Scalable Estimation of Multinomial Response Models with Random Consideration Sets*

Siddhartha Chib
Olin School of Business
Washington University,
in St. Louis, USA

Kenichi Shimizu
Department of Economics
University of Alberta,
Canada

September 1, 2024

Abstract

A common assumption in the fitting of unordered multinomial response models for J mutually exclusive categories is that the responses arise from the same set of J categories across subjects. However, when responses measure a choice made by the subject, it is more appropriate to condition the distribution of multinomial responses on a subject-specific consideration set, drawn from the power set of $\{1, 2, \ldots, J\}$. This leads to a mixture of multinomial response models governed by a probability distribution over the $J^* = 2^J - 1$ consideration sets. We introduce a novel method for estimating such generalized multinomial response models based on the fundamental result that any mass distribution over J^* consideration sets can be represented as a mixture of products of J component-specific inclusion-exclusion probabilities. Moreover, under time-invariant consideration sets, the conditional posterior distribution of consideration sets is sparse. These features enable a scalable MCMC algorithm for sampling the posterior distribution of parameters, random effects, and consideration sets. Under regularity conditions, the posterior distributions of the marginal response probabilities and the model parameters satisfy consistency. The methodology is demonstrated in a longitudinal data set on weekly cereal purchases that cover $J = 101$ brands, a dimension substantially beyond the reach of existing methods.

Keywords: Multinomial response, Bayesian computation, Dirichlet process mixture, Markov chain Monte Carlo, Metropolis-Hastings algorithm, Posterior consistency

*Email: chib@wustl.edu and kenichi.shimizu@ualberta.ca. Disclaimer: Researchers' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researchers and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

1 Introduction

A common assumption when fitting unordered multinomial response models, whether applied to cross-sectional or longitudinal data, is that the responses stem from the same set of J mutually exclusive categories across all subjects. However, this assumption may be questionable, especially when modeling the choices made by human subjects. For example, in fields such as economics and marketing, it is recognized that individuals may select from only a subset of the available alternatives, termed the "consideration set" (Manski, 1977; Honka et al., 2019). Neglecting this heterogeneity in the consideration sets can result in biased parameter estimates in the model (Bronnenberg and Vanhonacker, 1996; Chiang et al., 1998; Goeree, 2008; Draganska and Klapper, 2011; De los Santos, 2018; Morozov et al., 2021; Crawford et al., 2021). Such biases are problematic because these models are typically employed to understand the impact of covariates on outcomes and inform decision making.

To address this issue, it is necessary to generalize the standard multinomial response model by conditioning the distribution of responses on a latent subject-specific consideration set, which is drawn from the power set of $\{1, 2, \ldots, J\}$. This results in a mixture of multinomial models based on a probability distribution over consideration sets. However, the exponential size of this power set renders the estimation of this mixture of multinomial response models computationally infeasible in general.

In order to fix ideas, let $\mathcal{C}_i$ represent the latent consideration set for subject i. When J alternatives are available, $\mathcal{C}_i$ is a subset of $\{1, \ldots, J\}$, and there are $J^* = 2^J - 1$ possible consideration sets. A priori, $\mathcal{C}_i$ is assumed to be drawn from a probability mass function $\Pr(\mathcal{C}_i = c)$. When J is small, the direct approach proposed by Chiang et al. (1998) is effective. In this approach, all possible consideration sets $1, 2, \ldots, J^*$ are enumerated and assigned unknown probabilities $\pi_1, \pi_2, \ldots, \pi_{J^*}$, which can be estimated using MCMC methods under a Dirichlet prior. However, when J is large, the model has traditionally been estimated under the assumption that the distribution over consideration sets is determined by J independent attention probabilities. In this framework, it is assumed that each alternative appears independently in any given consideration set (Ben-Akiva and Boccara, 1995; Goeree, 2008; Manzini and Mariotti, 2014; Kawaguchi et al., 2021; Abaluck and Adams-Prassl, 2021). Specifically, let q_{ij} denote the probability that subject i considers the alternative j for $j = 1, \ldots, J$. The probability that $\mathcal{C}_i = c$ given $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{iJ})'$ is then modeled as:

$$\Pr\left(\mathcal{C}_i = c \mid \boldsymbol{q}_i\right) = \prod_{j \in c} q_{ij} \prod_{j \notin c} \left(1 - q_{ij}\right).$$

Although this model is appealing for handling the large J case, the distribution over consideration sets is unrealistic and leads to model misspecification (Crawford et al., 2021).

In another approach, the consideration sets are modeled as vectors of 0-1 binary variables (Van Nierop et al., 2010). This vector is then modeled by a multivariate probit model (Albert and Chib, 1993, Chib and Greenberg, 1998). Although this can generate correlation of items in consideration sets, inference is challenging because the number of parameters in the correlation matrix of the multivariate probit model increases quadratically in J.

Given the significant interest in incorporating consideration set heterogeneity in various fields - such as marketing (Van Nierop et al., 2010; Ching et al., 2014; Kawaguchi et al., 2021), economics (Goeree, 2008; Ching et al., 2009; Kashaev et al., 2019; Agarwal and Somaini, 2022), transportation science (Swait and Ben-Akiva, 1987; Paleti et al., 2021), and psychology (Traets et al., 2022) - there is a pressing need to develop a scalable estimation approach for estimating such generalized multinomial response models. The importance of accounting for consideration set heterogeneity becomes even more critical as J increases, which is precisely the case that current methods struggle to address. The method we propose is based on two key components. The first component is a representation of the probability masses $\pi_1, \pi_2, \ldots, \pi_{J^*}$ in terms of a weighted average of products of item-specific inclusion q_j and exclusion $1 - q_j$ probabilities, which is based on a result from Dunson and Xing (2009). We refer to this approach as a mixture of independent consideration models. To simulate the latent consideration sets, we introduce a straightforward and intuitive Metropolis-Hastings algorithm. It is important to highlight that, in this context, the consideration sets are latent, unlike in Dunson and Xing (2009), where the categorical variables are observed. This difference necessitates additional steps in both the theoretical derivations and the computational procedure. Another crucial feature of the method is the sparsity of the posterior distribution of the consideration sets, which occurs because sets that do not include the actual choices made by a subject must have a posterior probability of zero (Chiang et al., 1998). The scalability of the proposed approach is demonstrated through an application to marketing data involving $J = 101$ brands.

We establish two key theoretical results. First, under regularity conditions, as the number of subjects increases, we demonstrate that the posterior distribution of the marginal response probabilities is consistent. Second, under certain additional identification assumptions, the posterior distribution of the model parameters also achieves consistency.

In general, this paper contributes to the expanding literature on high-dimensional demand estimation in statistics and marketing: (Braun and McAuliffe, 2010; Chiong and Shum, 2019; Smith and Allenby, 2019; Loaiza-Maya and Nibbering, 2022; Jiang et al., 2024; Iaria and Wang, 2024; Ershov et al., 2024; Amano et al., 2018). Moreover, the proposed method can be interpreted as a generalized multinomial logit (MNL) model, with "structural zeros" incorporated in the first layer of its hierarchical structure. In the field of biostatistics, methodologies have been extensively explored to estimate microbial compositions that account for the sparsity due to excessive zero counts (e.g. Aitchison, 1982; Martín-Fernández et al., 2015; Liu et al., 2020; Cao et al., 2020; Paulson et al., 2013; Chen and Li, 2016; Tang and Chen, 2019). More recently, Zeng et al. (2023) introduced a zero-inflated probabilistic PCA model designed for high-dimensional, sparse microbiome data sets. Although our paper focuses on a different problem, the proposed method has the potential to be applied in similar contexts, particularly in scenarios where structural zeros exhibit complex dependency patterns.

The remainder of the paper is organized as follows. Section 2 introduces the model, and Section 3 establishes theoretical results. Section 4 considers the prior-posterior distribution. Section 5 develops the approach to posterior computation. Section 6 presents numerical simulations. An application to a marketing data set is given in Section 7.

2 The approach

Suppose that we have panel (longitudinal) data with n a priori independent subjects that contains multinomial (polychotomous) responses from a set $\mathcal{J} = \{1, \ldots, J\}$ of J mutually exclusive nominal categories/items as well as some covariates. Let $Y_{it} \in \mathcal{J}$ be the measured response for unit i at time t, where $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$. Let $\boldsymbol{\omega}_{it} = \{\boldsymbol{\omega}_{ijt}\}_{j\in\mathcal{J}}$, where $\boldsymbol{\omega}_{ijt}$ is the vector of covariates characterizing the category j for subject i at time t. Each subject i is associated with a latent consideration set $\mathcal{C}_i$, which is a subset of the entire set of alternatives $\mathcal{J}$. We model the distribution of the observed outcomes using a hierarchical approach. Specifically, we first specify the marginal distribution of the consideration sets and then define the response distribution conditional on a given consideration set. In this framework, we make the following assumptions.

Assumption 1: Consideration sets $\mathcal{C}_i$ vary over subjects but not over time, and the distribution over consideration sets, denoted by $\pi_c = \Pr(\mathcal{C}_i = c)$ for $c \in \mathcal{C}$, the set of all possible consideration sets minus the empty set, is free of covariates.

The assumption of time invariance is relatively mild and aids in inference. It also plays a role in the identification of model parameters. Covariates can be included in the model for consideration sets, but, as noted by Chiang et al. (1998), a covariate-dependent model is difficult to specify without increasing the risk of model mis-specification.

Assumption 2: For each $j \in \mathcal{J}$, the responses Y_{it} of subject i given $\mathcal{C}_i$ and random effects $\boldsymbol{b}_i$ are independent over time and follow the multinomial logit model.

