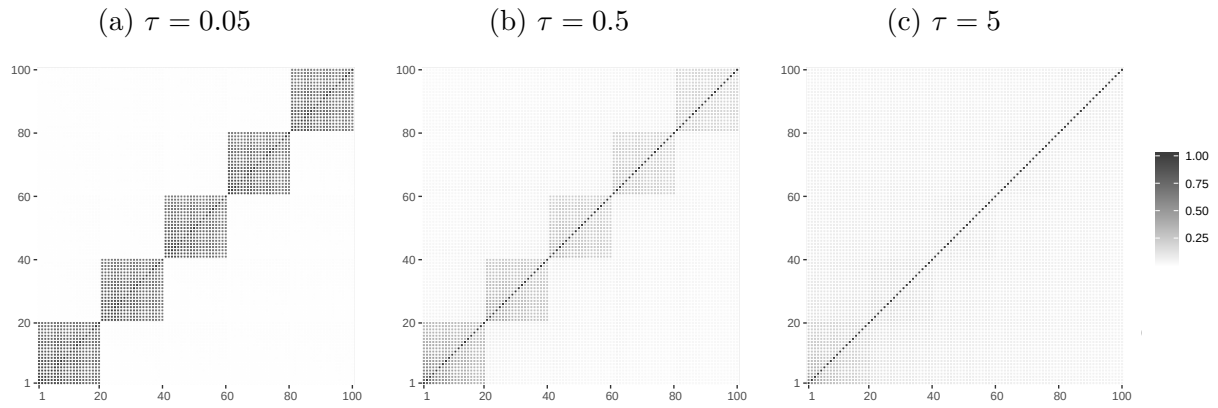


Figure 1: Three pairwise similarity matrices are shown based on $10,000$ draws from an LSP(ρ_n, τ) distribution with $n = 100$, ρ_n equal to the partition with five contiguous groups of twenty items each, and $\tau \in \{0.05, 0.5, 5\}$.

Moreover, we find that as τ gets small, $\Pr(\pi_n = \rho_n | \rho_n, \tau)$ not only converges to one but also does so monotonically. To understand why, observe that when ρ_n equals π_n, the conditional probabilities in (16) reduce to

$$p(g_i | g_{<i}, \rho_n, \tau) = \begin{cases} \tilde{c}_i \cdot \dfrac{\tau + n_k}{\tau C^{(i)} + \tau + n_k} & \text{if } i \text{ is assigned to group } k \\ \tilde{c}_i \cdot \dfrac{\tau + 1}{\tau C^{(i)} + \tau + 1} & \text{if } i \text{ starts a new group} \end{cases} \tag{18}$$

which are both positive, monotonically decreasing functions in τ. Since $\Pr(\pi_n = \rho_n | \rho_n, \tau)$ is defined as a product of the conditional probabilities $p(g_i | g_{<i}, \rho_n, \tau)$, it itself will be monotonically decreasing in τ.

The second property is marginal invariance. That is, it can be shown that the LSP distribution for a partition π_n can be obtained by marginalizing out item g_{n+1} from the LSP distribution for π_{n+1}.

Property 2 (Marginal Invariance)**.** *If $\pi_n \sim$ LSP(ρ_n, τ), then for any number of items*

n, location partition $\rho_n \in \mathcal{P}_n$*, scale parameter* $\tau > 0$*, and distribution* $p(s_{n+1})$ *such that* $\sum_{s_{n+1}} p(s_{n+1}) = 1$*,*

$$p(\pi_n|\rho_n, \tau) = \sum_{g_{n+1}=1}^{K+1} \sum_{s_{n+1}=1}^{C+1} p(\pi_{n+1}|\rho_n, s_{n+1}, \tau)p(s_{n+1}) \tag{19}$$

where $K = \max\{g_1, \ldots, g_n\}$ *and* $C = \max\{s_1, \ldots, s_n\}$*.*

Proof. See Appendix A.

This result is also consistent with Park and Dunson (2010) and Müller et al. (2011), who separately show that the only way for a Pólya urn induced partition distribution $p(\pi_n)$ to satisfy marginal invariance is if $w_0(\cdot)$ and $w_k(\cdot)$ take the form of an auxiliary probability model, as in (8) and (9).

3.3 A Covariate-Dependent LSP Model

In addition to having prior knowledge about the partition itself, researchers may have access to other data that could help inform how the items are split into groups. For example, if the partition represents a measure of economic competition between goods, then one may want to let the partitioning model be a function of covariates such as product characteristics, advertising levels, and in-store shelf position. This would allow products that are more similar in the covariate space to be more likely to be grouped together a priori.

Our approach to accommodating covariates in the LSP model again follows from Park and Dunson (2010) and Müller et al. (2011). Suppose the researcher has p continuous covariates for the n items to be partitioned. Let $\boldsymbol{X} \in \mathbb{R}^p$ be an $n \times p$ matrix and $\Lambda = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p_+$ be a vector of p scale parameters measuring the strength of the influence of $\boldsymbol{X}$ on the partitioning process. We then modify the original Pólya urn scheme in (7) to account for the additional information in $(\boldsymbol{X}, \Lambda)$.

$$\phi_i|\phi_{<i}, \rho_n, \boldsymbol{X}, \tau, \Lambda \sim \tilde{w}_0\big(\{\rho_n, \boldsymbol{X}\}, \{\tau, \Lambda\}\big)G_0(\phi_i) + \sum_{k=1}^{K^{(i)}} \tilde{w}_k\big(\{\rho_n, \boldsymbol{X}\}, \{\tau, \Lambda\}\big)\delta_{\phi_k^*}(\phi_i) \tag{20}$$

Here $\{\rho_n, \boldsymbol{X}\}$ contain the "location" terms and $\{\tau, \Lambda\}$ contain the "scale" terms. The new similarity functions $\tilde{w}_0(\cdot)$ and $\tilde{w}_k(\cdot)$ are positive and satisfy $\tilde{w}_0(\cdot) + \sum_k \tilde{w}_k(\cdot) = 1$, with functional forms given by:

$$\tilde{w}_0\big(\{\rho_n, \boldsymbol{X}\}, \{\tau, \Lambda\}\big) = w_0(s_i, \tau) \prod_{j=1}^{p} w_0^x(x_{ij}, \lambda_j) \tag{21}$$

$$\tilde{w}_k\big(\{\rho_n, \boldsymbol{X}\}, \{\tau, \Lambda\}\big) = w_k(\{s_i, S_k\}, \tau) \prod_{j=1}^{p} w_k^x(\{x_{ij}, X_{jk}\}, \lambda_j) \tag{22}$$

where $X_{jk} = \{x_{ij} : g_i = k\}$, $w_k(\cdot)$ and $w_0(\cdot)$ are the location-scale similarity functions defined in (8) and (9), and $w_k^x(\cdot)$ and $w_0^x(\cdot)$ are new functions that measure the similarity among each of the $j = 1, \ldots, p$ covariates.

The covariate similarity functions $w_k^x(\cdot)$ and $w_0^x(\cdot)$ are again defined as marginal distributions from an auxiliary probability model for each x_{ij}. With real-valued covariates, we specify the following conjugate normal-inverse-gamma family of models.

$$x_{ij}|\mu_\xi, \sigma_\xi, \lambda_j \sim \mathrm{N}(\mu_\xi, \sigma_\xi^2/\lambda_j) \tag{23}$$

$$\mu_\xi|\sigma_\xi, \lambda_j \sim \mathrm{N}(m, \sigma_\xi^2/\lambda_j) \tag{24}$$

$$\sigma_\xi^2 \sim \Gamma^{-1}(a, b) \tag{25}$$

Marginalizing over $\boldsymbol{\xi} = (\mu_\xi, \sigma_\xi^2)$ in a normal-inverse-gamma model gives rise to a noncentral t-distribution for x_{ij}:

$$w_0^x(x_{ij}, \lambda) \propto \int p(x_{ij}|\boldsymbol{\xi}) f_0(\boldsymbol{\xi}|\lambda_j) d\boldsymbol{\xi} = t_{2a}\left(m, \frac{(\lambda_j + 1)}{a\lambda_j} b\right) \tag{26}$$

$$w_k^x(\{x_{ij}, X_{jk}\}, \lambda_j) \propto \int p(x_{ij}|\boldsymbol{\xi}) f_k(\boldsymbol{\xi}|\lambda_j, X_{jk}) d\boldsymbol{\xi} = t_{2a+n_k}\left(\tilde{\mu}_j, \frac{(\lambda_j + n_k + 1)}{(a + n_k/2)(\lambda_j + n_k)} \tilde{\sigma}_j^2\right) \tag{27}$$

where

$$\tilde{\mu}_j = \frac{\lambda_j m + n_k \bar{x}_{jk}}{\lambda_j + n_k} \tag{28}$$

$$\tilde{\sigma}_j^2 = b + \sum_{i \in G_k} (x_{ij} - \bar{x}_{jk})^2 + \frac{n_k}{\lambda_j + n_k} (\bar{x}_{jk} - m)^2. \tag{29}$$

Because the covariate similarity functions are constructed by way of assuming a parametric distribution on x_{ij}, there is a notion that we are treating covariates as random variables. However, the use of conjugate parametric families here is done more out of convenience. That is, the objective is to construct a function $w_k^x(\cdot)$ such that similar values of the covariates induce higher values of $w_k^x(\cdot)$, and marginalized probability models in the form of (26) and (27) satisfy this condition.[1] Moreover, Müller et al. (2011) show that the similarity function must be in the form of a marginalized probability model in order for the induced partition distribution to satisfy marginal invariance. Thus, the form of the proposed covariate similarity functions can still be useful even if the researcher wants to treat the covariates as fixed.

In the presence of binary, categorical, or integer-valued covariates, other conjugate families such as the beta-binomial, Dirichlet-categorical, and gamma-Poisson can be used. Again, conjugate families are used only to simplify the integration present in (26) and (27). If the researcher wanted to write down a more judicious choice of $p(x_{ij}|\boldsymbol{\xi})$ for which conjugate priors did not exist, we follow Park and Dunson (2010) and suggest using an approximation to the marginal likelihood (e.g., the Laplace approximation).

The behavior of the covariate-dependent LSP (LSPx) distribution is illustrated in Figure 2. We consider an LSP$\big(\{\rho_n, \boldsymbol{X}\}, \{\tau, \Lambda\}\big)$ distribution with $n = 100$, ρ_n equal to the partition with five contiguous groups of twenty items each, $p = 1$, and $x_i \sim \text{N}(1(i > 50), .01)$. The covariates are then scaled to have mean zero and unit variance. We let $\tau \in \{0.05, 0.5, 5\}$ and $\lambda \in \{0.05, 0.5, 5\}$ and consider all nine possible pairs (τ, λ). For each pair, we generate 10,000 draws from the corresponding LSPx distribution and plot the associated pairwise similarity matrices. As shown in Figure 2, we find that when τ is small relative to λ (lower-diagonal), more weight is given to ρ_n. Conversely, when λ is small relative to τ (upper-diagonal), then more weight is given to the covariates $\boldsymbol{X}$. Finally, as τ and λ jointly increase (diagonal), then probability mass starts to spread more evenly across the full space of partitions.

[1] For example, the variance expression in (27) decreases as x_{ij} gets closer to $\bar{x}_{jk}$, implying that the density evaluation increases in the similarity between x_{ij} and X_{jk}.

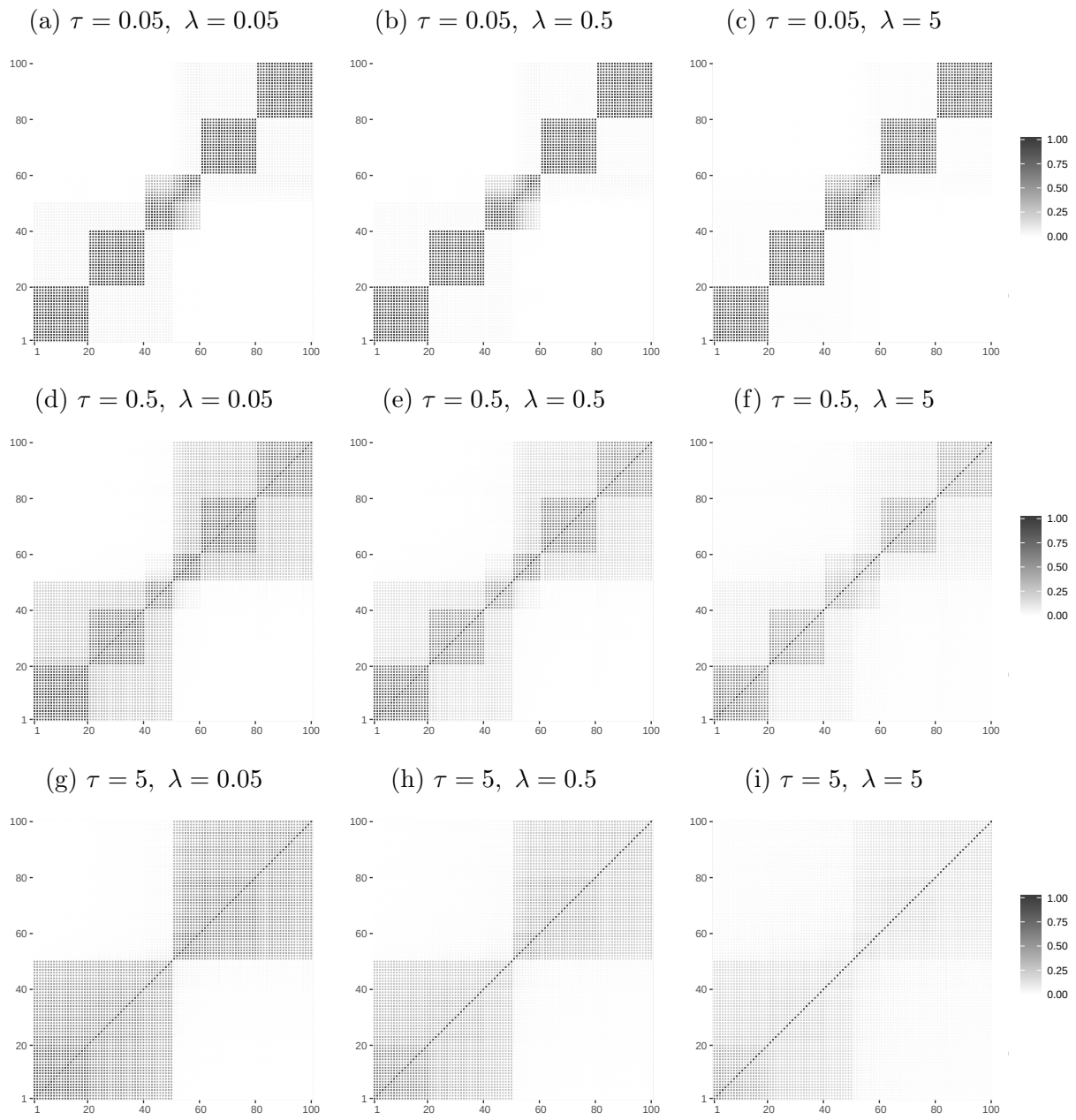


Figure 2: Pairwise similarity matrices are shown based on 10,000 draws from an LSPx distribution with $n = 100$, ρ_n equal to the partition with five contiguous groups of twenty items each. The covariates are drawn as $x_i \sim \mathrm{N}(1(i > 50), .01)$ and we consider all possible pairs of $\tau \in \{0.5, 5\}$ and $\lambda \in \{0.5, 5\}$.

3.4 Comparison to Other Partition Distributions

We now compare the LSP distribution to three other partition distributions: (1) the partition distribution induced by the standard DP; (2) the ddCRP of Blei and Frazier (2011); and (3) the EPA distribution of Dahl et al. (2017). The ddCRP and EPA distributions are of particular interest, as they can generate partitions centered around a prespecified grouping structure. The ddCRP distribution is parameterized by a mass parameter $\alpha > 0$, an $n \times n$ distance matrix $\boldsymbol{D} = \{d_{ij}\}$, and a decay function $f(d_{ij})$ controlling the degree to which the pairwise distances affect the resulting distribution over partitions. The EPA distribution is also indexed by a distance matrix $\boldsymbol{D}$ and decay function $f(d_{ij})$, as well as a discount parameter $\delta \in [0, 1)$ controlling the distribution of group sizes and a mass parameter $\alpha > -\delta$ controlling the number of groups. In both cases, the pairwise distances can be parameterized in order to center and scale the resulting distribution around a particular partition.

Comparisons among partition distributions are often made by assessing differences in the induced distribution over group numbers, sizes, or composition. Given that one of our primary goals is posterior sampling via random-walk-type MH proposals, we will study how well other models can be "centered" around a partition of interest. We do this with a simulation study in which we set the number of items to $n = 10$, the location partition to $\rho_n = (1, 1, 1, 1, 1, 2, 2, 2, 2, 2)$, and consider scale parameters in the range $\tau \in (0, 5)$. For each value of τ, we generate 10,000 draws from the LSP, EPA, ddCRP, and DP distributions and compare each draw to ρ_n using the adjusted Rand index (Hubert and Arabie, 1985), where a value of 1 indicates equality. For both the EPA and ddCRP distributions, we define pairwise distances based on ρ_n: if i and j are grouped together in ρ_n, then $d_{ij} = 0$, otherwise $d_{ij} = 1$. We also use an exponential decay function $f(d_{ij}) = e^{-d_{ij}/\tau}$ in both cases so that the pairwise distances have more influence on the partitioning process as τ gets small. The mass parameter in the EPA, ddCRP, and DP models is set to $\alpha = \max(\rho_n) = 2$ and the discount parameter for the EPA model is set to $\delta = 0$.

Figure 3 plots the average adjusted Rand index across τ for each partition distribution.

We find that the LSP, EPA, and ddCRP distributions behave similarly for larger values of τ, but differ as τ gets small. The location-scale consistency property of the LSP distribution guarantees that this similarity measure converges to 1 as τ gets small. In comparison, the similarity measure for the EPA and ddCRP distributions also increases as τ decreases, but does not have the same limiting behavior near 0. As expected, the DP partitioning distribution is uniform over τ because it is not parameterized in a way that allows it to be centered around a particular grouping structure.

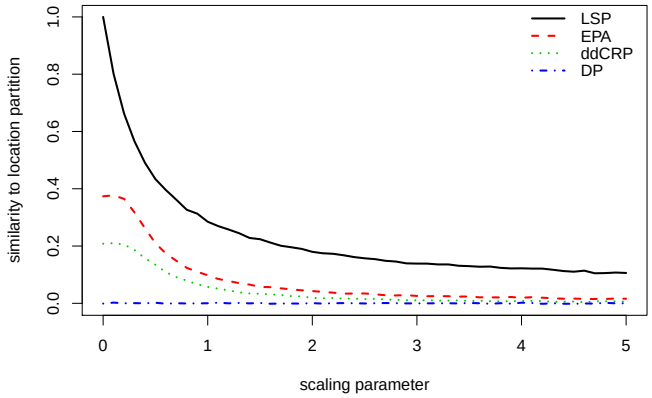


Figure 3: The LSP distribution is compared to the EPA (with $\alpha = 2$, $\delta = 0$, and exponential decay), ddCRP (with $\alpha = 2$ and exponential decay), and DP (with $\alpha = 2$) partition distributions. For each scaling parameter $\tau \in (0, 5)$, 10,000 random partitions are drawn from each distribution and then compared to ρ_n using the adjusted Rand index.

Because the EPA and ddCRP also contain other model parameters and require a parameterization of the distance function, the role of the scaling parameter may differ across the models we consider. It is then not clear as to whether the limiting behavior shown in Figure 3 is due to the statistical properties of the models or our own modeling choices (e.g., fixing $\alpha = 2$). We therefore carry out several robustness checks in Appendix B. For the ddCRP distribution, we vary the mass parameter α with τ and show that when τ is small, the mean similarity to ρ_n is still far from 1 for any $\alpha \in (0, 5)$. Similarly, for the EPA distribution, we vary the mass parameter α and discount parameter δ with τ and again show that the similarity to ρ_n remains far from 1 for any $\alpha \in (0, 5)$ and $\delta \in [0, 1)$. Although not reported,

we have also repeated the analysis using a reciprocal decay function for the EPA distribution and found little difference in the model's limiting behavior.

In summary, we find that the LSP distribution behaves similarly to distance-based partitioning models like the EPA and ddCRP distributions. These models are all parameterized in a way that allows the researcher to center the partitioning process around some fixed grouping structure. As a prior, the EPA and ddCRP models are generally more flexible than the LSP distribution since they are both indexed by a full $n \times n$ distance matrix. The single location partition in the LSP prior would (after appropriately reordering items) effectively correspond to a block-diagonal distance matrix. However, this can be overcome by using the LSPx model where covariates, in addition to the location partition, impact the prior probability of two items being grouped together. The idea is that the presence of multiple covariates creates a mixture over different grouping structures (as seen in equations 21 and 22) which in turn induces an $n \times n$ distance matrix that is no longer block-diagonal. The differences between the LSP and EPA models are most apparent in terms of posterior computation. Specifically, we find that the limiting behavior of the LSP distribution makes it especially well-suited for random-walk MH proposal schemes.

4 Posterior Computation

Given the LSP probability model, we now describe approaches for sampling from a joint posterior distribution of the form $p(\boldsymbol{\theta}, \pi_n|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \pi_n)p(\boldsymbol{\theta}, \pi_n)$. In particular, we use MCMC methods to construct a Markov chain $\{(\boldsymbol{\theta}^{(1)}, \pi_n^{(1)}), (\boldsymbol{\theta}^{(2)}, \pi_n^{(2)}), (\boldsymbol{\theta}^{(3)}, \pi_n^{(3)}), \dots\}$ whose stationary distribution is $p(\boldsymbol{\theta}, \pi_n|\boldsymbol{y})$. For simplicty, we assume that $\boldsymbol{\theta}$ is a priori independent of π_n and propose a Gibbs sampler which iteratively samples from the full conditional distributions of π_n and $\boldsymbol{\theta}$.

1. Draw $\pi_n|\boldsymbol{y}, \boldsymbol{\theta}$ using a MH update with LSP proposals
2. Draw $\boldsymbol{\theta}|\boldsymbol{y}, \pi_n$ using a Gibbs update

Assuming independence between $\boldsymbol{\theta}$ and π_n allows us to focus our attention on Step 1 while remaining agnostic towards the implementation of Step 2. Here, we simply assume that a conjugate prior is available for $\boldsymbol{\theta}$ which permits the use of a Gibbs draw.

That being said, $\boldsymbol{\theta}$ and π_n may exhibit some structural dependence in many demand applications. In the demand model considered in our empirical application, for example, the dimension the cross-price elasticity vector depends on π_n. Other examples of dependence would arise when modeling the within-group correlation parameters in nested logit models (McFadden, 1978; Train, 2009) or allowing the screening rules of Gilbride and Allenby (2004) to vary across product groups. In such cases, one can always factor the prior and proposal distributions as $p(\boldsymbol{\theta}, \pi_n) = p(\boldsymbol{\theta}|\pi_n)p(\pi_n)$ and then jointly accept or reject $(\boldsymbol{\theta}, \pi_n)$. This is the strategy we take in Section 5 and outlined in Appendix D. However, we simply maintain the assumption of independence here for ease of exposition.

4.1 Single LSP Proposal

We first consider one MH update using LSP proposals for the entire partition π_n. Let $(\boldsymbol{\theta}^{(1)}, \pi_n^{(1)})$ denote a pair of arbitrary starting values and simulate $(\boldsymbol{\theta}^{(r)}, \pi_n^{(r)})$ for $r = 2, \ldots, R$ according to the following algorithm.

Algorithm 1: Single LSP Proposal

---

1. Generate $\pi_n^* \sim q(\pi_n|\pi_n^{(r)}, v) = \text{LSP}(\pi_n^{(r)}, v)$. Set $\pi_n^{(r+1)} = \pi_n^*$ with probability

$$\mathcal{A}(\pi_n^*, \pi_n^{(r)}) = \min\left\{1, \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(r)}, \pi_n^*)p(\pi_n^*)}{p(\boldsymbol{y}|\boldsymbol{\theta}^{(r)}, \pi_n^{(r)})p(\pi_n^{(r)})} \times \frac{q(\pi_n^{(r)}|\pi_n^*, v)}{q(\pi_n^*|\pi_n^{(r)}, v)}\right\}.$$

   Otherwise set $\pi_n^{(r+1)} = \pi_n^{(r)}$.

2. Draw $\boldsymbol{\theta}^{(r+1)}|\boldsymbol{y}, \pi_n^{(r+1)}$ using a Gibbs update.

The parameterization of the LSP distribution allows us to propose partitions in a random-walk fashion. That is, we can generate a candidate partition π_n^* from a distribution centered

at $\pi_n^{(r)}$ with a step size of v. The LSP distribution can be highly asymmetric, so the acceptance ratio must include the ratio of transition probabilities $q(\pi_n^{(r)}|\pi_n^*, v)/q(\pi_n^*|\pi_n^{(r)}, v)$, which can be easily calculated from the LSP probability mass function described in (15) and (16).

As with any random-walk MH algorithm, practical convergence and proper mixing of the Markov chain is highly dependent on the choice of the tuning parameter v. If v is too small, the algorithm will move in small increments and may fail to fully explore regions of high posterior probability. If v is too large, the algorithm will reject a high proportion of proposed moves and underestimate posterior uncertainty. In general, the choice of an optimal step size depends on the dimension and shape of the target posterior. In our case, efficient tuning with respect to both is challenging as existing optimal scaling results (e.g., Roberts and Rosenthal, 2001) measure mixing efficiency as a function of the integrated autocorrelation time, which is not well defined on non-Euclidean spaces like $\mathcal{P}_n$. These results also only pertain to stationary distributions that are products of independent normal densities, which is far from the discrete posteriors studied here. Mixing can still be monitored by the algorithm's acceptance rate, but care must be taken in computing the acceptance rate to account for the fact that $\Pr(\pi_n^* = \pi_n^{(r)}) > 0$ when $\pi_n^* \sim \text{LSP}(\pi_n^{(r)}, v)$.