Based on Assumptions 1 and 2, the generalized multinomial logit model of interest has the hierarchical form:

$$
\begin{aligned}
&\text{Stage 1:} \quad \mathcal{C}_i \overset{iid}{\sim} \boldsymbol{\pi}, \\
&\text{Stage 2:} \quad \boldsymbol{b}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{D}), \qquad\qquad (1)\\
&\text{Stage 3:} \quad \Pr(Y_{it} = j \mid \boldsymbol{\beta}, \boldsymbol{\omega}_{it}, \mathcal{C}_i, \boldsymbol{b}_i) = \begin{cases} \dfrac{\exp\left(\boldsymbol{x}_{ijt}'\boldsymbol{\beta} + \boldsymbol{z}_{ijt}'\boldsymbol{b}_i\right)}{\sum_{\ell \in \mathcal{C}_i} \exp\left(\boldsymbol{x}_{i\ell t}'\boldsymbol{\beta} + \boldsymbol{z}_{i\ell t}'\boldsymbol{b}_i\right)} & \text{if } j \in \mathcal{C}_i \\ 0 & \text{otherwise} \end{cases} \quad t = 1, \ldots, T_i,
\end{aligned}
$$

for $i = 1, \ldots, n$, where $\boldsymbol{\pi} = \{\pi_c : c \in \mathcal{C}, 0 \le \pi_c \le 1, \sum_{c \in \mathcal{C}} \pi_c = 1\}$ denotes the collection of probabilities associated with all possible consideration sets, and $\boldsymbol{b}_i$ are random effects normally and independently distributed across subjects with zero mean and unknown covariance matrix $\boldsymbol{D}$. The covariates are denoted by $\boldsymbol{\omega}_{it} = \{\boldsymbol{x}_{ijt}, \boldsymbol{z}_{ijt}\}_{j \in \mathcal{J}}$, where $\boldsymbol{x}_{ijt} \in \mathbb{R}^{d_x}$ and $\boldsymbol{z}_{ijt} \in \mathbb{R}^{d_z}$. Stage 1 can be interpreted as introducing another layer of random effects, where heterogeneity arises from the random consideration sets.

Letting $\Pr(\boldsymbol{Y}_i | \boldsymbol{\theta}, \boldsymbol{\omega}_{it}, \mathcal{C}_i = c)$ denote the distribution of outcomes $\boldsymbol{Y}_i = (Y_{1i}, \ldots, Y_{iT_i})$ of subject i marginalized over the random effects, the distribution of responses takes the finite mixture form:

$$\Pr(\boldsymbol{Y}_i|\boldsymbol{\theta}, \boldsymbol{\omega}_{it}) = \sum_{c \in \mathcal{C}} \pi_c \Pr(\boldsymbol{Y}_i|\boldsymbol{\theta}, \boldsymbol{\omega}_{it}, \mathcal{C}_i = c),$$

Note that the number of terms in the summation increases exponentially in J. Also note that we adopt the logit specification for simplicity, but our approach can in principle be used with other link functions, such as the probit. However, starting with an alternative link is not really necessary because the marginal model of the outcomes is already a generalized multinomial logit due to the mixing over the consideration sets.

Assumptions 1 and 2 imply time-invariant consideration sets, conditional independence of responses, and full support of the conditional response probabilities on consideration sets. These conditions, along with additional assumptions detailed below, establish the point identification of the model parameters (Aguiar and Kashaev, 2024). Furthermore, in Theorem 2 of Section 3, we demonstrate the posterior consistency of the model parameters under these assumptions.

2.1 The latent consideration sets

To fix notation, let $\mathcal{C}$ represent the collection of all possible consideration sets, which corresponds to the power set of $\mathcal{J} = \{1, \ldots, J\}$, excluding the empty set. The consideration set for subject i is indicated by $\mathcal{C}_i = c$, where $c \in \mathcal{C}$. For example, when $J = 3$, $\mathcal{C} = \{\{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \text{and} \{1,2,3\}\}$, and c is one of these elements. Furthermore, by $\boldsymbol{C}_i = (C_{i1}, \ldots, C_{iJ})'$, we mean a $J \times 1$ multivariate binary vector where $C_{ij} = 1$ if the category j is in the consideration set and 0 otherwise. In the example of $J = 3$, $\mathcal{C}_i = \{1\}$ is equivalent to $\boldsymbol{C}_i = (1, 0, 0)'$ and $\mathcal{C}_i = \{1, 3\}$ is equivalent to $\boldsymbol{C}_i = (1, 0, 1)'$ etc. In the following, we use the two notations interchangeably depending on the context. Researchers sometimes include an outside option in the model that is always considered by each subject. We can incorporate this into our framework by adding a $(J+1)$th category and fixing $C_{iJ+1} = 1$ for all i. Our goal is to put a probability distribution on $\mathcal{C}$ that is rich enough to accommodate dependencies while maintaining scalability.

2.2 Dimensionality reduction via tensor decomposition

We now review the factor decomposition technique that we employ to specify the distribution over consideration sets. Dunson and Xing (2009) consider modeling large contingency tables that, for example, represent DNA sequences, each of which is defined as a collection of J categorical variables, each having d_j possible values $j = 1, \ldots, J$, where J is large. A realization of the contingency table can be expressed as a vector $(a_1, \ldots, a_J)'$, where $a_j \in \{1, \ldots, d_j\}$ for $j = 1, \ldots, J$. The true distribution of the contingency tables is a probability tensor $\boldsymbol{\pi} = \{\pi_{a_1 a_2 \cdots a_J}, a_j = 1, \ldots, d_j, j = 1, \ldots, J\}$, where $0 \leq \pi_{a_1 a_2 \cdots a_J} \leq 1$ and $\sum_{a_1=1}^{d_1} \cdots \sum_{a_J=1}^{d_J} \pi_{a_1 a_2 \cdots a_J} = 1$. Note that consideration sets can be seen as contingency tables with $d_j = 2$ for all j. Generally, there are a large number of elements in the tensor $\boldsymbol{\pi}$, $d_1 \times \cdots \times d_J$, when J is large. Dunson and Xing (2009) show that $\boldsymbol{\pi}$ can be expressed as a finite mixture of rank 1 tensors. We describe this result for the special case that corresponds to modeling consideration sets.

Lemma 1 (Exact matching of consideration set probabilities)**.** *Let $\boldsymbol{\pi}$ be the probability mass distribution over the consideration sets: It is a collection of probabilities $\{\pi_c = \Pr(\mathcal{C}_i = c) : c \in \mathcal{C}\}$, where $0 \leq \pi_c \leq 1$ and $\sum_{c \in \mathcal{C}} \pi_c = 1$. Then there are $K \in \mathbb{Z}^+$, $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K) \in \Delta^{K-1}$, $\boldsymbol{q}_h = (q_{h1}, \ldots, q_{hJ})'$, $\quad h = 1, \ldots, K, \quad q_{hj} \in [0,1]$ such that for each $c \in \mathcal{C}$,*

$$\pi_c = \sum_{h=1}^{K} \omega_h \left\{ \prod_{j \in c} q_{hj} \prod_{j \notin c} (1 - q_{hj}) \right\}. \tag{2}$$

It states that a mixture of K independent consideration models can model an arbitrary distribution over the $J^* = 2^J - 1$ possible consideration sets. Within each component h, items are included in or excluded from a consideration set c according to an independent consideration model defined by a vector of attention probabilities $\boldsymbol{q}_h = (q_{h1}, \ldots, q_{hJ})'$. Therefore, the number of parameters needed to model the probabilities in $\boldsymbol{\pi}$ is reduced from J^* to $K \times J + (K-1)$, which scales linearly with J, providing the basis for scalability.

2.3 Infinite mixture of independent consideration models

Building on this result, we model the J-dimensional latent vectors $\{\boldsymbol{C}_i\}$ via a mixture of independent probabilities. In practice, the number of components K in (2) is unknown. To overcome this issue, following Dunson and Xing (2009), we use a Dirichlet process (DP) prior (Ferguson, 1973), to induce an infinite mixture model. It is important to emphasize that our problem differs from Dunson and Xing (2009) in the sense that while the categorical variables (contingency tables) are observed in their paper, the corresponding objects (the consideration sets) are latent in the current problem. This distinction introduces additional challenges both in the theory in Section 3 and in computation, since now an efficient algorithm is required to sample the latent consideration sets efficiently even in high-dimensional scenarios; see our proposed approach in Section 5. An alternative approach that begins by estimating K could offer computational advantages. However, existing methods for consistently estimating K, such as those proposed by Kwon and Mbakop (2021), may not be directly applicable in our context, where the variables modeled by the mixture are latent. In contrast, our posterior consistency results only require that the prior on K has positive mass for all positive integers. This flexibility allows for posterior inference on model parameters and their functions (e.g., predictions) to automatically account for uncertainty regarding the value of K.

We now describe our approach. Assume that $\{\boldsymbol{C}_i\}$ is i.i.d. with density $f(\,\cdot\,|G) = \int \prod_{j=1}^{J} q_j^{C_{ij}} (1-q_j)^{1-C_{ij}}\, dG(\boldsymbol{q})$. The discrete mixing distribution G is modeled by a DP prior with a concentration parameter α and a specified base probability measure G_0 that depends on a hyperparameter $\underline{\boldsymbol{\phi}}_q$. Equivalently, by using the stick breaking construction (Sethuraman, 1994), we have the following representation: $\boldsymbol{C}_i$'s are i.i.d. with the density for the infinite mixture of independent consideration models:

$$\Pr(\boldsymbol{C}_i = \boldsymbol{c}_i) = \sum_{h=1}^{\infty} \omega_h \prod_{j=1}^{J} \left\{ q_{hj}^{c_{ij}} (1-q_{hj})^{1-c_{ij}} \right\}, \tag{3}$$

where $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{iJ})'$, $\omega_1 = V_1, \omega_h = V_h \prod_{\ell<h}(1-V_\ell), h = 2, \ldots, \infty$, $V_h \overset{iid}{\sim} \text{Beta}(1, \alpha)$, and $\boldsymbol{q}_h \overset{iid}{\sim} G_0(\,\cdot\,|\underline{\boldsymbol{\phi}}_q), h = 1, \ldots, \infty$, with $\boldsymbol{q}_h = (q_{h1}, \ldots, q_{hJ})'$ being the vector of attention probabilities specific to the component h. A priori, the first few weights dominate and cover most of the probability mass, which are then adjusted by the data. Although the model (3) includes infinitely many components, typically only a small number of distinct values for $\boldsymbol{q}_h$ are imputed.