Even without access to formal optimal scaling results, we can still provide some guidance on how to choose the scale parameter in order to account for differences in the dimension of the parameter space $\mathcal{P}_n$ across analyses. For example, if the step size $v = 1$ leads to good mixing for an analysis with $n = 10$ products, it may likely be too large of a step size for an analysis $n = 50$ products, regardless of the shape of the posterior. This problem also arises when choosing the prior scale parameter τ, as the probability mass placed on the location partition will change as a function of n.

Formally, consider two distinct LSP distributions: an $\text{LSP}(\rho_m, v)$ and an $\text{LSP}(\rho_n, v)$ where $v > 0$ is a common scale parameter and $\rho_m \in \mathcal{P}_m$, $\rho_n \in \mathcal{P}_n$, and $(m, n) \in \mathbb{N}$. If $m = n$ but $\rho_m \neq \rho_n$, then any differences between $\Pr(\pi_n = \rho_n|\rho_n, v)$ and $\Pr(\pi_m = \rho_m|\rho_m, v)$ can be attributed to differences in the group composition of ρ_m and ρ_n. However, if $m < n$

then regardless of the structure of ρ_n and ρ_m, $\Pr(\pi_n = \rho_n|\rho_n, v)$ will likely be smaller than $\Pr(\pi_m = \rho_m|\rho_m, v)$ simply because $\Pr(\pi_n = \rho_n|\rho_n, v)$ is normalized over a higher-dimensional domain which scales the LSP probabilities downward. While the same argument could be applied to any well-defined neighborhood of ρ_n, we choose to focus on the probability assigned to ρ_n alone in order to simplify the optimization below.

To account for this dimensional scaling of the LSP distribution, we propose using scale parameters of the form $v = s \cdot f(n)$ where $s > 0$ and $f(n)$ is a decreasing function of n. The idea is that as n increases, any downward pressure on the LSP probabilities can be countered with a smaller scale parameter. We then choose v so that, on average, the probability assigned to the location partition is constant across dimensions. Doing so allows us to roughly hold fixed the amount of information imposed by the prior for different n as well as control the step sizes of LSP proposals across dimensions. More formally, we want to find the scale parameter v which solves the following for any $(m, n) \in \mathbb{N}$.

$$\min_{v>0} \left| \frac{1}{|\mathcal{P}_m|} \sum_{\rho_m \in \mathcal{P}_m} \Pr(\pi_m = \rho_m|\rho_m, v) - \frac{1}{|\mathcal{P}_n|} \sum_{\rho_n \in \mathcal{P}_n} \Pr(\pi_n = \rho_n|\rho_n, v) \right| \tag{30}$$

Formal optimization of (30) would require enumerating the entire space of partitions $\mathcal{P}_n$ for various n, which becomes unwieldly for even a moderate number of items. Instead, we generate Monte Carlo estimates of the high dimensional sums by sampling n^2 partitions from $\mathcal{P}_n$ for each $n \in \{25, 50, \ldots, 200\}$ and then averaging $\Pr(\pi_n = \rho_n|\rho_n, v)$ across draws within $\mathcal{P}_n$. Figure 4 plots these averaged LSP probabilities with $s = 1$ and five types of scaling functions: logarithmic $f(n) = 1/\log(n)$, linear $f(n) = 1/n$, linearithmic $f(n) = 1/(n\log(n))$, polynomial $f(n) = 1/n^a$, and exponential $f(n) = 1/a^n$. We find that the linearithmic function $v = 1/(n\log(n))$ scales the LSP distribution best, as it minimizes the change in probabilities across dimensions.

To reiterate, we are not claiming that the choice of $v = 1/(n\log(n))$ will guarantee optimal posterior mixing. Rather it only serves as a guide for how to control the neighborhood of LSP proposals across analyses with different n. Given a specific data set, more experimentation

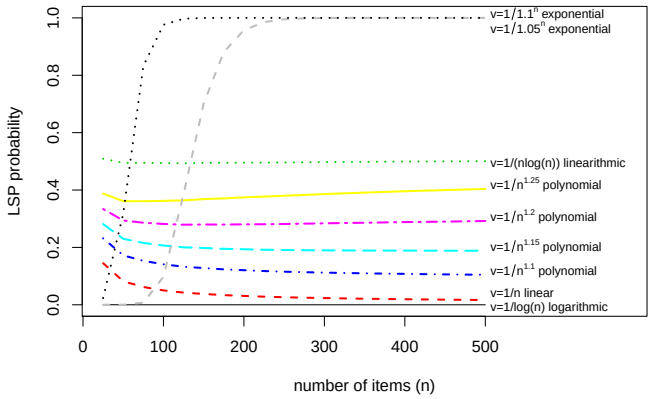


Figure 4: Monte Carlo averages of $\Pr(\pi_n = \rho_n | \rho_n, v)$ are plotted across dimensions n for different scaling functions $v = f(n)$.

can be done in choosing s, for example, to ensure proper mixing.

4.2 Sequential Block LSP Proposals

The algorithm described above is subject to the same scaling limitations as any random-walk MH algorithm. That is, one may be concerned that for large n, the algorithm will get stuck in local modes of the posterior and yield low acceptance rates. Given the challenges of deriving optimal scaling results for Markov chains on non-Euclidean discrete spaces, we propose a second algorithm that uses a blocking strategy to improve mixing. The idea of blocking is to split the parameter vector (in our case the partition) into mutually exclusive blocks and then propose updates one block at a time.

The theoretical advantages of block sampling are well documented, both for Gibbs samplers (Liu et al., 1994; Roberts and Sahu, 1997) and more general MH-based algorithms (Sargent et al., 2000; Turek et al., 2017). These advantages have also been illustrated empirically. For example, Chib and Ramamurthy (2010) use block sampling to address the high dimensionality and multi-modality in dynamic stochastic general equilibrium (DSGE) models for macroeconomic data. Additionally, Musalem et al. (2009) and Chen and Yang (2007) discuss advantages of block sampling high-dimensional vectors of latent individual

choices when estimating individual-level consumer preferences from store-level sales data.

In our context, the idea is to divide π_n into L contiguous blocks $\pi_n = (B_1, \ldots, B_L)$ and use a sequence of MH updates with LSP proposals for each $B_\ell | B_{-\ell}$. For example, we could divide a partition of $n = 20$ items into $L = 4$ equally-sized blocks.

$$\pi_n = (\underbrace{g_1, g_2, g_3, g_4, g_5}_{B_1}, \underbrace{g_6, g_7, g_8, g_9, g_{10}}_{B_2}, \underbrace{g_{11}, g_{12}, g_{13}, g_{14}, g_{15}}_{B_3}, \underbrace{g_{16}, g_{17}, g_{18}, g_{19}, g_{20}}_{B_4})$$

In this case, each of the $\ell = 1, \ldots, 4$ MH steps would propose a new partition $\pi_n^* = (B_1, \ldots, B_\ell^*, \ldots, B_L)$ such that only the elements in block ℓ are allowed to change.

Formally, let $\text{LSP}_{\{\underline{b}_\ell, \bar{b}_\ell\}}(\rho_n, \tau)$ denote a block LSP distribution that is defined for all items $i \in B_\ell = \{\underline{b}_\ell, \ldots, \bar{b}_\ell\}$. Partitions can be sampled from the $\text{LSP}_{\{\underline{b}_\ell, \bar{b}_\ell\}}(\rho_n, \tau)$ distribution as follows.

1. Set $g_i = s_i$ for all $i < \underline{b}_\ell$.
2. Sample g_i according to (7) for $i = \underline{b}_\ell, \ldots, \bar{b}_\ell$.
3. Set $g_i = s_i$ for all $i > \bar{b}_\ell$ and, if necessary, relabel g_i so that: (i) the sampled partition conforms to the order restriction of $\mathcal{P}_n$; and (ii) if $s_i > s_j$ for any $j \leq \underline{b}_\ell$, then $g_i > g_j$.

The final step ensures that a partition sampled from $\text{LSP}_{\{\underline{b}_\ell, \bar{b}_\ell\}}(\rho_n, \tau)$ is a valid partition in $\mathcal{P}_n$ and preserves the grouping for all items not in B_ℓ. Examples of relabeling according to these two conditions are provided in Figure 5.

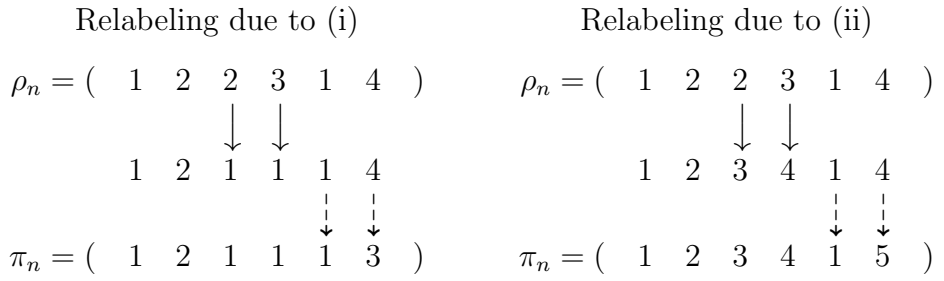


Figure 5: Examples of relabeling required when block sampling partitions. Solid lines indicate probabilistic sampling and dashed lines indicate deterministic relabeling.

An MCMC routine that uses sequential block LSP proposals to draw π_n is outlined below. The sampler integrates over the block configuration by randomly generating a new number

of blocks L and associated cutpoints $\{(\underline{b}_1, \overline{b}_1), \ldots, (\underline{b}_L, \overline{b}_L)\}$ in each iteration.

Algorithm 2: Block LSP Proposals

---

1. (a) Randomly generate the number of blocks L and cutpoints $\{(\underline{b}_1, \overline{b}_1), \ldots, (\underline{b}_L, \overline{b}_L)\}$.

   (b) For each block $\ell = 1, \ldots, L$ generate $\pi_n^* \sim q(\pi_n | \pi_n^{(r)}, v, \underline{b}_\ell, \overline{b}_\ell) = \text{LSP}_{\{\underline{b}_\ell, \overline{b}_\ell\}}(\pi_n^{(r)}, v)$. Set $\pi_n^{(r+1)} = \pi_n^*$ with probability

$$\mathcal{A}(\pi_n^*, \pi_n^{(r)}) = \min \left\{ 1, \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^{(r)}, \pi_n^*) p(\pi_n^*)}{p(\boldsymbol{y}|\boldsymbol{\theta}^{(r)}, \pi_n^{(r)}) p(\pi_n^{(r)})} \times \frac{q(\pi_n^{(r)} | \pi_n^*, v, \underline{b}_\ell, \overline{b}_\ell)}{q(\pi_n^* | \pi_n^{(r)}, v, \underline{b}_\ell, \overline{b}_\ell)} \right\}.$$

   Otherwise set $\pi_n^{(r+1)} = \pi_n^{(r)}$.

2. Draw $\boldsymbol{\theta}^{(r+1)} | \boldsymbol{y}, \pi_n^{(r+1)}$ using a Gibbs update.

The block configuration moves in step 1(a) will leave the stationarity distribution $p(\boldsymbol{\theta}, \pi_n | \boldsymbol{y})$ invariant as long as the proposal mechanism is independent of the output from previous iterations. We generate block configurations as follows: first set $\underline{b}_1 = 1$ and draw $\overline{b}_1 \sim \text{Unif}\{\underline{b}_1 + 1, n\}$; if $\overline{b}_1 < n$, then set $\underline{b}_\ell = \overline{b}_{\ell-1} + 1$ and draw $\overline{b}_\ell \sim \text{Unif}\{\underline{b}_\ell + 1, n\}$ for each $\ell > 1$ until $\overline{b}_\ell = n$. The last piece is the transition probabilities in step 1(b), which take the form

$$q(\pi_n | \rho_n, \tau, \underline{b}_\ell, \overline{b}_\ell) = \prod_{i \in B_\ell} p(g_i | g_{<i}, \rho_n, \tau, \underline{b}_\ell, \overline{b}_\ell) \tag{31}$$

where each $p(g_i | g_{<i}, \rho_n, \tau, \underline{b}_\ell, \overline{b}_\ell)$ comes from the LSP probability mass function in (16).

One potential drawback of block sampling is that it requires L (instead of 1) likelihood evaluations per iteration. The gains in efficiency from block LSP proposals are therefore bound by the time it takes to evaluate the likelihood, prior, and proposal densities. In cases with lots of data or where evaluating the likelihood requires the inversion of large matrices or integration over irregular regions, block LSP proposals will yield much longer run times than single LSP proposals. However, alternative block configuration sampling schemes can always be used to alleviate the additional computational burden. For example, minimum block sizes can be imposed or distributions other than the uniform can be used when sampling

the cutpoints $(\underline{b}_\ell, \overline{b}_\ell)$ in order to keep the number of blocks L small. We propose one such modification in the following simulation study.

4.3 Simulation Study with Alternative Proposals

We compare single LSP and block LSP proposals to two alternative proposal mechanisms. The first is standard Gibbs sampling where each item-group indicator variable g_i (rather than the entire partition π_n) is drawn from its full conditional distribution. The second is the split-merge algorithm of Jain and Neal (2004). Both of these proposals are characterized by incremental moves. That is, when generating a candidate partition $\pi_n^* = (g_1^*, \ldots, g_n^*)$, at most one element (or group of elements) of π_n^* can differ from $\pi_n^{(r)}$. This can pose significant mixing problems for models that place high posterior probability on partitions that are separated by valleys of low posterior probability, as measured by the number of incremental moves it would take to move from one to the other. To this extent, we expect LSP proposals to offer an advantage over existing methods in their ability to navigate complicated posterior distributions. This is because the step size v in LSP proposals permits the generation of candidate partitions that can be radically different from the partition in the current state.

For the purpose of illustration, we compare the effectiveness of different proposal distributions using data generated according to the following regression model.

$$y_t = h(\boldsymbol{x}_t, \boldsymbol{\beta}, \pi_n) + \varepsilon_t = \sum_{k=1}^{K} \Big(\sum_{j \in G_k} x_{jt}\beta_j \Big)^2 + \varepsilon_t, \;\; \varepsilon_t \sim \mathrm{N}(0, \sigma^2) \tag{32}$$

Here the n-dimensional covariate vector $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{nt})$ is partitioned into $K \leq n$ groups, which introduces nonlinearities into the conditional mean function $h(\boldsymbol{x}_t, \boldsymbol{\beta}, \pi_n)$. We assume a uniform prior for π_n so that the conditional posterior reduces to

$$p(\pi_n | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \sigma^2) \propto \prod_{t=1}^{T} \mathrm{N}\left(h(\boldsymbol{x}_t, \boldsymbol{\beta}, \pi_n), \sigma^2\right). \tag{33}$$

We simulate $D = 25$ data sets from the model in (32), each with $n = 6$ covariates and $T = 20$ observations. For each data set, we fix $\beta_1 = \cdots = \beta_n = 1$ and the error variance

$\sigma^2 = 1$, and generate the true partition from an LSP distribution with $\rho_n = (1, 1, 1, 1, 1, 1)$ and $\tau = n$. We generate the covariates x_{it} from either a Unif(-1,1) or a Unif(0,2) distribution, with each distribution giving rise to likelihood surfaces of different complexities.

Each Markov chain is run for $R = 2000$ iterations and we discard the first 50% of draws as burn-in. The step size for both types of LSP proposals is taken to be $v = 1/(n \log(n))$. To address concerns of dependence on parameter starting values, the initial value of the partition for each chain is randomly drawn from an LSP distribution with ρ_n equal to the partition with n groups and $\tau = n$. For the Gibbs sampling routine, we also fix the number of groups to be the true number of groups for each data set. The performance of the Gibbs sampler will thus be overstated, as the number of groups is almost always unknown.

The top panel of Figure 6 plots a sample log likelihood for each distribution of the covariates. The x-axis is sorted according to the adjusted Rand index where each $\pi_n \in \mathcal{P}_n$ is compared to the partition with one group – i.e., the partition with one group is farthest on the left and the partition with n groups is farthest on the right. The vertical black line marks the position of the partition used to generate the data. The bottom panel of Figure 6 plots the similarity between the true partition and $\pi_n^{(r)}$ for each post-burn-in draw using the adjusted Rand index.[2]

We find that the performance of the samplers is highly dependent on the complexity of the likelihood surface. When the covariates are in the range (-1,1), the likelihood is relatively flat over $\mathcal{P}_n$ and the samplers do equally well in exploring the conditional posterior. However, the likelihood exhibits significant peaks and valleys when the covariates are positive. While this poses a problem for samplers that use (group-wise) incremental moves, LSP proposals remain effective in navigating the complicated likelihood surface to find regions of high probability. An alternative explanation for low values of the adjusted Rand index is that the

[2]The Rand index measures the similarity between two partitions on the basis of how often items get assigned into the same/different groups. The *adjusted* Rand index is just the Rand index accounting for chance of co-clustering under some null model. Specifically, it measures the normalized difference of the Rand index and its expected value under the null hypothesis that the induced contingency table (from comparing the two partitions) is generated from a hypergeometric distribution (Hubert and Arabie, 1985).

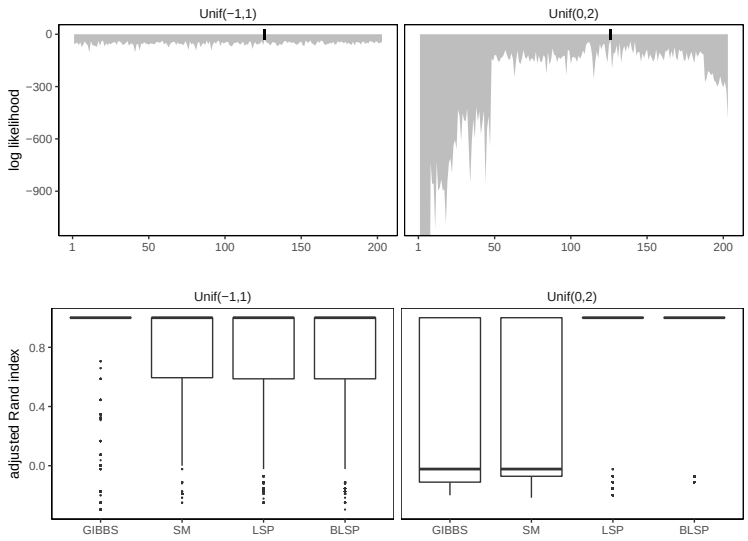


Figure 6: The top panel plots a sample log likelihood for each set of covariates. The vertical line indicates the position of the partition used to generate the data. The bottom panel plots the Rand-similarity between the true partition and each post-burn-in draw $\pi_n^{(r)}$.

Markov chain has yet to reach its stationary distribution. Formally testing for convergence here is challenging because the high-dimensional discrete parameter space precludes the use of traditional convergence diagnostic tools. However, Figure 6 still indicates the relative speed at which the different samplers navigate to regions of high posterior probability. It has also been our experience that the patterns shown in Figure 6 change very little as the number of iterations grows.

We repeat the simulation study described above in a higher-dimensional setting where $n \in \{25, 50, 75, 100\}$. For each value of n, we generate $D = 25$ data sets where the amount of information is either high ($T = 10n$) or low ($T = 5n$). The step size is set to $v = 1/(n \log(n))$ and we let $R = 500n$ so that the run length increases with the dimension of $\mathcal{P}_n$. For block LSP proposals, we add an additional constraint to the block configuration sampling scheme. In each iteration of the chain, the block structure will have one group (reducing to a single LSP proposal) with probability θ or will be drawn from the scheme proposed in Section 4.2 with probability $1-\theta$, where we fix $\theta = 0.5$. This strategy takes advantage of gains in mixing associated with block sampling while minimizing the costs of longer run times.

Each sampler is then evaluated in two ways: (1) how well it recovers the true partition, which we measure by computing the adjusted Rand index between the posterior draws associated with each data set $\pi_{n,d}^{(r)}$ and the true partition $\pi_{n,d}^{\text{true}}$; (2) how well it recovers the true number of groups, as measured by the difference between the posterior draws $K_d^{(r)}$ and the true value $K_d^{\text{true}} = \max(\pi_{n,d}^{\text{true}})$. The results are shown in Table 1.

Table 1: Results from a high-dimensional simulation study. The LSP and block LSP samplers are evaluated based on how well they can recover the true partition (measured by computing the average adjusted Rand index (ARI) between posterior draws of π_n and the true partition) as well as the true number of groups (measured by computing the average difference between posterior draws of K and the true number of groups).

		Partitions: $ARI(\pi_{n,d}^{(r)}, \pi_{n,d}^{\text{true}})$				Groups: $K_d^{(r)} - K_d^{\text{true}}$			
		LSP		block LSP		LSP		block LSP	
Observations	Dimension	Mean	SD	Mean	SD	Mean	SD	Mean	SD
High	$n = 25$	0.98	0.05	1.00	0.01	-0.02	0.18	0.00	0.00
	$n = 50$	0.98	0.04	0.99	0.02	0.36	0.76	0.33	0.77
	$n = 75$	0.95	0.04	0.98	0.03	0.04	0.21	0.04	0.20
	$n = 100$	0.90	0.10	0.96	0.05	2.82	2.27	2.22	2.10
Low	$n = 25$	0.82	0.30	0.88	0.27	0.33	0.64	0.17	0.49
	$n = 50$	0.53	0.33	0.65	0.30	5.96	1.45	5.09	1.98
	$n = 75$	0.33	0.19	0.41	0.24	1.90	2.46	0.72	1.37
	$n = 100$	0.22	0.15	0.25	0.13	6.63	2.35	6.93	2.03

We find that LSP proposals do well in navigating high-dimensional posteriors with a sufficient amount of data. However, when the amount of data is small relative to n, the posterior of π_n tends to concentrate in areas farther away from the true partition and overestimate the number of groups K. Although these results are conditional on the likelihood defined by (32) and the uniform prior on π_n, our approach is likely best suited for empirical settings where T is large relative to n. In even higher-dimensional settings where $T < n$ and/or n is in the hundreds or thousands, further item-level restrictions would be necessary to ensure good mixing and convergence.

5 An Application to Store-Level Category Demand

Category managers face the task of setting price and promotion schedules, arranging end-aisle displays, allocating shelf-space among national and store brands, and forecasting category demand. These activities all depend on the estimation of a demand system that relates prices to quantities purchased among the set of goods within the category. The challenge is that categories are often broadly defined (e.g., juice, salty snacks, cereal), spanning a large set of products and product attributes. This gives rise to product subcategories, which are usually more homogeneous in at least one attribute dimension like flavor (e.g., pretzels).

When optimizing marketing actions for a category, managers must consider if and how demand is related across subcategories. For instance, if price changes of potato chips tend to have appreciable effects on the demand for other salty snacks, then optimal marketing actions must be solved for as a function of all products in the entire salty snacks category. However, if demand is completely isolated by subcategory (e.g., demand for potato chips is unaffected by price changes to pretzels and vice-versa), then the manager could simplify the problem by separately solving for optimal policies within each subcategory. The extent to which subcategories are related or isolated is ultimately an empirical question, and one we will address using the partitioning methodology outlined above.

We start with a flexible category demand model which regresses the log of total purchase volume for products $i = 1, \ldots, n$ on the set of log prices for all related goods as well as other product-specific covariates z_{jt} for $j = 1, \ldots, p$ using $t = 1, \ldots, T$ weeks of data.

$$\log y_{it} = \sum_{j=1}^{n} \beta_{ij} \log p_{jt} + \sum_{j=1}^{p} \psi_{ij} z_{jt} + \varepsilon_{it} \tag{34}$$

This results in a system of n demand equations that are related through the joint error vector $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{nt}) \sim \mathrm{N}(0, \Sigma)$. Although the errors could also exhibit some dependence over time, the typical approach in the demand modeling literature is to maintain the assumption of independence while accounting for various time trends in the set of controls z_{jt} (e.g., Montgomery, 1997; Wedel and Zhang, 2004). In our application, we include year dummies

in addition to product intercepts as controls. The log-linear demand specification in (34) is popular in practice for three reasons: (1) the model parameters β_{ij} represent price elasticities which measure the percent change in demand for product i given a one percent increase in price for product j; (2) the system is flexible, as it can admit substitution patterns consistent with both substitutable ($\beta_{ij} > 0$) and complementary ($\beta_{ij} < 0$) goods; (3) the model is linear in β_{ij} so estimation of model parameters is straightforward.

Model flexibility becomes especially important as the size and scope of the product category grows. For example, while all potato chips may be substitutes, it may be the case that some potato chips and pretzels are complements. If the goal is to forecast demand for the entire salty snack category, a model assuming unit demand among substitutable goods (e.g., logit-based demand systems) would be inappropriate. However, flexibility comes at a cost of estimating the n^2 parameters making up the full $n \times n$ price elasticity matrix $\boldsymbol{B} = \{\beta_{ij}\}$. Since many pricing/promotion schedules are customized to the store level, it is also important to be able to generate precise parameter estimates and demand forecasts even for a relatively short panel of observations. The demands on the data can therefore be significant when the number of products in a given category is large.