For the baseline distribution G_0, we assume that $q_{hj} \sim G_{0j}$ independently for $j = 1, \ldots, J$ and $h = 1, \ldots, \infty$. Specifically, we assume that $q_{hj} \sim \text{Beta}(\underline{a}_{q_j}, \underline{b}_{q_j})$, independently over $j = 1, \ldots, J$, for $h = 1, \ldots, \infty$, and we define $\underline{\boldsymbol{\phi}}_q = (\underline{\boldsymbol{a}}_q, \underline{\boldsymbol{b}}_q)$ with $\underline{\boldsymbol{a}}_q = (\underline{a}_{q_1}, \ldots, \underline{a}_{q_J})'$ and $\underline{\boldsymbol{b}}_q = (\underline{b}_{q_1}, \ldots, \underline{b}_{q_J})'$. Note that $\underline{\boldsymbol{\phi}}_q = (\underline{\boldsymbol{a}}_q, \underline{\boldsymbol{b}}_q)$ are the hyperparameters chosen by the user. We discuss this in more detail in the Supplementary Material. We complete the model specification by assuming the prior distribution for the DP concentration parameter $\alpha \sim \text{Gamma}(\underline{a}_\alpha, \underline{b}_\alpha)$, where $(\underline{a}_\alpha, \underline{b}_\alpha)$ are the hyperparameters chosen by the user. For smaller values of α, ω_h decreases toward zero more rapidly as h increases, so that the prior favors a sparse representation with most of the weight on a few components. We allow the data to inform us about α and, therefore, an appropriate degree of sparsity.

3 Theoretical results

In this section, we establish posterior consistency results under the proposed approach. For simplicity, let $T_i = T$, $\forall i$. We consider the framework where $T \geq 1$ is fixed and $n \to \infty$. Theorem 1 states that the posterior of the marginal response probabilities is consistent, and in Theorem 2, we establish that the posterior of the model parameters is consistent when T is large enough and the model does not include random effects. Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{D}\}$ denote the parameters in the response model. Also, recall that the distribution over the consideration sets is denoted by $\boldsymbol{\pi} = \{\pi_c : c \in \mathcal{C}\}$, where $0 \leq \pi_c \leq 1$ and $\sum_{c\in\mathcal{C}} \pi_c = 1$. Define the probability that the sequence of items $\boldsymbol{y} = (y_1, \ldots, y_T)' \in \mathcal{J}^T$ is chosen conditional on covariates $\boldsymbol{w}_i = \{\boldsymbol{w}_{i1}, \ldots, \boldsymbol{w}_{iT}\}$ taking some specific value $\boldsymbol{w} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T\} \in \mathbb{R}^{TJ(d_x+d_z)}$:

$$p_{\boldsymbol{\theta},\boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{w}) \equiv \sum_{c\in\mathcal{C}} \pi_c \Pr\left(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{w}, c\right),$$

where the response probability given a consideration set c is

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{w}, c) = \int \prod_{t=1}^{T} \Pr(Y_{it} = y_t \mid \boldsymbol{\beta}, \boldsymbol{w}_t, \mathcal{C}_i = c, \boldsymbol{b}_i)\phi(\boldsymbol{b}_i|0, \boldsymbol{D})d\boldsymbol{b}_i,$$

where the integrand is defined in (1). The data set contains responses $\boldsymbol{y}_i = \{y_{it}\}$ and covariates $\boldsymbol{w}_i = \{\boldsymbol{w}_{it}\}$: $\boldsymbol{D}^n = \{(\boldsymbol{y}_i, \boldsymbol{w}_i) : i = 1, \ldots, n\}$. The covariates $\boldsymbol{w}_i$ are i.i.d. and follow an unknown distribution with density g^* with support $\mathcal{W} \subset \mathbb{R}^{TJ(d_x+d_z)}$. We do not model the covariate distribution. Conditional on covariates, responses are generated from the collection of the data-generating response probabilities $\boldsymbol{p}^* = \{p_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w})\}_{\boldsymbol{y}\in\mathcal{J}^T,\boldsymbol{w}\in\mathcal{W}}$, where $\boldsymbol{\theta}^*$ denotes the true response model parameter and $\boldsymbol{\pi}^* = \{\pi^*_c : c \in \mathcal{C}\}$ denotes the true probability mass function over consideration sets. We emphasize that $\boldsymbol{\pi}^*$ does not have to be a finite mixture. The joint probability measure implied by $\boldsymbol{p}^*$ and g^* is denoted by F_0. For $\varepsilon > 0$, define a Kullback-Leibler neighborhood of $\boldsymbol{p}^*$ as

$$KL_\varepsilon(\boldsymbol{p}^*) = \left\{ (\boldsymbol{\theta}, \boldsymbol{\pi}) : \int \sum_{\boldsymbol{y} \in \mathcal{J}^T} \log \left(\frac{p_{\boldsymbol{\theta}^*, \boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w})}{p_{\boldsymbol{\theta}, \boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{w})} \right) p_{\boldsymbol{\theta}^*, \boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w}) g^*(\boldsymbol{w}) d\boldsymbol{w} < \varepsilon \right\}.$$

It is essentially a set of $(\boldsymbol{\theta}, \boldsymbol{\pi})$ that makes $p_{\boldsymbol{\theta}, \boldsymbol{\pi}}$ close to $p_{\boldsymbol{\theta}^*, \boldsymbol{\pi}^*}$.

Given a $K \in \mathbb{Z}^+$, define $\boldsymbol{\phi}_{1:K} = \{\omega_h, \boldsymbol{q}_h : h = 1, \ldots, K\}$, the collection of all component-specific parameters, where $\boldsymbol{q}_h = (q_{h1}, \ldots, q_{hJ})'$. Note that by Lemma 1, there exist $\{K, \tilde{\boldsymbol{\phi}}_{1:K}\}$, which may not be unique, such that $\pi_c^* = \sum_{h=1}^K \tilde{\omega}_h \left\{ \prod_{j \in c} \tilde{q}_{hj} \prod_{j \notin c} \left(1 - \tilde{q}_{hj}\right) \right\}$, for all $c \in \mathcal{C}$, and the KL divergence is zero at $\{\boldsymbol{\theta}^*, K, \tilde{\boldsymbol{\phi}}_{1:K}\}$. In the following lemma, we establish that the KL divergence can be made arbitrarily small in sufficiently small neighborhoods of $(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\phi}}_{1:K})$. Define the model induced probability for a consideration set $c \in \mathcal{C}$: $\pi(c|K, \boldsymbol{\phi}_{1:K}) = \sum_{h=1}^K \omega_h \prod_{j \in c} q_{hj} \prod_{j \notin c}(1 - q_{hj})$, and the model induced marginal response probability as

$$p(\boldsymbol{y}|\boldsymbol{w}; \boldsymbol{\theta}, K, \boldsymbol{\phi}_{1:K}) = \sum_{c \in \mathcal{C}} \pi(c|K, \boldsymbol{\phi}_{1:K}) \Pr(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{w}, c).$$

Lemma 2. *Suppose: (i) $\boldsymbol{\beta}^* \in interior(\mathcal{B})$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_x}$ and $\boldsymbol{D}^*$ is positive definite, and (ii) $\mathcal{W}$ is compact. Then $\forall \varepsilon > 0$, $\exists$ an open neighborhood $\mathcal{O}$ of $\boldsymbol{\theta}^*$, $K \in \mathbb{Z}^+$, and an open neighborhood $\mathcal{P}^K$ such that for any $\boldsymbol{\theta} \in \mathcal{O}$ and $\boldsymbol{\phi}_{1:K} \in \mathcal{P}^K$,*

$$\int \sum_{\boldsymbol{y} \in \mathcal{J}^T} \log \left(\frac{p_{\boldsymbol{\theta}^*, \boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{w}; \boldsymbol{\theta}, K, \boldsymbol{\phi}_{1:K})} \right) p_{\boldsymbol{\theta}^*, \boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w}) g^*(\boldsymbol{w}) d\boldsymbol{w} < \varepsilon.$$

The proof can be found in Appendix. Let $\Pi(\cdot)$ denote the prior for the response model parameter $\boldsymbol{\theta}$ and the distribution of consideration sets $\boldsymbol{\pi}$.

Theorem 1. *Suppose conditions (i) and (ii) of Lemma 2. Suppose (iii) for any open neighborhood $\mathcal{O}$ of $\boldsymbol{\theta}^*$, and for any $K, \boldsymbol{\phi}_{1:K}$, and an open neighborhood $\mathcal{P}^K$ of $\boldsymbol{\phi}_{1:K}$, $\Pi(\boldsymbol{\theta} \in \mathcal{O}, \boldsymbol{\phi}_{1:K} \in \mathcal{P}^K, K) > 0$. Then, for all weak neighborhood $\mathcal{U}$ of $\boldsymbol{p}^*$, as $n \to \infty$, $\Pi\left(\mathcal{U}|\boldsymbol{D}^n\right) \to 1$ a.s. F_0^∞.*

Proof of Theorem 1. By Schwartz's theorem (Ghosal and van der Vaart, 2017, ch.6), the result follows if we show that $\Pi(KL_\varepsilon(\boldsymbol{p}^*)) > 0$. By Lemma 2, there exist open neighborhoods $\mathcal{O}$ and $\mathcal{P}^K$ on which the KL divergence can be made sufficiently small. The lemma combined with a prior that places positive mass on open neighborhoods (condition iii) implies that $\Pi(KL_\varepsilon(\boldsymbol{p}^*)) > 0$. □

It states that asymptotically, the posterior converges in a sense that the model-induced response probability is consistent with the true data-generating counterpart. A similar result is obtained in Theorem 2 of Dunson and Xing (2009), but in the context of observed categorical variables and no covariates. In contrast, in our more general setup, we have latent categorical variables (consideration sets) in the first layer of the hierarchical model (1), which must be integrated out, and covariates are also present. These differences add to the complexity of the theoretical analysis, requiring additional steps to prove posterior consistency, such as the need to incorporate the KL-divergence in the analysis (Lemma 2).