One way to improve the precision of elasticity estimates is to impose prior restrictions on the elasticity parameters (Blattberg and George, 1991; Montgomery and Rossi, 1999). Here we propose a "grouped isolation" restriction in which demand is unaffected by changes in price of products in different groups. The partition $\pi_n = (g_1, \ldots, g_n)$ then imposes equality restrictions on cross-price elasticities such that $\beta_{ij} = 0$ if $g_i \neq g_j$ and is left unrestricted otherwise. Conditional on any single partition, the model imposes very strict restrictions on cross-price effects. Therefore, the benefits of dimension reduction may be limited if the shrinkage patterns induced by the partition are misspecified. By letting the partition be a model parameter, we can both learn about the structure of demand from the data as well as average over any uncertainty in π_n when forecasting demand.[3]

[3]It could also be argued that a partition with one group, in which everything relates to everything else, is always the "true structure." However, we focus on situations where it is desirable to impose some

An alternative approach to reducing the dimension of the cross-price elasticity matrix is to rely on other shrinkage and regularization methods. Classical examples include ridge regression (Hoerl and Kennard, 1970), the lasso (Tibshirani, 1996), and the elastic net (Zou and Hastie, 2005). Bayesian regularization arises through the specification of various sparsity-inducing priors on the regression coefficients. Examples include spike-and-slab priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1997), Student-t priors (Tipping, 2001), Laplacian priors (Park and Casella, 2008; Hans, 2009), orthant normal priors (Hans, 2011), Horseshoe priors (Carvalho et al., 2010), and spike-and-slab lasso priors (Ročková and George, 2018).

While each of these methods produces different sparsity patterns, a common theme is that they all assume independence between regression coefficients a priori. When applied to demand models in the form of (34), this implies that the shrinkage imposed on one cross-price elasticity β_{ij} is entirely independent of the shrinkage imposed on any other elasticity a priori. In contrast, our approach shrinks cross-price elasticities to zero at the *group level*, which effectively sets blocks of cross-price elasticities to zero. Our approach is therefore closest to other group-level regularization methods, such as the group lasso (Yuan and Lin, 2006) and sparse-group lasso (Simon et al., 2013). These methods have been shown to work well when the covariates exhibit a natural grouping structure. One key difference is that we also treat the partitioning of items into groups as a model parameter.

The regularization literature has also focused on fixed shrinkage points set to zero. This may not be ideal in a demand modeling context because the Slutsky equation suggests that price effects are comprised of both substitution effects and income effects (Deaton and Muellbauer, 1980). Thus, even when substitution effects are zero, non-zero income effects imply non-zero price effects. Although we have assumed zero shrinkage points to simplify the analysis, our restrictions on β_{ij} can be extended to include (non-zero) group-level parameters that provide differential shrinkage across product groups.

restrictions on model parameters either because of data limitations or to improve the precision and reliability of estimates.

5.1 Data Description

We apply the demand model described above to store-level data from the IRI Marketing data set (Bronnenberg et al., 2008). We use five years of weekly price and purchase volume data from one store in Eau Claire, Wisconsin. Four years of data ($T = 208$) are used for estimation and one year ($T = 52$) is used for prediction. We focus on the salty snacks category and include UPCs from all subcategories: potato chips (PTOCHP), pretzels (PRETZL), tortilla chips (TTACHP), corn snacks (CRNSNK), cheese snacks (CHESNK), popcorn (POPCRN), and other salted snacks (OTHER). We aggregate UPCs to the brand level within each subcategory due to the high collinearity of prices within a brand's product line (e.g., Classic Lay's vs. Barbecue Lay's). Weekly volume for each brand is taken to be the sum of UPC-level volumes, while weekly prices are volume-weighted averages of UPC-level prices. Brands with low within-subcategory market share are also discarded due to the abundance of missing data. A description of the resulting $n = 40$ products (which correspond to 454 unique UPCs) is provided in Table C.1.

5.2 Prior Specification

A fully unrestricted log-linear model and a log-linear model subject to isolation restrictions are fit to the data. Conjugate but diffuse priors are placed on parameters in the unrestricted multivariate regression model: $\psi_{ij} \sim \mathrm{N}(0, 100)$, $\mathrm{vec}(\boldsymbol{B})|\Sigma \sim \mathrm{N}(\bar{\boldsymbol{\beta}}, \Sigma \otimes A^{-1})$, $\Sigma \sim \mathrm{IW}(\nu, V)$, $\bar{\boldsymbol{\beta}} = 0$, $A^{-1} = 10I$, $\nu = n + 3$, and $V = \nu I$. Note that the prior for all elasticity parameters is centered at 0 with a variance of 10, which implies that the prior places roughly 95% of its mass on the range (-6,6). We believe this is reasonable as own-price elasticities do not usually fall below -6% and cross-price elasticities do not usually exceed 6%. Prior information about the sign of the elasticities can also be imposed by specifying elements of $\bar{\boldsymbol{\beta}}$ to be non-zero. Doing so can help improve the practical validity of the model by ensuring that the sign of the estimated elasticities conforms with economic theory (e.g., substitutes exhibiting positive cross-price elasticities). However, given the wide assortment of goods we study, not all goods

may be strict substitutes so we maintain a relatively uninformative prior with $\bar{\boldsymbol{\beta}} = 0$.

In the restricted model, the joint prior for the elasticities and partition is specified as

$$\boldsymbol{\beta}_{\pi_n}|\pi_n \sim \mathrm{N}(\bar{\boldsymbol{\beta}}_{\pi_n}, a^{-2}I_{\pi_n}) \tag{35}$$

$$\pi_n \sim \mathrm{LSP}(\rho_n, \tau) \tag{36}$$

where I_{π_n} denotes a diagonal matrix of dimension equal to the number of unrestricted elasticities conditional on π_n. For example, if there are $n = 4$ goods and π_n contains $K = 2$ groups each having two products, then the dimension of $\boldsymbol{\beta}_{\pi_n}$ will be $2^2 + 2^2 = 8$. In general, the dimension of $\boldsymbol{\beta}_{\pi_n}$ increases as the number of groups K decreases since fewer groups implies more within-group cross-elasticities. The dimension of this conditional posterior is important because any MH algorithm that jointly updates $(\boldsymbol{\beta}_{\pi_n}, \pi_n)$ will evaluate an acceptance probability that includes the ratio of prior densities evaluated at the proposed and current values: $p(\boldsymbol{\beta}^*_{\pi^*_n}|\pi^*_n)p(\pi^*_n)/\big(p(\boldsymbol{\beta}_{\pi_n}|\pi_n)p(\pi_n)\big)$. However, if $\dim(\boldsymbol{\beta}^*_{\pi^*_n}) > \dim(\boldsymbol{\beta}_{\pi_n})$ and these conditional priors are specified to be diffuse, then the ratio of prior densities will favor $\boldsymbol{\beta}_{\pi_n}$ simply because it contains fewer elements than $\boldsymbol{\beta}^*_{\pi^*_n}$. In other words, diffuse priors will give more weight to models with fewer dimensions, effectively penalizing models with fewer groups and many unrestricted elasticities. This effect can be countered by specifying a more informative conditional prior $p(\boldsymbol{\beta}_{\pi_n}|\pi_n)$ or by choosing the LSP hyperparameters to favor models with fewer groups (e.g., letting ρ_n have one group and choosing a small value of τ).

In our restricted models, we specify the same prior for ψ_{ij} and Σ as the unrestricted models. We then let $\bar{\boldsymbol{\beta}}_{\pi_n}|\pi_n \sim \mathrm{N}(0, 10I_{\pi_n})$ and consider a variety of prior specifications for π_n. The first is the prior distribution on partitions that is induced by the DP with concentration parameter $\alpha = 1$. This prior will be relatively flat over the space of pairwise assignment probabilities, but will tend to favor partitions with a few large groups and many small groups. Then we consider different LSP priors where ρ_n is either a partition with one group (LSP-one), a partition based on predefined subcategories (LSP-category), or a partition based on brands (LSP-brand). In each case, we make the prior informative by

setting $\tau = 0.1/(n\log(n))$. Finally, we specify a covariate-dependent LSP prior with two covariates: end-aisle display advertising frequency and in-store circulator feature advertising frequency. We also let the location partition for this prior have one group. Moreover, the covariates are rescaled to have mean zero and unit variance, and the scaling parameter for both covariates is set to $\lambda = 0.1/(n\log(n))$.

5.3 Computation

To facilitate posterior sampling, we first rewrite the multivariate normal likelihood as a seemingly unrelated regression (SUR) model.

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_{1,\pi_n} & & \\ & \ddots & \\ & & \boldsymbol{X}_{n,\pi_n} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_n \end{pmatrix} + \begin{pmatrix} \boldsymbol{Z}_1 & & \\ & \ddots & \\ & & \boldsymbol{Z}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\psi}_1 \\ \vdots \\ \boldsymbol{\psi}_n \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix} \tag{37}$$

The design matrix for product i, denoted $\boldsymbol{X}_{i,\pi_n}$, now contains the columns of $\boldsymbol{X}$ for all other products j such that $g_i = g_j$. The benefit of the SUR likelihood representation is that we now have a model that is linear in the vector of unrestricted elasticities $\boldsymbol{\beta}_{\pi_n}$. Assuming a normal prior on $\boldsymbol{\psi}$ and a conditionally normal prior on $\boldsymbol{\beta}_{\pi_n}|\pi_n$ gives rise to closed form expressions for the associated full conditional distributions.

MCMC methods are used to sample from the posterior of each model. For the unrestricted model, we use a Gibbs sampler that draws $\boldsymbol{\beta}|\boldsymbol{\psi}, \Sigma, \boldsymbol{y}$ using the normal posterior for conjugate multivariate regression models and then draws $\boldsymbol{\psi}, \Sigma|\boldsymbol{\beta}, \boldsymbol{y}$ using the normal posterior for conjugate SUR models (Rossi et al., 2005). For the restricted models, the same Gibbs step is used to draw $\boldsymbol{\psi}, \Sigma|\boldsymbol{\beta}, \boldsymbol{y}$. However, to sample from the posterior of $\boldsymbol{\beta}_{\pi_n}, \pi_n|\boldsymbol{\psi}, \Sigma, \boldsymbol{y}$, we use a joint MH proposal described in Appendix D. Each chain is run for $R = 500,000$ iterations and then thinned by keeping every 100th draw to reduce autocorrelation. The first 50% of draws are discarded as burn-in.

5.4 Results

Table 2 reports in-sample and predictive fit statistics. In particular, we report the posterior mean and standard deviation of the root mean-squared error (RMSE) statistic. We find the in-sample fit to be similar across models, suggesting the restricted models can retain flexibility when the partition is estimated. In terms of predictive fit, however, we find the restricted models outperform the unrestricted model by roughly 5-6 percentage points. Among the restricted models, the LSP-one prior performs best, although the differences are relatively small.[4]

Table 2: Model Fit Statistics

	In-Sample RMSE		Predictive RMSE	
Model	Mean	SD	Mean	SD
1. Unrestricted	0.542	0.004	0.668	0.011
2. Restricted w/ DP prior	0.538	0.002	0.611	0.005
3. Restricted w/ LSP-one prior	0.535	0.002	0.607	0.006
4. Restricted w/ LSP-category prior	0.542	0.002	0.614	0.005
5. Restricted w/ LSP-brand prior	0.542	0.002	0.613	0.004
6. Restricted w/ LSPx prior	0.535	0.002	0.612	0.005

One benefit of imposing isolation restrictions is a reduction in the number of estimated parameters. For example, while there are $40^2 = 1600$ price elasticities to be estimated in the unrestricted model, there are only 135 elasticities on average (a posteriori) in the restricted model with the LSP-one prior. We find this dimension reduction improves the precision of demand forecasts, as the standard deviations of the predictive RMSEs are roughly cut in half relative to the unrestricted model. Dimension reduction also provides gains in efficiency of the elasticity estimates themselves. Figure 7 compares the estimated own-price elasticities from the unrestricted model and the restricted model with the LSP-one prior. While we find a high degree of correlation between the posterior means (≈ 0.97), the restricted model leads to a reduction in the posterior standard deviation for all products.

[4]We also fit LSP-one priors where τ is scaled down and up by a factor of 1000. The fit of the more informative prior is similar to the original LSP-one prior, while the fit of the diffuse prior is similar to that of the DP prior.