In the context of semiparametric estimation of dynamic discrete choice models, Norets and Shimizu (2024) obtained a similar result. Our proof strategy is similar, but different, due to the presence of random effects, continuous covariates, a different model, and also the need to work with the KL-divergence. The compactness assumption (ii) is common in Bayesian nonparametric estimation. Condition (iii) of Theorem 1 is satisfied by our prior: the DP prior for ω_h's and the Beta prior for q_{hj}'s as shown by Dunson and Xing (2009).

Theorem 1 establishes posterior consistency in terms of marginal response probabilities. However, there could be multiple pairs of $(\boldsymbol{\theta}, \boldsymbol{\pi})$ that are consistent with the true response probabilities. This issue of partial identification has attracted much attention recently (Masatlioglu et al., 2012, Cattaneo et al., 2020, Barseghyan et al., 2021a, Lu, 2022), and some articles have studied conditions/setting in which the model is point-identified (Dardanoni et al. (2020), Abaluck and Adams-Prassl, 2021, Barseghyan et al., 2021b). Aguiar and Kashaev (2024) provide conditions under which the distribution of consideration sets and the conditional response probabilities are non-parametrically identified when there are no random effects (i.e. $\boldsymbol{\theta} = \boldsymbol{\beta}$) and consideration sets are time-invariant. We employ one of their identifying conditions: the panel is "long enough" so that one can pin down the two sources of variation in responses.

The next theorem builds on this identification result and establishes that our posterior distribution is consistent in the sense that it contracts to within arbitrarily small ball around $(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*)$ given by the following distance function: $d((\boldsymbol{\beta}, \boldsymbol{\pi}), (\boldsymbol{\beta}', \boldsymbol{\pi}')) = \max\{||\boldsymbol{\pi} - \boldsymbol{\pi}'||_1, ||\boldsymbol{\beta} - \boldsymbol{\beta}'||_2\}$. The difference from Theorem 1 is that the added identifying assumption enables us to conclude that the posterior is consistent for $(\boldsymbol{\beta}^*, \boldsymbol{\pi}^*)$ rather than for the marginal response probability $\boldsymbol{p}_{\boldsymbol{\beta}^*, \boldsymbol{\pi}^*}$.

Theorem 2. *In addition to the assumptions in Theorem 1, suppose that $\boldsymbol{\beta} \in \mathcal{B} \subseteq \mathbb{R}^{d_x}$, where $\mathcal{B}$ is compact, and that the number of periods T is such that the largest integer that is smaller than or equal to $\frac{T-3}{2}$ is larger than J. Then, $\forall \varepsilon > 0$, as $n \to \infty$, $\Pi\left((\boldsymbol{\beta}, \boldsymbol{\pi}) : d((\boldsymbol{\beta}, \boldsymbol{\pi}), (\boldsymbol{\beta}^*, \boldsymbol{\pi}^*)) < \varepsilon | \boldsymbol{D}^n\right) \to 1$ a.s. F_0^∞.*

Proof of Theorem 2. The proof is by Schwartz's theorem. The identification assumption together with Assumptions 1-2 ensures that $p_{\boldsymbol{\beta}, \boldsymbol{\pi}} \neq p_{\boldsymbol{\beta}', \boldsymbol{\pi}'}$ whenever $(\boldsymbol{\beta}, \boldsymbol{\pi}) \neq (\boldsymbol{\beta}', \boldsymbol{\pi}')$ (Aguiar and Kashaev, 2024). Identifiability, continuity of $p_{\boldsymbol{\beta}, \boldsymbol{\pi}}$ in $(\boldsymbol{\beta}, \boldsymbol{\pi})$ for the total variation norm

(Lemma SA.3), and compactness of the parameter space ensure the existence of consistent tests (Van der Vaart, 2000, Lemma 10.6). The approximation result (Lemma 2) without random effects can be established as a special case, and together with the regularity conditions on the prior distribution, the KL-support condition holds. □

The differences from Aguiar and Kashaev (2024) include the fact that the authors work with a more general nonparametric setting for the conditional response distribution given consideration sets while we impose the logistic specification. Another point is that inferential framework is not available in their paper while one can easily conduct inference of model parameters and their functions using the posterior distributions in our approach.

4 Inference

4.1 Data structure

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})'$ and $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})'$ be the sequence of random responses made by unit i over T_i periods and its observed counterpart. Let $\boldsymbol{\omega}_i = (\boldsymbol{\omega}'_{i1}, \ldots, \boldsymbol{\omega}'_{iT_i})'$ be the covariates for the subject i observed over time. Define

$$p(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_i, \boldsymbol{C}_i) = \prod_{t=1}^{T_i} \Pr(Y_{it} = y_{it} | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i), \tag{4}$$

where $\Pr(Y_{it} = y_{it} | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i)$ is

$$\Pr(Y_{it} = j | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i) = \frac{\exp\left(\boldsymbol{x}'_{ijt}\boldsymbol{\beta} + \boldsymbol{z}'_{ijt}\boldsymbol{b}_i\right)}{\sum_{\ell \in \mathcal{C}_i} \exp\left(\boldsymbol{x}'_{i\ell t}\boldsymbol{\beta} + \boldsymbol{z}'_{i\ell t}\boldsymbol{b}_i\right)} \text{ if } j \in \mathcal{C}_i, \text{ and } 0 \text{ otherwise.} \tag{5}$$

Note that $\boldsymbol{C}_i$ is the conditioning variable on the left side of (4), while $\mathcal{C}_i$ is on the right side. Although the two objects represent the same information, the J-dimensional vector $\boldsymbol{C}_i$ is easier to use when we discuss posterior sampling of individual consideration sets. Hence, we use $\boldsymbol{C}_i$ to define the individual's contribution to the likelihood. Let $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\}$ and $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ denote the random and observed sequences of the responses made by all units, and let $\boldsymbol{\omega} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n\}$ be the observed covariates. Then the likelihood conditional on the common fixed-effects $\boldsymbol{\beta}$, the random effects $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)'$, the covariates $\boldsymbol{\omega}$, and the latent consideration sets $\boldsymbol{C} = (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n)$ is given by

$$p(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{C}) = \prod_{i=1}^{n} p(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_i, \boldsymbol{C}_i). \tag{6}$$

We complete the model by specifying standard prior distributions for the parameters in the response model: $\boldsymbol{\beta} \sim \mathcal{N}_{d_x}\left(\boldsymbol{0}, \underline{\boldsymbol{V}}_{\beta}\right)$ and $\boldsymbol{D}^{-1} \sim \text{Wishart}\,(\underline{v}, \underline{\boldsymbol{R}})$, indepdently, a normal distribution for $\boldsymbol{\beta}$, and an inverse Wishart distribution for $\boldsymbol{D}$ with degrees-of-freedom parameter $\underline{v}$ and scale matrix $\underline{\boldsymbol{R}}$. The hyperparameters $(\underline{\boldsymbol{V}}_{\beta}, \underline{v}, \underline{\boldsymbol{R}})$ are chosen by the user.

4.2 Posterior distribution

For the mixture model on the latent consideration sets $\boldsymbol{C} = (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n)$, let $S_i \in \{1, 2, \ldots\}$ be the latent cluster assignment such that $C_{ij} | S_i = h \sim \text{Bernoulli}(q_{hj})$, independently $j = 1, \ldots, J$, for $i = 1, \ldots, n$. We have the latent consideration sets $\boldsymbol{C}$, the common fixed-effects $\boldsymbol{\beta}$, the random effects $\boldsymbol{b}$, the corresponding covariance matrix $\boldsymbol{D}$, the DP parameters $\boldsymbol{V} = (V_1, V_2, \ldots)$ as well as $\boldsymbol{Q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots)$, the DP cluster assignment variables $\boldsymbol{S} = (S_1, \ldots, S_n)$, and the DP concentration parameter α. Let $\pi(\cdot)$ denote the prior density. Then, from the Bayes theorem, the posterior density of interest is

$$p\big(\boldsymbol{C},\ \boldsymbol{S}, \boldsymbol{V}, \boldsymbol{Q}, \alpha, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{D} \big| \boldsymbol{y}, \boldsymbol{\omega}\big) \propto p\big(\boldsymbol{y} \big| \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{C}\big) \cdot p(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{D}) \cdot p\big(\boldsymbol{C}, \boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{V}, \alpha\big), \tag{7}$$

where the first term is given by (6) and only the last term is associated with the DP prior.

5 Computation

MCMC methods can be applied to efficiently sample the posterior distribution. The algorithm we present is scalable and is constructed from simple and intuitive steps. Posterior inferences are based on the sample of draws produced by the algorithm. The posterior sample consists of $\{V_h^{(g)}\}$, $\{\boldsymbol{q}_h^{(g)}\}$, $\{S_i^{(g)}\}$, $\alpha^{(g)}$, $\{\boldsymbol{C}_i^{(g)}\}$, $\boldsymbol{\beta}^{(g)}$, $\{\boldsymbol{b}_i^{(g)}\}$, and $\boldsymbol{D}^{(g)}$ for $g = 1, \ldots, G$, where G is the number of MCMC draws (beyond a suitable burn-in).

5.1 Simulation of consideration sets

We now focus on sampling the conditional distribution of consideration sets. The other steps in the MCMC simulation follow from standard calculations and are given in the Supplementary Material. From Equation (7), the full conditional distribution of $\boldsymbol{C}_i$ is

$$\pi(\boldsymbol{C}_i|\boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{q}_{S_i}, S_i, \boldsymbol{y}_i, \boldsymbol{\omega}_i) \propto \ p\big(\boldsymbol{Y}_i = \boldsymbol{y}_i \big| \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_i, \boldsymbol{C}_i\big) \cdot \prod_{j=1}^{J} q_{S_i j}^{C_{ij}} (1 - q_{S_i j})^{1 - C_{ij}}, \tag{8}$$

where the proportionality sign is with respect to $\boldsymbol{C}_i$, and the first term is defined in (4). Importantly, consideration sets that exclude any observed response made by subject i receive zero posterior probability (see Table 1 for an example). This is because the first term on the left-hand side of (8) is zero for these consideration sets. This desirable feature of our approach is based on Chiang et al. (1998). In contrast, in many existing methods, every consideration set receives a strictly positive probability, as pointed out by Crawford et al. (2021). Now, due to the independence structure in (8) over $j = 1, \ldots, J$,

$$\pi(C_{ij}|\boldsymbol{C}_i \setminus \{j\}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{q}_{S_i}, S_i, \boldsymbol{y}_i, \boldsymbol{\omega}_i) \propto \; p\big(\boldsymbol{Y}_i = \boldsymbol{y}_i \big| \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_i, \boldsymbol{C}_i\big) \cdot q_{S_i j}^{C_{ij}}(1 - q_{S_i j})^{1 - C_{ij}},$$

where $\boldsymbol{C}_i \setminus \{j\}$ denotes $\boldsymbol{C}_i$ without the coordinate j. To sample from this distribution, we employ the Metropolis-Hastings (M-H) algorithm (Chib and Greenberg, 1995). An effective implementation of this approach is detailed in Algorithm 1.

Algorithm 1: M-H step for Sampling Consideration Sets

Input: The current draws at the gth iteration $\{\boldsymbol{C}_i^{(g)}\}, \{\boldsymbol{q}_h^{(g)}\}, \{S_i^{(g)} = h\}, \boldsymbol{\beta}^{(g)}, \{\boldsymbol{b}_i^{(g)}\}$

Output: The updated consideration sets $\{\boldsymbol{C}_i^{(g+1)}\}$

for $i \in \{1, \ldots, n\}$ **do**

  for $j \in \{1, \ldots, J\}$ **do**

    1) Propose $\tilde{C}_{ij} \sim \text{Bernoulli}(q_{hj}^{(g)})$ and define $\boldsymbol{C}_i^{(1)} = (C_{i1}^{(g+1)}, \ldots, C_{ij-1}^{(g+1)}, \tilde{C}_{ij}, C_{ij+1}^{(g)}, \ldots, C_{iJ}^{(g)})'$

    2) Accept $\tilde{C}_{ij}$ with probability

$$\min\left\{\frac{p\big(\boldsymbol{Y}_i = \boldsymbol{y}_i \big| \boldsymbol{\beta}^{(g)}, \boldsymbol{b}_i^{(g)}, \boldsymbol{\omega}_i, \boldsymbol{C}_i^{(1)}\big)}{p\big(\boldsymbol{Y}_i = \boldsymbol{y}_i \big| \boldsymbol{\beta}^{(g)}, \boldsymbol{b}_i^{(g)}, \boldsymbol{\omega}_i, \boldsymbol{C}_i^{(0)}\big)}, 1\right\},$$

    where $\boldsymbol{C}_i^{(0)} = (C_{i1}^{(g+1)}, \ldots, C_{ij-1}^{(g+1)}, C_{ij}^{(g)}, C_{ij+1}^{(g)}, \ldots, C_{iJ}^{(g)})'$.

    Otherwise, set $C_{ij}^{(g+1)} = C_{ij}^{(g)}$

In Step 1 of Algorithm 1, we generate a proposal from a one-dimensional Bernoulli distribution. In Step 2, given the current state $\boldsymbol{C}_i^{(0)}$ and the proposed state $\boldsymbol{C}_i^{(1)}$, the acceptance probability is computed as the ratio of the likelihood contributions for subject i. This Metropolis-Hastings step is valid because the likelihood $p\big(\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_i, \boldsymbol{C}_i\big)$ is uniformly bounded. See Chib and Greenberg (1995) (p. 330, "the third algorithm") for more discussion. In practice, we update the states in a random order within each MCMC iteration. In addition, the computational burden is minimized by parallelizing the loop on the n subjects. Finally, the proposed Metropolis-Hastings step exhibits an important "sparsity property," which is elaborated on in the Supplementary Material.

5.2 Numerical illustration

We illustrate posterior probabilities on consideration sets in our approach, using a synthetic panel data with $n = 100$ subjects with time periods $T \in \{1, 2, \ldots, 10\}$. Let $J = 4$ so that it is possible to present all the the $2^J - 1 = 15$ consideration sets as in the first column of Table 1. The table also shows the posterior probabilities of the consideration sets for a randomly chosen subject $i = A$ whose true consideration set is $\mathcal{C}_A^* = \{1, 3, 4\}$.

Table 1: Posterior probability of consideration sets for the unit $i = A$

	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$	$T=6$	$T=7$	$T=8$	$T=9$	$T=10$
$\{1\}$	0.074	0	0	0	0	0	0	0	0	0
$\{2\}$	0	0	0	0	0	0	0	0	0	0
$\{3\}$	0	0	0	0	0	0	0	0	0	0
$\{4\}$	0	0	0	0	0	0	0	0	0	0
$\{1,2\}$	0.06	0	0	0	0	0	0	0	0	0
$\{1,3\}$	0.167	0.593	0.832	0.837	0	0	0	0	0	0
$\{1,4\}$	0.062	0	0	0	0	0	0	0	0	0
$\{2,3\}$	0	0	0	0	0	0	0	0	0	0
$\{2,4\}$	0	0	0	0	0	0	0	0	0	0
$\{3,4\}$	0	0	0	0	0	0	0	0	0	0
$\{1,2,3\}$	0.13	0.136	0.071	0.069	0	0	0	0	0	0
$\{1,2,4\}$	0.045	0	0	0	0	0	0	0	0	0
$\{1,3,4\}$	0.251	0.179	0.077	0.079	0.8	0.88	0.855	0.931	0.951	0.964
$\{2,3,4\}$	0	0	0	0	0	0	0	0	0	0
$\{1,2,3,4\}$	0.211	0.092	0.02	0.015	0.2	0.12	0.145	0.069	0.049	0.036
y_{iT}	1	3	1	1	4	1	1	4	1	4
Acc.Rate.	0.885	0.819	0.814	0.757	0.64	0.648	0.646	0.674	0.673	0.683

The results are based on a synthetic panel data with $J = 4$ and $n = 100$. The true consideration set is $\mathcal{C}_A^* = \{1, 3, 4\}$. The row $y_{A,T}$ shows the actual response made by subject A at time T. Acc. Rate denotes the acceptance rate of consideration sets in the M-H step.

The first column ($T = 1$) shows the results for the first period at which the subject's response was 1. The consideration sets that do not contain the item 1 receive a posterior probability of zero. With one period, the posterior support has been already reduced from 15 points to 8, which is a desirable feature of our approach. In the second period, the subject's response was 3, and the posterior support in the column ($T = 2$) is now only 4. As T increases, the posterior concentrates at the true consideration set $\{1, 3, 4\}$.

6 Monte Carlo Simulation

We demonstrate the performance of the proposed approach through simulation studies, starting with a small number of alternatives ($J = 4$). This allows for the enumeration of all possible consideration sets, making it easier to present the results. In addition, a simulation study with $J = 100$ is provided in the Supplementary Material, and we further demonstrate the scalability of our approach in an application involving $J = 101$ alternatives.

Our goal is to empirically validate the findings of Theorem 2 and demonstrate that the proposed approach can effectively assess consideration dependence. We consider a balanced panel where $T_i = T = 3$ for all i. Although this panel length is smaller than the requirement in Theorem 2, it still demonstrates good convergence. To simulate the data, we first specify the distribution of the consideration sets $\boldsymbol{\pi}^* = \{\pi_c^* = \Pr(\mathcal{C}_i = c) : c \in \mathcal{C}\}$. We assume that product consideration is segmented and dependent so that the first two and the last two products are more likely to be considered together, each with a relatively high probability: $\pi^*(\{1, 2\}) = \pi^*(\{3, 4\}) = 0.25$. As motivation, the first two products could be seen as "non-vegetarian options," and the last two as "vegetarian." The other consideration sets are assigned the same small probability of 0.0385. Figure 1 shows $\boldsymbol{\pi}^*$ in red. Note that the true $\boldsymbol{\pi}^*$ itself is not defined as a finite mixture, but in the fitting, we use the infinite mixture model. The true consideration sets $\mathcal{C}_i^*$, for $i = 1, \ldots, n$ are generated from this $\boldsymbol{\pi}^*$. We then generate outcomes from the logit model with $V_{ijt} = \delta_j^* + \beta^* x_{ijt}$, where $(\delta_1^*, \delta_2^*, \delta_3^*, \delta_4^*)' = (0.5, -0.5, 0.3, 0)'$ and $\beta^* = 1$, and $x_{ijt} \overset{iid}{\sim} N(0, 1)$.

6.1 Consistency

We compare the performance between the proposed infinite mixture of independent consideration models ($K = \infty$) and the model that assumes independent consideration ($K = 1$).

The analysis is repeated 50 times. Table 2 shows that as n increases, the posterior distribution of both the response parameter $\boldsymbol{\beta} = (\delta_1, \delta_2, \delta_3, \beta)$ and the distribution of the consideration sets $\boldsymbol{\pi}$ converges to the true values, under the proposed model ($K = \infty$). In contrast, when $K = 1$, we do not observe sufficient evidence of posterior consistency. In particular, MSEs are an order of magnitude larger in some cases than those under $K = \infty$, due to misspecification.