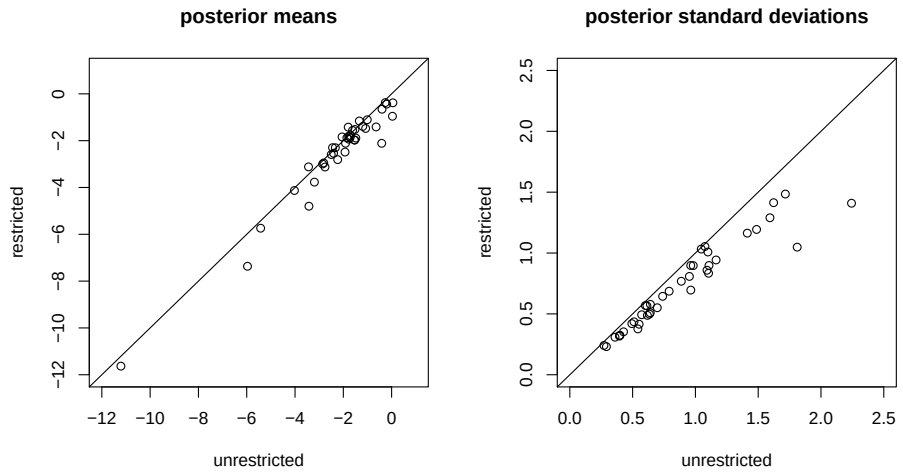


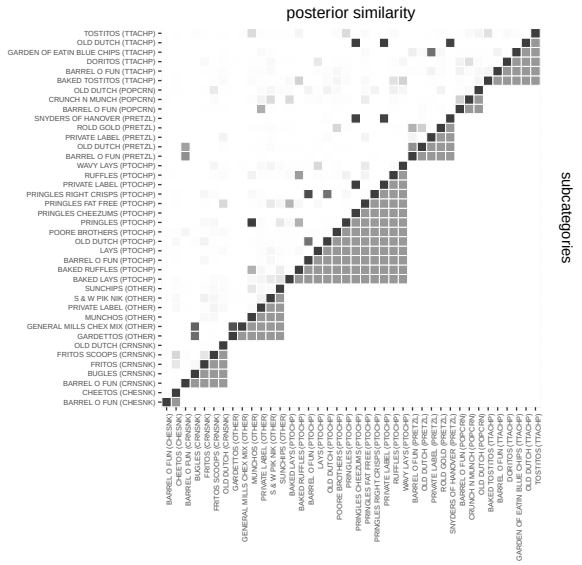
Figure 7: Posterior means and standard deviations of the set of $n = 40$ own-price elasticities are plotted for the unrestricted model and the restricted model with an LSP prior centered around the partition with one group.

Precision is important here because own-price elasticity parameters are a key input into retailer optimal pricing problems (Montgomery, 1997; DellaVigna and Gentzgow, 2017). Specifically, when a monopolist retailer sets prices to maximize total category profits, log-linear demand models give rise to optimal markups of the form $\beta_{ii}/(1 + \beta_{ii})$. Given the curvature of this function near zero, large posterior standard deviations in the own elasticities can lead to a long upper tail in the distribution of optimal prices. More precise estimates of β_{ii} can thus ensure more stable and reliable pricing.

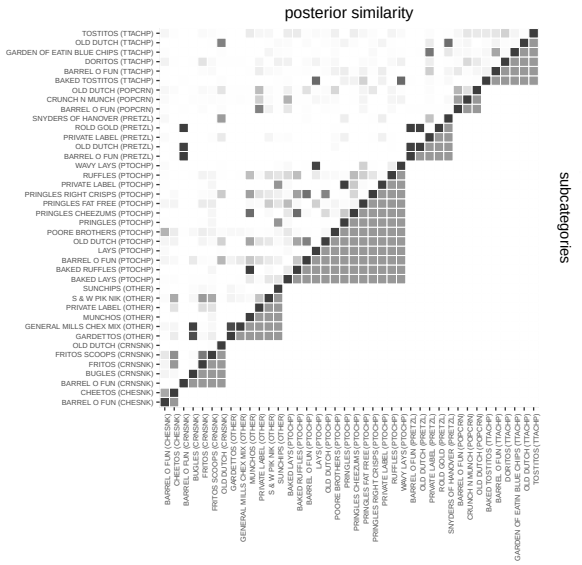
Next, we examine the posterior distribution of partitions under the various restricted models. Figure 8 plots the posterior pairwise similarity matrices for restricted models with a DP prior, LSP-one prior, LSP-category prior, and LSP-brand prior. In each plot, the upper left corner shows how often two products are grouped together in the posterior and the bottom right corner shows the composition of product subcategories. This is done just to help visualize the extent to which demand is or is not isolated across subcategories.

In general, we find evidence that demand is not perfectly isolated across subcategories. The greatest cross-subcategory clustering is induced by the DP and LSP-one priors, which both place appreciable mass on all pairwise clustering probabilities a priori. For example,

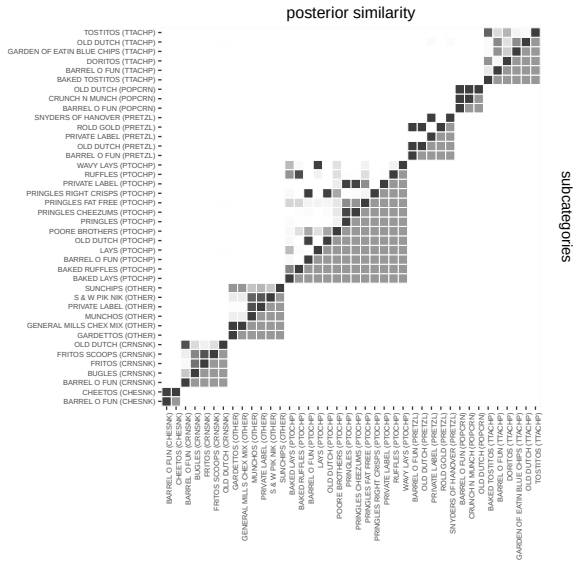
(a) DP prior



(b) LSP prior



(c) LSP-category prior



(d) LSP-brand prior

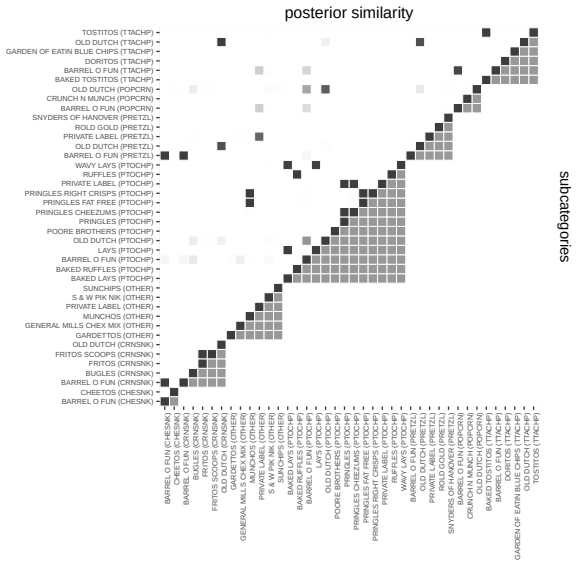


Figure 8: Posterior pairwise similarity matrices are plotted for the restricted demand models with a DP prior, LSP prior, LSP-category prior, and LSP-brand prior. The upper left corner shows posterior similarity and the bottom right corner shows the composition of product subcategories defined in the data.

the LSP-one prior leads to a posterior similarity matrix that places positive probability on 64% of all off-diagonal elements. This is in stark contrast to the LSP-category and LSP-brand priors, which only place positive posterior probability on 17% and 11% of off-diagonal elements, respectively.

The differences in the posterior similarity matrices in Figure 8 also illustrate the role of the prior partitioning model on posterior inference. In our data set, the number of observations is still relatively small so the prior will more strongly inform posterior clustering. Retail managers applying this methodology to shallow data sets will then want to make sure that their partitioning prior reflects both current knowledge of market structure as well as the strength of that belief. One benefit of the LSP prior is that its location-scale parameterization directly facilitates this prior elicitation: market structure can be represented through ρ_n and the strength of belief is reflected by τ. If one wants to guard against too much prior influence, then priors can be placed on ρ_n and τ.

Figure 9 plots the induced posterior distribution over the number of groups K for the same four restricted models. We find the ordering to be consistent with our expectation. For example, there are $K = 21$ groups (unique brands) in the LSP-brand prior, which is close where the corresponding posterior places most of its mass. In contrast, the LSP-one prior concentrates its posterior mass around partitions with $K = 13$ groups, showing how it can penalize models with many groups. The LSP-category prior imposes a similar restriction, but to a lesser extent as there are seven unique categories in the data. The posterior corresponding to the DP prior is centered closer to $K = 22$ and has a larger spread than the other distributions. This is likely because the DP prior induces a distribution over K that is more flat than the LSP prior.

Lastly, we examine the results from the restricted model with the covariate-dependent LSP prior. Figure 10 depicts the influence of feature and display advertising on the posterior partitioning behavior. For both covariates, we find a negative relationship between the distance in covariate space between two products and their posterior probability of being

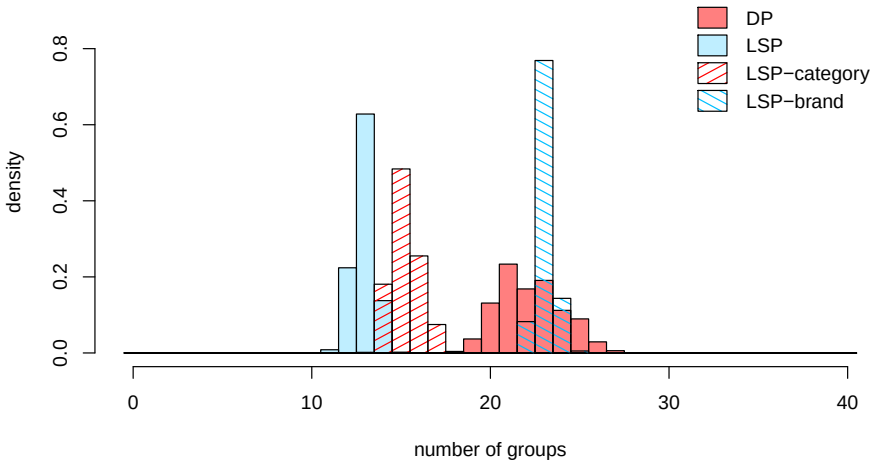


Figure 9: The posterior distribution over the number of groups K is plotted for the restricted models with a DP prior, LSP prior, LSP-category prior, and LSP-brand prior.

grouped together. That is, if two products tend to have similar frequencies of feature or display advertising (i.e., low covariate distance), then they have a higher chance of being grouped together a posteriori. In particular, this correlation is -0.18 for feature advertising and -0.21 for display advertising. The negative correlations appear reasonable, as co-occuring store circulator ads for two products may actually generate cross-price effects between them.

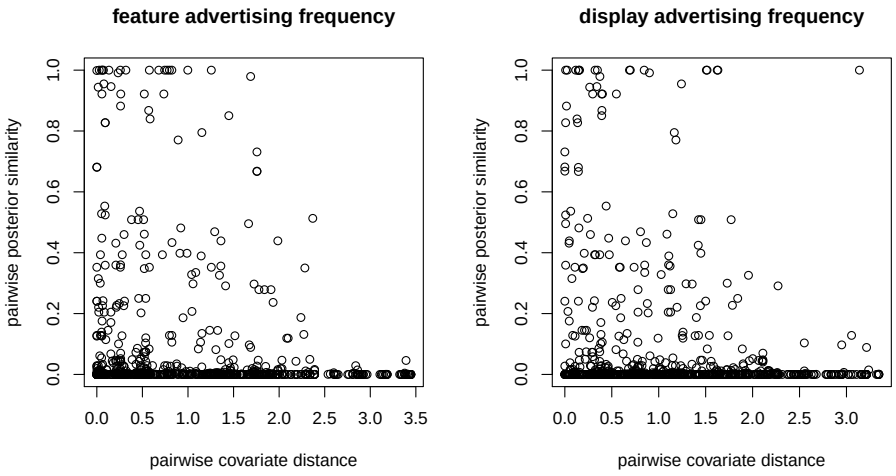


Figure 10: Pairwise posterior grouping probabilities are plotted against pairwise covariate similarity for the restricted model with the covariate-dependent LSP prior.

6 Discussion

This paper presents a Bayesian method of joint inference for the class of partitioned demand models. We build on previous nonparametric Bayesian models for random partitions to construct the LSP distribution, which is a formal probability distribution on $\mathcal{P}_n$ indexed by a location partition ρ_n and scale parameter τ. The LSP model serves two purposes. First, it is used as a proposal distribution within a random-walk MH algorithm. We find this approach to be especially effective in situations where incremental updates of the item-group indicator variables are either inefficient or intractable. The LSP distribution is also used as a prior which can be used to incorporate other covariate information.

The value of the LSP distribution is then illustrated empirically in the context of an aggregate demand model applied to data in the salty snack category. We include products that span many subcategories, which usually presents challenges in inference due to limited data and quadratic growth in the number of demand parameters. This problem of modeling and forecasting demand across a wide assortment of goods is common in retail settings. Our solution is to partition the cross-elasticity matrix into groups of related and unrelated goods. This induces equality restrictions on sets of cross-elasticity parameters while still retaining model flexibility when the partition is estimated. We find that imposing these restrictions improves demand forecasts, increases the precision of elasticity estimates, and allows us to learn about the structure of category competition.

There are several limitations and possible extensions of the current work. First, we have assumed throughout that the number of observations exceeds the number of products. In many practical large-scale demand settings, this may not be the case. It would then be useful to explore how other regularization priors, which are commonly used in high-dimensional settings, could be used to generate similar group-level shrinkage patterns for really large-scale demand problems. Additionally, while our empirical application has focused exclusively on store-level sales data, partitioned demand models can also be estimated with household-level choice data. When combined with a household-level partitioned demand model, the LSP

distribution could be used as a hierarchical prior to model heterogeneity in the partition along with the heterogeneity in usual set of demand parameters. However, the required number of observations per household may be high in order to get traction on the household-level posterior distributions.