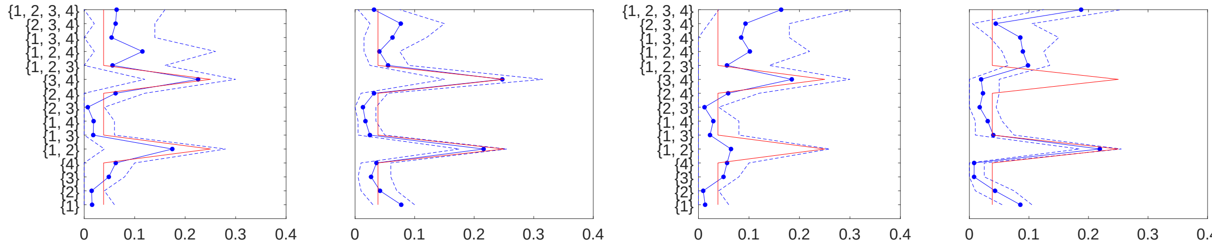
Table 2: Simulation results with $J = 4$

		β		δ_1		δ_2		δ_3		π	
$K = \infty$	$n = 50$	0.07	(0.23)	0.507	(0.66)	0.503	(0.66)	0.173	(0.45)	0.491	(0.03)
	$n = 200$	0.012	(0.11)	0.151	(0.34)	0.125	(0.35)	0.048	(0.24)	0.294	(0.02)
$K = 1$	$n = 50$	0.047	(0.21)	1.813	(0.67)	1.597	(0.69)	0.199	(0.45)	0.754	(0.03)
	$n = 200$	0.023	(0.09)	2.471	(0.3)	2.382	(0.31)	0.048	(0.22)	0.733	(0.02)

For β and δ, we show the averages of the mean squared errors (MSEs) based on 50 replications, using the posterior means as point estimators. For π, we show the average of L_1 norm between the posterior mean and $\boldsymbol{\pi}^*$. The average posterior standard deviations are in parentheses.

The vertical axes of Figure 1 list the 15 consideration sets, with the true distribution of the consideration sets, $\boldsymbol{\pi}^*$, highlighted in red. Each panel of the figure displays the posterior mean (solid with dots, blue) along with the 95% credible intervals (dashed, blue), based on one realized data set. The first two panels illustrate that under the proposed approach ($K = \infty$), as the sample size n increases, the discrepancy between the posterior mean and the true distribution diminishes. In contrast, the two right panels show that when $K = 1$, even as n increases, the posterior does not adequately converge to the truth. This is because the model does not account for the true consideration dependence.

In the Supplementary Material, we conduct experiments with random effects. This case is not covered by Theorem 2 of our paper, but the proposed approach is found to deliver consistent estimates for both $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{D})$ and $\boldsymbol{\pi}$, where $\boldsymbol{D}$ is the covariance matrix of random effects, but the restrictive approach with $K = 1$ leads to larger MSEs/biases.

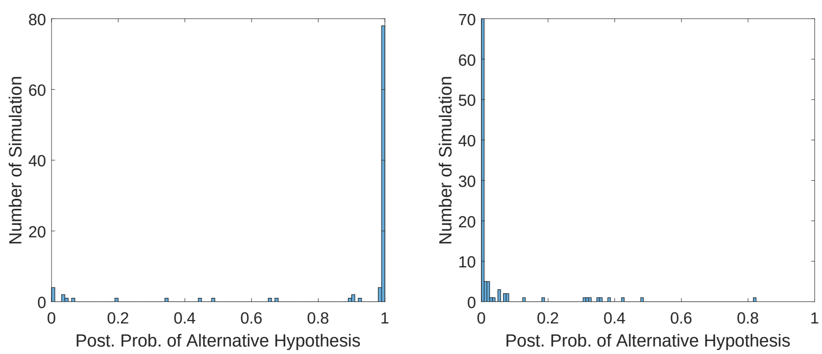


(a) $K = \infty, n = 50$ (b) $K = \infty, n = 200$ (c) $K = 1, n = 50$ (d) $K = 1, n = 200$

Figure 1: The true distribution over consideration sets (solid, red), posterior mean (solid with dots, blue), 95% equal-tailed credible interval (dashed, blue). Each plot is based on one realization of simulated data. $J = 4$, $T = 3$.

6.2 Testing for dependent consideration

From the MCMC output, it is possible to assess the degree of consideration dependence using the method proposed by Dunson and Xing (2009), but now applied to the latent consideration sets. The null hypothesis tests for independent consideration, formulated as $H_0 : \omega_1 = 1$. We utilize the interval null of $H_0 : \omega^* > 1 - \varepsilon$ with $\omega^* = \max\{\omega_k : k = 1, \ldots, k^*\}$ and $\varepsilon > 0$ is a small value. The Bayes factor in favor of the alternative hypothesis, $H_1 : \omega^* \leq 1 - \varepsilon$, is defined as $\frac{\Pr(H_1|\boldsymbol{D}^n)\Pr(H_1)}{\Pr(H_0|\boldsymbol{D}^n)\Pr(H_0)}$, which can be estimated using $\hat{\Pr}(H_1|\boldsymbol{D}^n)$, the portion of the posterior sample such that $\omega^* \leq 1 - \varepsilon$, and $\hat{\Pr}(H_0|\boldsymbol{D}^n) = 1 - \hat{\Pr}(H_1|\boldsymbol{D}^n)$. In the simulations and the application, $\underline{a}_\alpha = \underline{b}_\alpha = 1/4$ is fixed to produce $\Pr(H_0) \approx 0.5$.



(a) Case 1 (dependent consid.) (b) Case 2 (independent consid.)

Figure 2: Histograms of estimated posterior probabilities of H_1 in each of the 100 simulations under (a) case 1 (dependent consideration - H_1 is true) and (b) case 2 (independent consideration - H_0 is true). $\varepsilon = 0.1, n = 50$.

We use the current data-generating process as the first case (dependent consideration). In the second case, the consideration is independent. We generate $C_{ij} \overset{iid}{\sim} \text{Bernoulli}(\gamma_j)$, for $j = 1, 2, 3$ with $(\gamma_1, \gamma_2, \gamma_3) = (0.2, 0.15, 0.35)$ and fix $C_{i4} = 1$, for $i = 1, \ldots, n$. Figure 2 (a) provides a histogram showing the estimated posterior probabilities of H_1 under the first case (H_1 is true) across the 100 simulated data sets using $\varepsilon = 0.1$. The method appropriately assigns a value close to one to $\Pr(H_1|\boldsymbol{D}^n)$ in most cases, with only 8/100 having an estimated $\Pr(H_1|\boldsymbol{D}^n) < 0.5$. Figure 2 (b) provides the results for case 2. The posterior probability assigned to H_1 is close to zero for most simulations. We find similar results with random effects, as shown in the Supplementary Material.

7 Application to Cereal Consumption in Midwest

7.1 Data Description

In this section, we apply our approach to a manually constructed longitudinal data set that includes $J = 101$ cereal brands, which represents a significantly larger number than was feasible before. For comparison, J was 4 in Chiang et al. (1998), 10 in Van Nierop et al. (2010), and 5 in Aguiar and Kashaev (2024). We constructed the data set by integrating Nielsen Consumer Panel data with Retail Scanner Data, focusing on weekly shopping trips in 2019 in stores operated by a single anonymous retailer primarily based in the United States Midwest. Although data from 2020 are available, we chose to use the most recent pre-pandemic year to avoid potential biases introduced by pandemic-related shopping behavior. This particular retailer was selected because it consistently stocked more than 100 cereal brands throughout the sample period. Furthermore, we limited our analysis to a single retailer to prevent inconsistencies in brand definitions between different retailers, which would have required speculative alignment of brand names from various sources. The final data set includes $J = 101$ brands and $n = 1880$ households, covering 25,849 purchases in 239 stores during the 52-week period in 2019. See Figure 3, for the locations of these stores with relative purchase volumes. The average number of shopping trips per household (T_i) is 13.7, and the price P_{ijt} of each brand $j \in 1, \ldots, J$ is represented by a size-weighted price index constructed from prices at the UPC level. For the analysis, we used the first 10 months of data for the estimation and reserved the last two months for the prediction outside the sample. Further details on data preparation are provided in the Supplementary Material.

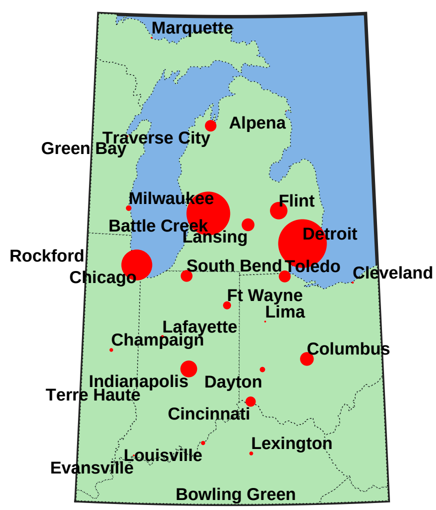


Figure 3: Locations of the 239 stores under the chosen retailer. Circle sizes correspond to purchases (percentages).

Conditional on the consideration set $\{\mathcal{C}_i\}$, in the most general version of the model, we enter the fixed effects and random effects in the MNL model $V_{ijt} = \delta_j + P_{ijt}(\beta + b_i)$, where $i \in \{1, \ldots, 1880\}$ indexes households, and $t \in \{1, \ldots, T_i\}$ indexes purchase occasions. In this model, δ_j represents the brand-specific fixed effect for brand j, with the normalization $\delta_J = 0$. The parameter β is the common fixed effect, and $b_i \sim \mathcal{N}(0, D)$ is the random effect for household i. We consider four variants of the MNL, differentiated by the inclusion of random effects and/or consideration set heterogeneity, as detailed in models (1)–(4) of Table 3. In addition, models (5) and (6) assume an independent consideration structure (i.e., $K = 1$). Each of these cases is estimated using the simulation method developed in Section 4, by omitting the components not present in the full hierarchical model (MNL_RC).

7.2 Empirical Results

We obtained 10,000 MCMC draws for each of the six models in Matlab on a desktop with a 4.9GHz processor and 64GB RAM. Broadly speaking, the estimated parameters of the response model from the approaches (1)-(4) shown in Table 3 are similar to those in the literature. For instance, when consideration set heterogeneity is incorporated, the magnitude of the slope parameter β on price increases and the number of significant brand-specific terms δ_j's decreases (See Table B1 for the list of estimated δ_j's), which is consistent with the previous studies (Stopher, 1980, Swait and Ben-Akiva, 1986, Chiang et al., 1998, Van Nierop et al., 2010). Moreover, when we control for consideration sets, the posterior mean of $D^{1/2}$ decreases, which is in accordance with the literature that finds that the amount of preference heterogeneity is overestimated if consumer's limited information is not accounted for (Chiang et al., 1998, Morozov et al., 2021). The confirmation of these findings under large J was made possible because of the proposed scalable approach.

Under the independent consideration assumption ($K = 1$), i.e., (5) and (6), the estimated parameters are similar to the proposed flexible approach i.e., (3) and (4) except that the estimated $D^{1/2}$ under (6) is slightly larger than (2), which contradicts with the previous studies. In general, it is possible that the obtained estimates under $K = 1$ are biased, as shown in simulation studies in Section 6. We conduct the test for independent considera-

Table 3: Estimation results

	(1) MNL		(2) MNL_R		(3) MNL_C		(4) MNL_RC		(5) MNL_C_K1		(6) MNL_RC_K1	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
Random effects on price												
β	-0.69***	0.02	-0.77***	0.03	-0.73***	0.02	-0.82***	0.04	-0.73***	0.02	-0.85***	0.04
$D^{1/2}$	——	——	0.99	0.02	——	——	0.96	0.04	——	——	1.07	0.04
Brand-specific fixed-effects												
# of "significant" params.	100		98		76		70		73		68	
Computational time												
min. per 1,000 MCMC iters.	64		67		122		125		121		122	
Test for indep. consid.	–		–		Reject H_0		Reject H_0		–		–	
random effects	No		Yes		No		Yes		No		Yes	
Consideration sets	No		No		$K=\infty$		$K=\infty$		$K=1$		$K=1$	

$K=\infty$ ($K=1$) refers to the proposed infinite mixture of independent consideration models (the model under the independent consideration). The first panel shows posterior means of the mean β of the random effects on price and the standard deviations $D^{1/2}$ with their posterior standard deviations. Three stars indicate that the corresponding 99% credible interval does not include 0. The second panel shows the number of brand-specific fixed effects whose 95% posterior credible intervals do not include 0 (out of 100 terms). See Table B1 for the estimated δ_j's under MNL_RC. The third panel shows computational time (minutes) on a desktop with a 4.9GHz processor and 64GB RAM. The last panel shows the results for the test for independent consideration with the null hypothesis H_0 : independent consideration. The results are based on 10,000 posterior draws.

tion studied in Section 6. Under both (3) and (4), the estimated posterior probability of the alternative hypothesis (dependent consideration) is very close to one, and we conclude that the considerations of cereal products in this particular market are dependent.

7.2.1 Computational time

Table 3 also shows the computational time per 1,000 MCMC draws. The extra burden of estimating latent consideration sets using our proposed approach is reasonable. For instance, when consideration sets are estimated along with random effects, the computational time roughly doubles (67 mins. for MNL_R and 125 mins. for MNL_RC). Not surprisingly, compared to the fully flexible estimator, the estimators that assume independent consideration take less computational time but only slightly.

7.2.2 Estimated parameters in the mixture model

We next investigate the clustering of subjects according to the proposed mixture model. The posterior mode of the number of nonempty clusters under the full-specification (MNL_RC) is six. The Supplementary Material shows further estimation results on the number of clusters and the DP concentration parameter α. To understand how households are clustered, we computed the posterior mean of the event that a given pair of households (i, i') are clustered together i.e. $\{S_i = S_{i'}\}$. This results in a $n \times n$ "similarity matrix," which can be found in the Supplementary Material.

An examination of how households are clustered reveals interesting points. Take household A as an example whose actual choices consist of $\{60, 69, 73\}$. Define an estimator $\hat{\mathcal{C}}_i$ of the consideration set for household i as the set of brands j whose posterior probability that $C_{ij} = 1$ is greater than 0.2658, the prior median of q_{hj}. This results in the estimated set $\hat{\mathcal{C}}_A = \{3, 60, 68, 69, 73, 79, 101\}$. The upper panel of Table 4 lists the three households with the highest posterior similarity to subject A. There are several observations.

Table 4: Households clustered with $i \in \{A, B\}$.

i'	Similarity	Chosen brands	Estimated $\hat{\mathcal{C}}_{i'}$
Household $i = A$			
$i' = A$	1.00	$\{60, 69, 73\}$	$\{3, 60, 68, 69, 73, 79, 101\}$
$i' = 1340$	0.60	$\{3, 68, 73\}$	$\{3, 25, 26, 59, 60, 61, 63, 66, 68, 73, 77, 79, 101\}$
$i' = 818$	0.59	$\{60, 68, 73\}$	$\{3, 25, 26, 59, 60, 61, 63, 66, 68, 73, 79, 101\}$
$i' = 1187$	0.59	$\{7, 39, 73\}$	$\{3, 7, 25, 26, 39, 59, 60, 61, 63, 66, 68, 73, 77, 79, 101\}$
Household $i = B$			
$i' = B$	1.00	$\{4, 7, 8, 17, 27, 53, 66, 68, 77, 79, 92, 101\}$	$\{3, 4, 7, 8, 17, 25, 27, 53, 59, 60, 63, 66, 68, 73, 77, 79, 92, 101\}$
$i' = 566$	0.97	$\{5, 7, 8, 26, 27, 39, 60, 61, 77, 101\}$	$\{3, 5, 7, 8, 26, 27, 39, 60, 61, 77, 101\}$
$i' = 31$	0.96	$\{26, 47, 51, 59, 60, 62, 63, 68, 72, 73, 92, 101\}$	$\{26, 47, 51, 59, 60, 62, 63, 68, 72, 73, 92, 101\}$
$i' = 50$	0.96	$\{3, 6, 7, 25, 59, 66, 68, 81, 101\}$	$\{3, 6, 7, 25, 59, 60, 66, 68, 81, 101\}$

Similarity is defined as the posterior mean of $1\{S_{i'} = S_i\}$, which corresponds to the ith row (equivalently ith column) of the similarity matrix presented in the Supplementary Material. The estimated consideration set $\hat{\mathcal{C}}_i$ is defined as the set of brands j whose posterior probability that $C_{ij} = 1$ is greater than 0.2658, the prior median of q_{hj}. The result is from the MNL_RC model.

First, the actual choices of the households tend to overlap within a cluster; all four subjects chose 73. Second, the estimated consideration sets $\hat{\mathcal{C}}_{i'}$ are similar between households in a cluster. For example, household A did not choose brands 3 and 68, but other households did, and they are in $\hat{\mathcal{C}}_A$. Third, the stronger the purchase overlap, the higher the chance of being in the same cluster. The lower panel of Table 4 shows the results for household B. All of the four households in the cluster purchased at least three from the brands 7, 8, 27, 66, 68, 77, 92, and 101, and show higher similarity scores (≥ 0.96). In this way, our algorithm discovers the probabilistic grouping patterns in the choice data.

7.2.3 Predictive performance

We examine the predictive performance of the model using the last 2 months of the observations that we keep as an out-of-sample period. Denote the set of subjects who made purchases during the out-of-sample period by $\mathcal{O} \subset \{1, \dots, n\}$. There are 1079 households in $\mathcal{O}$. For each $i \in \mathcal{O}$, we predict $\boldsymbol{Y}_i^f = \{Y_{iT_i+s} : s = 1, \dots, h_i\}$, given the newly available covariates $\boldsymbol{\omega}_i^f = \{\boldsymbol{\omega}_{iT_i+s} : s = 1, \dots, h_i\}$, where h_i denotes the forecast horizon for the subject i. Let $\boldsymbol{y}_i^f = \{y_{iT_i+s} : s = 1, \dots, h_i\}$ be the actual set of responses for the subject $i \in \mathcal{O}$. The predictive likelihood for this subject is

$$
\begin{aligned}
p(\boldsymbol{y}_i^f | \boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{\omega}_i^f) &= \int \prod_{s=1}^{h_i} \Pr(Y_{iT_i+s} = y_{iT_i+s} | \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{iT_i+s}, \mathcal{C}_i) d\pi(\boldsymbol{\delta}, \boldsymbol{\beta}, \{\boldsymbol{b}_i\}, \{\mathcal{C}_i\} | \boldsymbol{y}, \boldsymbol{\omega}) \\
&\approx \frac{1}{G} \sum_{g=1}^{G} \prod_{s=1}^{h_i} \Pr(Y_{iT_i+s} = y_{iT_i+s} | \boldsymbol{\delta}^{(g)}, \boldsymbol{\beta}^{(g)}, \boldsymbol{b}_i^{(g)}, \boldsymbol{\omega}_{iT_i+s}, \mathcal{C}_i^{(g)}),
\end{aligned}
$$

where the response probability conditional on a consideration set is given in (5). Figure 4 gives the log-predictive likelihood for each household under the (MNL_R) and (MNL_RC) models. This figure shows that including the consideration set heterogeneity generally improves the predictions. Importantly, this is perhaps the first such predictive comparison in the large J case, made possible by the scalable procedure developed in this paper. See Supplementary Material for more discussion on the prediction.