This paper has also only considered single-layer, non-overlapping partitions. Other types of partition structures (e.g., multi-layer partitions and/or partitions with overlapping clusters) are common in areas of regression tree modeling and network analysis. The advantages offered by these methods are likely to be problem and model specific. For example, if the goal is to simply predict demand or find zero or non-zero cross price effects, then our model is flexible enough and the particular topology of the random grouping structure is less important. We also believe that the ideas presented here could be used to modify more flexible partitioning models such as the Indian buffet process (Griffiths and Ghahramani, 2011), which allows items to appear in multiple groups.

Furthermore, in economic demand models it may also be the case that the covariates included in the demand equation or LSPx prior (e.g., prices or promotion incidence and frequency) are set strategically by the firm. That is, the firm may choose to coordinate prices or promotions across products in anticipation of some marketplace response. To formally control for these supply-side effects, we would need to integrate a supply-side model (i.e., a likelihood function for the covariates) with the demand equation and then generate samples from the resulting joint posterior. While simultaneous models of supply and demand have been developed in the literature (e.g., Yang et al., 2003), price is usually the only strategic variable of the firm. Extending this work into the class of partitioned demand models would allow the firm to also affect the structure of product categories as perceived by the consumer.

Finally, more work is needed on the convergence and scaling properties of Markov chains on high-dimensional discrete spaces like $\mathcal{P}_n$. The challenge is that many traditional diagnostic statistics (e.g., Gelman and Rubin, 1992) are based on normal-theory approximations to the posterior, which seem unreasonable given the non-Euclidean dimension of the posteriors

studied here. Visual inspection of trace plots can be a useful diagnostic tool, but can also be unreliable as the number of items grows. For example, it is common in high-dimensional settings for the Markov chain to not visit any partition more than once. This also challenges the use of many conventional posterior summary statistics, such as maximum a posteriori estimates. Recent work by Wade and Ghahramani (2018) develops posterior credible balls for partition parameters to more formally characterize uncertainty, which would be a useful extension to our empirical setting. We leave these extensions for future work.

A Proofs

Property 1 (Location-Scale Consistency)**.** *If $\pi_n \sim LSP(\rho_n, \tau)$, then for any number of items n and location partition $\rho_n \in \mathcal{P}_n$,*

$$\lim_{\tau \to 0} \Pr(\pi_n = \rho_n | \rho_n, \tau) = 1.$$

Proof. Since each item-group assignment variable g_i is generated sequentially, we use mathematical induction to show that $g_i = s_i$ as $\tau \to 0$ for $i = 1, \ldots, n$. Since $g_1 = s_1 = 1$ trivially, we use $i = 2$ as a base case.

BASE CASE: There are two cases to consider.

(i) Suppose $s_2 = s_1$. We must show that as τ approaches zero, the probability that item 2 starts a new group goes to zero ($w_0(\cdot) \to 0$) and the probability that it joins the first group goes to one ($w_1(\cdot) \to 1$).

$$w_0(s_2, \tau) = \tilde{c}_2 \cdot \frac{\tau + 1(s_2 = C^{(2)} + 1)}{\tau C^{(2)} + \tau + 1} = \tilde{c}_2 \cdot \frac{\tau}{\tau C^{(2)} + \tau + 1} \to 0$$
$$w_1(\{s_2, S_1\}, \tau) = \tilde{c}_2 \cdot \frac{\tau + n_{S_1}^{s_2}}{\tau C^{(2)} + \tau + n_1} = \tilde{c}_2 \cdot \frac{\tau + 1}{\tau C^{(2)} + \tau + 1} \to 1$$

(ii) Suppose $s_2 = C^{(2)} + 1 \neq s_1$. We must show that as τ approaches zero, the probability that item 2 starts a new group goes to one ($w_0(\cdot) \to 1$) and the probability that it joins the first group goes to zero ($w_1(\cdot) \to 0$).

$$w_0(s_2, \tau) = \tilde{c}_2 \cdot \frac{\tau + 1(s_2 = C^{(2)} + 1)}{\tau C^{(2)} + \tau + 1} = \tilde{c}_2 \cdot \frac{\tau + 1}{\tau C^{(2)} + \tau + 1} \to 1$$
$$w_1(\{s_2, S_1\}, \tau) = \tilde{c}_2 \cdot \frac{\tau + n_{S_1}^{s_2}}{\tau C^{(2)} + \tau + n_1} = \tilde{c}_2 \cdot \frac{\tau}{\tau C^{(2)} + \tau + 1} \to 0$$

INDUCTIVE STEP: Assume that $g_i = s_i$ for $i = 1, \ldots, j-1$ where $j < n + 1$. We wish to show that $g_j = s_j$. There are again two cases to consider.

(i) Suppose $s_j = c$ where $c \in \{1, \ldots, C^{(j)}\}$. We must show that as τ approaches zero, the probability that item j starts a new group goes to zero ($w_0(\cdot) \to 0$), the

probability that it joins group c goes to one ($w_c(\cdot) \to 1$), and the probability it joins any other group $k \neq c$ goes to zero ($w_k(\cdot) \to 0$).

$$w_0(s_j, \tau) = \tilde{c}_j \cdot \frac{\tau + 1(s_j = C^{(j)} + 1)}{\tau C^{(j)} + \tau + 1} = \tilde{c}_j \cdot \frac{\tau}{\tau C^{(j)} + \tau + 1} \to 0$$
$$w_c(\{s_j, S_c\}, \tau) = \tilde{c}_j \cdot \frac{\tau + n_{S_c}^{s_j}}{\tau C^{(j)} + \tau + n_c} = \tilde{c}_j \cdot \frac{\tau + n_c}{\tau C^{(j)} + \tau + n_c} \to 1$$
$$w_k(\{s_j, S_k\}, \tau) = \tilde{c}_j \cdot \frac{\tau + n_{S_k}^{s_j}}{\tau C^{(j)} + \tau + n_k} = \tilde{c}_j \cdot \frac{\tau}{\tau C^{(j)} + \tau + n_k} \to 0$$

(ii) Suppose $s_j = C^{(j)} + 1$. We must show that as τ approaches zero, the probability that item j starts a new group goes to one ($w_0(\cdot) \to 1$) and the probability that it joins group k goes to zero ($w_k(\cdot) \to 0$) for any $k = 1, \ldots, K^{(j)}$.

$$w_0(s_j, \tau) = \tilde{c}_j \cdot \frac{\tau + 1(s_j = C^{(j)} + 1)}{\tau C^{(j)} + \tau + 1} = \tilde{c}_j \cdot \frac{\tau + 1}{\tau C^{(j)} + \tau + 1} \to 1$$
$$w_k(\{s_j, S_k\}, \tau) = \tilde{c}_j \cdot \frac{\tau + n_{S_k}^{s_j}}{\tau C^{(j)} + \tau + n_k} = \tilde{c}_j \cdot \frac{\tau}{\tau C^{(j)} + \tau + n_k} \to 0$$

□

Property 2 (Marginal Invariance)**.** *If $\pi_n \sim$ LSP(ρ_n, τ), then for any number of items n, location partition $\rho_n \in \mathcal{P}_n$, scale parameter $\tau > 0$, and distribution $p(s_{n+1})$ such that $\sum_{s_{n+1}} p(s_{n+1}) = 1$,*

$$p(\pi_n | \rho_n, \tau) = \sum_{g_{n+1}=1}^{K+1} \sum_{s_{n+1}=1}^{C+1} p(\pi_{n+1} | \rho_n, s_{n+1}, \tau) p(s_{n+1})$$

where $K = \max\{g_1, \ldots, g_n\}$ and $C = \max\{s_1, \ldots, s_n\}$.

Proof. First pick an arbitrary value of $s_{n+1} \in \{1, \ldots, C + 1\}$. By the sequential nature of

the Pólya-urn scheme, we have

$$\begin{aligned}
\sum_{g_{n+1}=1}^{K+1} p(\pi_n, g_{n+1}|\rho_n, s_{n+1}, \tau) &= \sum_{g_{n+1}=1}^{K+1} p(g_{n+1}|\pi_n, \rho_n, s_{n+1}, \tau) p(\pi_n|\rho_n, s_{n+1}, \tau) \\
&= p(\pi_n|\rho_n, \tau) \sum_{g_{n+1}=1}^{K+1} p(g_{n+1}|\pi_n, \rho_n, s_{n+1}, \tau) \\
&= p(\pi_n|\rho_n, \tau)\big[w_0(\cdot) + w_1(\cdot) + \ldots + w_K(\cdot)\big] \\
&= p(\pi_n|\rho_n, \tau).
\end{aligned}$$

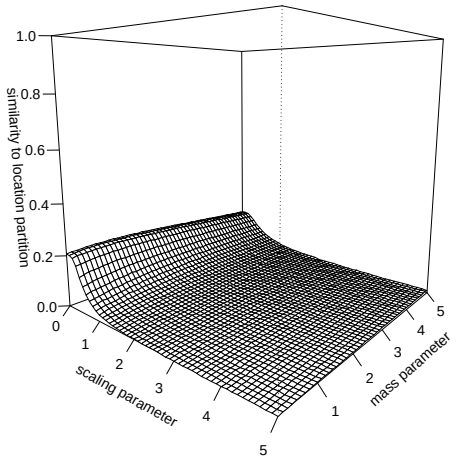
Since $p(\pi_n|\rho_n, \tau)$ does not depend on s_{n+1} and $\sum_{s_{n+1}=1}^{C+1} p(s_{n+1}) = 1$, it follows that

$$\begin{aligned}
\sum_{g_{n+1}=1}^{K+1} \sum_{s_{n+1}=1}^{C+1} p(\pi_n, g_{n+1}|\rho_n, s_{n+1}, \tau) p(s_{n+1}) &= \sum_{s_{n+1}=1}^{C+1} \left[\sum_{g_{n+1}=1}^{K+1} p(\pi_n, g_{n+1}|\rho_n, s_{n+1}, \tau) \right] p(s_{n+1}) \\
&= \sum_{s_{n+1}=1}^{C+1} p(\pi_n|\rho_n, \tau) p(s_{n+1}) \\
&= p(\pi_n|\rho_n, \tau) \sum_{s_{n+1}=1}^{C+1} p(s_{n+1}) \\
&= p(\pi_n|\rho_n, \tau)
\end{aligned}$$

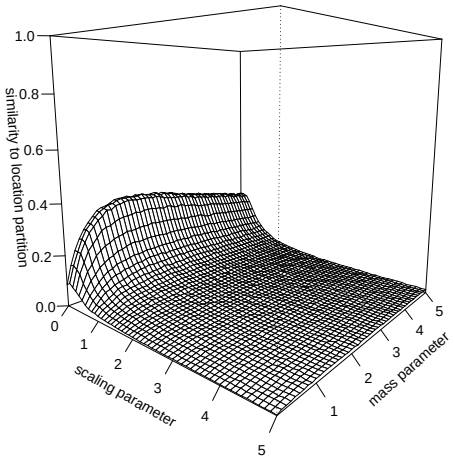
as desired. □

B Behavior of the ddCRP and EPA Models

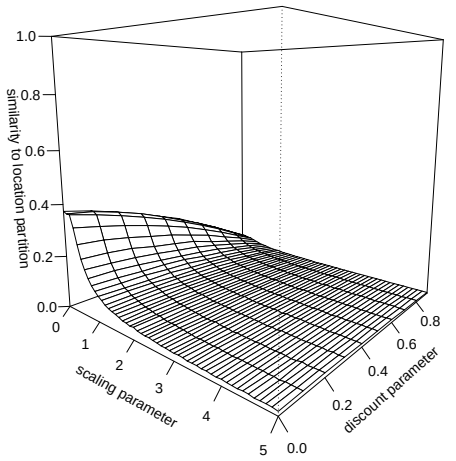
(a) ddCRP



(b) EPA: discount parameter = 0



(c) EPA: mass parameter = 2



(d) EPA: scaling parameter = 0

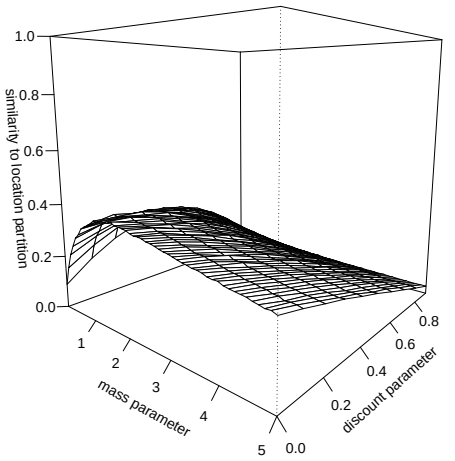


Figure B.1: Each plot shows the extent to which the ddCRP and EPA distributions can be centered around a location partition ρ_n. For each partition distribution, 10,000 random partitions of length $n = 10$ are drawn and then compared to $\rho_n = (1, 1, 1, 1, 1, 2, 2, 2, 2, 2)$ using the adjusted Rand index. The EPA and ddCRP distributions are both parameterized by an exponential decay function and by the pairwise distance matrix induced by ρ_n. The surface of each plot then shows the averaged adjusted Rand index across values of the scaling, mass, or discount parameters.

C Product Descriptions

Table C.1: Salty Snack Product Descriptions

	Brand	Subcategory	Subcategory Volume Share	Feature Frequency	Display Frequency
1	BARREL O FUN	CHESNK	23.24	1.56	31.97
2	CHEETOS	CHESNK	76.76	13.13	29.43
3	BARREL O FUN	CRNSNK	13.12	0.48	66.46
4	BUGLES	CRNSNK	19.17	18.63	20.53
5	FRITOS	CRNSNK	38.42	22.14	30.84
6	FRITOS SCOOPS	CRNSNK	26.97	0.18	34.04
7	OLD DUTCH	CRNSNK	2.32	0.00	1.91
8	GARDETTOS	OTHER	12.06	19.41	11.24
9	GENERAL MILLS CHEX MIX	OTHER	47.07	16.84	18.49
10	MUNCHOS	OTHER	1.26	0.00	0.00
11	PRIVATE LABEL	OTHER	2.15	0.00	0.00
12	S & W PIK NIK	OTHER	3.48	0.00	0.00
13	SUNCHIPS	OTHER	33.99	21.05	47.58
14	BAKED LAYS	PTOCHP	2.48	0.48	19.25
15	BAKED RUFFLES	PTOCHP	2.45	6.70	19.40
16	BARREL O FUN	PTOCHP	4.69	5.96	27.02
17	LAYS	PTOCHP	27.14	35.09	58.54
18	OLD DUTCH	PTOCHP	9.07	4.20	12.03
19	POORE BROTHERS	PTOCHP	3.78	0.96	21.25
20	PRINGLES	PTOCHP	14.89	5.47	6.34
21	PRINGLES CHEEZUMS	PTOCHP	1.81	5.26	3.83
22	PRINGLES FAT FREE	PTOCHP	1.40	0.00	0.00
23	PRINGLES RIGHT CRISPS	PTOCHP	1.76	4.88	4.81
24	PRIVATE LABEL	PTOCHP	8.80	9.61	11.35
25	RUFFLES	PTOCHP	9.93	13.24	33.17
26	WAVY LAYS	PTOCHP	11.81	34.98	51.96
27	BARREL O FUN	PRETZL	4.07	0.00	25.84
28	OLD DUTCH	PRETZL	12.42	11.31	0.69
29	PRIVATE LABEL	PRETZL	27.19	13.46	14.99
30	ROLD GOLD	PRETZL	45.32	11.09	37.28
31	SNYDERS OF HANOVER	PRETZL	11.00	3.40	16.78
32	BARREL O FUN	POPCRN	28.08	0.00	2.75
33	CRUNCH N MUNCH	POPCRN	25.56	20.93	0.00
34	OLD DUTCH	POPCRN	46.37	1.68	0.00
35	BAKED TOSTITOS	TTACHP	3.03	0.96	20.57
36	BARREL O FUN	TTACHP	18.21	2.92	32.96
37	DORITOS	TTACHP	40.57	25.00	58.92
38	GARDEN OF EATIN BLUE CHIPS	TTACHP	1.14	3.53	7.02
39	OLD DUTCH	TTACHP	5.07	7.49	14.74
40	TOSTITOS	TTACHP	31.98	12.90	37.84

D MH Step for the Isolated Demand Model

1. Generate the candidate partition

$$\pi_n^* \sim q_1(\pi_n|\pi_n^{(r)}, v) = \mathrm{LSP}(\pi_n^{(r)}, v)$$

where $v = 1/(n\log(n))$. Then conditional on π_n^*, generate $\boldsymbol{\beta}_{\pi_n}^*$ from its full conditional distribution

$$\boldsymbol{\beta}_{\pi_n}^* \sim q_2(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}_{\pi_n^*}, \pi_n^*, \Sigma) = \mathrm{N}(\tilde{\boldsymbol{\beta}}, (\tilde{\boldsymbol{X}}'_{\pi_n^*}\tilde{\boldsymbol{X}}_{\pi_n^*} + A_{\pi_n^*})^{-1})$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}'_{\pi_n^*}\tilde{\boldsymbol{X}}_{\pi_n^*}+A_{\pi_n^*})^{-1}(\tilde{\boldsymbol{X}}'_{\pi_n^*}\tilde{\boldsymbol{y}}+A_{\pi_n^*}\bar{\boldsymbol{\beta}}_{\pi_n^*})$, $\tilde{\boldsymbol{X}}_{\pi_n^*} = ((U^{-1})'\otimes I)\boldsymbol{X}_{\pi_n^*}$, and $\Sigma = U'U$.

2. Set $(\pi_n^{(r+1)}, \boldsymbol{\beta}_{\pi_n}^{(r+1)}) = (\pi_n^*, \boldsymbol{\beta}_{\pi_n}^*)$ with probability

$$\begin{aligned}&\mathcal{A}(\pi_n^*, \boldsymbol{\beta}_{\pi_n}^*, \pi_n^{(r)}, \boldsymbol{\beta}_{\pi_n}^{(r)})\\ &= \min\left\{1, \frac{p(\boldsymbol{y}|\boldsymbol{X}_{\pi_n^*}, \boldsymbol{\beta}_{\pi_n}^*, \pi_n^*, \Sigma)p(\boldsymbol{\beta}_{\pi_n}^*|\pi_n^*)p(\pi_n^*)}{p(\boldsymbol{y}|\boldsymbol{X}_{\pi_n^{(r)}}, \boldsymbol{\beta}_{\pi_n}^{(r)}, \pi_n^{(r)}, \Sigma)p(\boldsymbol{\beta}_{\pi_n}^{(r)}|\pi_n^{(r)})p(\pi_n^{(r)})} \times \frac{q_2(\boldsymbol{\beta}_{\pi_n}^{(r)}|\pi_n^{(r)})q_1(\pi_n^{(r)}|\pi_n^*)}{q_2(\boldsymbol{\beta}_{\pi_n}^*|\pi_n^*)q_1(\pi_n^*|\pi_n^{(r)})}\right\}.\end{aligned}$$

Otherwise, set $(\pi_n^{(r+1)}, \boldsymbol{\beta}_{\pi_n}^{(r+1)}) = (\pi_n^{(r)}, \boldsymbol{\beta}_{\pi_n}^{(r)})$.

References

Ainslie, A. and Rossi, P. E. (1998). Similarities in Choice Behavior Across Product Categories. *Marketing Science*, 17(2):91–106.

Airoldi, E. M., Costa, T., Bassetti, F., and Leisen, F. (2014). Generalized species sampling priors with latent beta reinforcements. *Journal of the American Statistical Association*, 109(508):1466–1480.

Barry, D. and Hartigan, J. (1992). Product Partition Models for Change Point Problems. *The Annals of Statistics*, 20(1):260–279.

Blackwell, D. and MacQueen, J. B. (1973). Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1(2):353–355.

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.

Blei, D. M. and Frazier, P. I. (2011). Distance Dependent Chinese Restaurant Processes. *Journal of machine Learning research*, 12(Aug):2461–2488.

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

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

Chen, Y. and Yang, S. (2007). Estimating Disaggregate Models Using Aggregate Data through Augmentation of Individual Choice. *Journal of Marketing Research*, 44(4):613–621.

Chib, S. and Ramamurthy, S. (2010). Tailored randomized block MCMC methods with application to DSGE models. *Journal of Econometrics*, 155(1):19–38.

Chib, S., Seetharaman, P. B., and Strijnev, A. (2002). Analysis of multi-category purchase incidence decisions using IRI market basket data. In *Advances in Econometrics*, pages 57–92.

Dahl, D. B. (2003). An Improved Merge-Split Sampler for Conjugate Dirichlet Process Mixture Models. *Technical Report*.

Dahl, D. B., Day, R., and Tsai, J. W. (2017). Random Partition Distribution Indexed by Pairwise Information. *Journal of the American Statistical Association*, 0(0):1–12.

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

DellaVigna, S. and Gentzgow, M. (2017). Uniform Pricing in US Retail Chains. *NBER Working Paper*.

Dotson, J. P., Howell, J. R., Brazell, J. D., Otter, T., Lenk, P. J., MacEachern, S., and Allenby, G. M. (2018). A Probit Model with Structured Covariance for Similarity Effects and Source of Volume Calculations. *Journal of Marketing Research*, 55(1):35–47.

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ewens, W. J. (1972). The Sampling Theory of Selectively Neutral Alleles. *Theoretical Population Biology*, 3:87–112.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230.

Ferreira, J. A., Loschi, R. H., and Costa, M. A. (2014). Detecting changes in time series: A product partition model with across-cluster correlation. *Signal Processing*, 96:212–227.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Gilbride, T. J. and Allenby, G. M. (2004). A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules. *Marketing Science*, 23(3):391–406.

Goldman, S. 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.

Green, P. J. and Richardson, S. (2001). Modelling Heterogeneity with and without the Dirichlet Process. *Scandinavian Journal of Statistics*, 28:355–375.

Griffiths, T. L. and Ghahramani, Z. (2011). The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, 12:1185–1224.

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

Hans, C. (2011). Elastic Net Regression Modeling With the Orthant Normal Prior. *Journal of the American Statistical Association*, 106(496):1383–1393.

Hansen, K., Singh, V., and Chintagunta, P. (2006). Understanding Store-Brand Purchase Behavior Across Categories. *Marketing Science*, 25(1):75–90.

Hartigan, J. (1990). Partition Models. *Communications in Statistics - Theory and Methods*, 19(8):2745–2756.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression:Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Hubert, L. and Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, 2(1):193–218.

Hui, S. K. and Bradlow, E. T. (2012). Bayesian Multi-Resolution Spatial Analysis with Applications to Marketing. *Quantitative marketing and economics*, 10(4):419–452.

Jain, S. and Neal, R. M. (2004). A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

Kim, D. S., Bailey, R. A., Hardt, N., and Allenby, G. M. (2017). Benefit-Based Conjoint Analysis. *Marketing Science*, 36(1):54–69.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes. *Biometrika*, 81(1):27–40.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238.

McFadden, D. (1978). *Modelling Choice of Residential Location*. Amsterdam: North-Holland.

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. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Monteiro, J. V. D., Assuncao, R. M., and Loschi, R. H. (2011). Product partition models with correlated parameters. *Bayesian Analysis*, 6(4):691–726.

Montgomery, A. L. (1997). Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data. 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.

Müller, P., Quintana, F., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics.

Müller, P., Quintana, F., and Rosner, G. L. (2011). A Product Partition Model With Regression on Covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278.

Musalem, A., Bradlow, E. T., and Raju, J. S. (2009). Bayesian Estimation of Random-Coefficients Choice Models Using Aggregate Data. 24(3):490–516.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Page, G. L. and Quintana, F. A. (2016). Spatial product partition models. *Bayesian Analysis*, 11:265–298.

Park, J. H. and Dunson, D. B. (2010). Bayesian Generalized Product Partition Model. *Statistica Sinica*, 20:1203–1226.

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

Pitman, J. (1995). Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, 102:145–158.

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In *Statistics, Probability and Game Theory IMS Lecture Notes - Monograph Series*, pages 245–267. Institute of Mathematical Statistics.

Pitman, J. and Yor, M. (1997). The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *The Annals of Probability*, 25(2):855–900.

Pudney, S. E. (1981). An Empirical Method of Approximating the Separable Structure of Consumer Preferences. 48(4):561–577.

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, 136(8):2407–2429.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(2):557–574.

Roberts, G. O. and Rosenthal, J. S. (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, 16(4):351–367.

Roberts, G. O. and Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(2):291–317.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing*. John Wiley and Sons Ltd.

Ročková, V. and George, E. I. (2018). The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444.

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000). Structured Markov Chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 9(2):217–234.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Song, I. and Chintagunta, P. K. (2006). Measuring Cross-Category Price Effects with Aggregate Store Data. *Management Science*, 52(10):1594–1609.

Strotz, R. H. (1957). The Empirical Implications of a Utility Tree. *Econometrica*, 25(2):269–280.

Thomassen, Ø., Smith, H., Seiler, S., and Schiraldi, P. (2017). Multi-Category Competition and Market Power: A Model of Supermarket Pricing. *American Economic Review*, 107(8):2308–2351.

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

Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, 2 edition.

Turek, D., de Valpine, P., Paciorek, C. J., and Anderson-Bergman, C. (2017). Automated Parameter Blocking for Efficient Markov Chain Monte Carlo Sampling. *Bayesian Analysis*, 12(2):465–490.

Wade, S. and Ghahramani, Z. (2018). Bayesian Cluster Analysis: Point Estimation and Credible Balls. *Bayesian Analysis*, 13(2):559–626.

Wedel, M. and Zhang, J. (2004). Analyzing Brand Competition across Subcategories. *Journal of Marketing Research*, 41(4):448–456.

Yang, S., Chen, Y., and Allenby, G. M. (2003). Bayesian Analysis of Simultaneous Demand and Supply. *Quantitative Marketing and Economics*, 1(3):251–275.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68(1):49–67.

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.