In this empirical section, we demonstrated that the added computational effort of estimating latent consideration sets under the proposed approach is reasonable. The estimated response model parameters and the improved prediction when consideration set heterogeneity is taken into account are consistent with the previous studies, but their confirmation

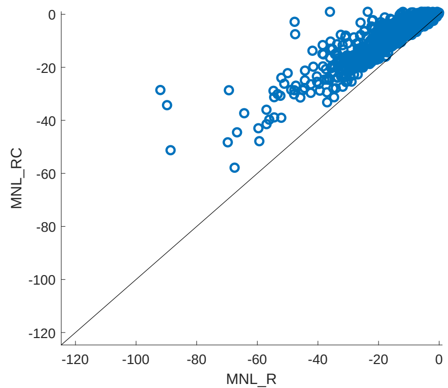


Figure 4: Log-predictive likelihoods (circles) for the 1079 households that made purchases in the out-of-sample period. The x-coordinate of each circle is the log-predictive likelihood under MNL_R, and the y-coordinate is under MNL_RC. The he 45-degree line is plotted as a solid line.

was made possible under the large J because of the proposed scalable approach.

8 Conclusion

In this article, we proposed a scalable modeling and estimation scheme for multinomial response models with uncertain consideration sets. The approach relies on a factor decomposition technique to flexibly model the distribution over the latent consideration sets. We showed that our approach can be applied to situations beyond the reach of existing methods, such as an application with $J = 101$ brands.

As described in the introduction, the latent consideration set models have been applied in many fields such as economics, marketing, psychology, and transportation science, but in the absence of restrictive assumptions such as independence of considerations, its application has been limited to data sets with small numbers of alternatives due to the curse of dimensionality. The methodology proposed in this paper thus provides new opportunities for empirical researchers in multiple disciplines to estimate large-scale multinomial response models with latent consideration sets which would otherwise be infeasible.

A Proof of Lemma 2

Proof of Lemma 2. Recall that $p_{\boldsymbol{\theta},\boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{w}) \equiv \sum_{c\in\mathcal{C}} \pi_c \Pr(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{w}, c)$. For any $\boldsymbol{y} \in \mathcal{J}^T$, if $p_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w}) = 0$, the integrand in the KL divergence is $\log(0)0$ which is defined to be zero. Therefore, without loss of generality, suppose that for all $\boldsymbol{y}$, there is $c_y \in \mathcal{C}$ that contains all the elements of $\boldsymbol{y}$ and have $\pi^*_{c_y} > 0$. By Lemma 1, we can find a finite mixture of independent consideration models that exactly matches the true distribution of consideration sets; i.e. $\exists(K, \tilde{\boldsymbol{\phi}}_{1:K})$ such that $\pi^*_c = \sum_{h=1}^K \tilde{\omega}_h \prod_{j\in c} \tilde{q}_{hj} \prod_{j\notin c}(1 - \tilde{q}_{hj})$ for each $c \in \mathcal{C}$. Hence we have

$$
\begin{aligned}
&\int \sum_{\boldsymbol{y}} \log\left(\frac{p_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}, K, \boldsymbol{\phi}_{1:K})}\right) p_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w}) g^*(\boldsymbol{w}) d\boldsymbol{w} \\
&= \int \sum_{\boldsymbol{y}} \left\{ \log\left(\frac{p_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w})}{p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}^*, K, \tilde{\boldsymbol{\phi}}_{1:K})}\right) + \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}^*, K, \tilde{\boldsymbol{\phi}}_{1:K})}{p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}, K, \boldsymbol{\phi}_{1:K})}\right) \right\} p_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w}) g^*(\boldsymbol{w}) d\boldsymbol{w},
\end{aligned}
$$

and the first term in the brackets is zero. Hence, it suffices to show that the integral of the second term is continuous in $(\boldsymbol{\theta}, \boldsymbol{\phi}_{1:K})$ at $(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\phi}}_{1:K})$. In the Supplementary Material, we prove that the response probability is continuous in $\boldsymbol{\phi}_{1:K}$ (Lemma SB1) and it is continuous also in $\boldsymbol{\theta}$ (Lemma SB2). Let $(\boldsymbol{\theta}^m, \boldsymbol{\phi}^m_{1:K})$ be a sequence of parameter values converging to $(\boldsymbol{\theta}^*, \tilde{\boldsymbol{\phi}}_{1:K})$. Then

$$
\lim_{m\to\infty} \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}^*, K, \tilde{\boldsymbol{\phi}}_{1:K})}{p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}^m, K, \boldsymbol{\phi}^m_{1K})}\right) = 0.
$$

The result will follow from the dominated convergence theorem if there is an integrable (with respect to $p_{\boldsymbol{\theta}^*,\boldsymbol{\pi}^*}(\boldsymbol{y}|\boldsymbol{w}) g^*(\boldsymbol{w})$) upper bound of $|\log p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}^m, K, \boldsymbol{\phi}^m_{1:K})|$. Note that

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}^m, K, \boldsymbol{\phi}^m_{1:K}) &= \sum_{c\in\mathcal{C}} \pi(c|K, \boldsymbol{\phi}^m_{1K}) \Pr(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}^m, \boldsymbol{w}, c) \\
&\geq \pi(c_y|K, \boldsymbol{\phi}^m_{1K}) \Pr(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}^m, \boldsymbol{w}, c_y),
\end{aligned}
$$

where $\pi(c_y|K,\boldsymbol{\phi}^m_{1K}) = \sum_{h=1}^K \omega_h^m \prod_{\ell\in c_y} q^m_{h\ell} \prod_{\ell\notin c_y}(1-q^m_{h\ell})$. First, since $\boldsymbol{\phi}^m_{1K} \to \tilde{\boldsymbol{\phi}}_{1:K}$ and $\pi^*_{c_y} = \sum_{h=1}^K \tilde{\omega}_h \prod_{\ell\in c_y} \tilde{q}_{h\ell} \prod_{\ell\notin c_y}(1-\tilde{q}_{h\ell}) > 0$, the first term is bounded below by some $\ell_1(\boldsymbol{y}) > 0$ for sufficiently large m. Second, since $\boldsymbol{\theta}^m \to \boldsymbol{\theta}^*$ and $\Pr(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}^*, \boldsymbol{w}, c_y) > 0$ (as $\boldsymbol{\beta}^*$ is in a compact set, $\boldsymbol{D}^*$ is positive definite, and $\mathcal{W}$ is compact), $\Pr(\boldsymbol{Y}_i = \boldsymbol{y}|\boldsymbol{\theta}^m, \boldsymbol{w}, c_y)$ is bounded below by some $\ell_2(\boldsymbol{y}, \boldsymbol{w}) > 0$ for sufficiently large m. Finally, $1 \geq p(\boldsymbol{y}|\boldsymbol{w};\boldsymbol{\theta}^m, K, \boldsymbol{\phi}^m_{1:K}) \geq \inf_{\boldsymbol{w}\in\mathcal{W}} \min_{\boldsymbol{y}\in\mathcal{J}^T} \ell_1(\boldsymbol{y})\ell_2(\boldsymbol{y},\boldsymbol{w}) > 0$, for all $(\boldsymbol{y},\boldsymbol{w}) \in \mathcal{J}^T \times \mathcal{W}$. $\square$

B Additional table for the application

Table B1: Estimates of the 100 brand fixed-effects from the real data on cereal market

	brand	mean	s.d		brand	mean	s.d
1	BEAR NAKED FIT GRN	-0.19	0.29	51	KELLOGGS FROOT LOOPS	-0.37*	0.08
2	BEAR NAKED GRN	0.37	0.18	52	KELLOGGS FROOT LOOPS MARSHMALLOW	-1.5*	0.17
3	BETTER OATS	-0.93*	0.22	53	KELLOGGS FROSTED FLAKES	0.09	0.06
4	CREAM OF WHEAT	0.14	0.1	54	KELLOGGS FROSTED MINIWHEATS	0.51*	0.06
5	CTL BR	-0.44*	0.06	55	KELLOGGS FROSTED MINIWHT LTTLE BTS	-0.52*	0.08
6	GENERAL MILLS APPLE CINNAMON CHEERIOS	-1.22*	0.14	56	KELLOGGS KRAVE	0.54*	0.08
7	GENERAL MILLS BLUEBERRY CHEX	-0.38*	0.14	57	KELLOGGS RAISIN BRAN	-0.12	0.07
8	GENERAL MILLS BREAKFAST PACK	-2.09*	0.52	58	KELLOGGS RAISIN BRAN CRUNCH	-0.02	0.07
9	GENERAL MILLS CHEERIOS	0.1	0.06	59	KELLOGGS RICE KRISPIES	-0.42*	0.07
10	GENERAL MILLS CHEERIOS OAT CRUNCH CNMN	-0.48*	0.11	60	KELLOGGS RICE KRISPIES TREATS	-0.1	0.32
11	GENERAL MILLS CHOCOLATE CHEERIOS	-1.7*	0.22	61	KELLOGGS SPECIAL K	-0.59*	0.17
12	GENERAL MILLS CHOCOLATE CHEX	-0.55*	0.13	62	KELLOGGS SPECIAL K CHOCOLATY DELGHT	0.09	0.11
13	GENERAL MILLS CHOCOLATE PNUT BTR CHEERIO	-1.13*	0.18	63	KELLOGGS SPECIAL K CINNAMON PECAN	-0.71*	0.17
14	GENERAL MILLS CINNAMON CHEX	-1.25*	0.18	64	KELLOGGS SPECIAL K FRUIT & YOGURT	0.14	0.11
15	GENERAL MILLS CINNAMON TOAST CRUNCH	-0.04	0.06	65	KELLOGGS SPECIAL K PROTEIN	-0.24*	0.11
16	GENERAL MILLS CINNAMON TOAST CRUNCH CHRS	-1.63*	0.19	66	KELLOGGS SPECIAL K RED BERRY	0.14	0.08
17	GENERAL MILLS COCOA PUFFS	-0.66*	0.09	67	KELLOGGS SPECIAL K VANILLA ALMOND	-0.05	0.12
18	GENERAL MILLS COOKIECRISP	-1.08*	0.16	68	KELLOGGS STBY KRISPIES US OLYMPC TM	-1.63*	0.19
19	GENERAL MILLS CORN CHEX	-0.47*	0.11	69	MOM BERRY COLOSSAL CRN	-0.87*	0.34
20	GENERAL MILLS FIBER ONE	-0.07	0.22	70	MOM CINNAMON TOASTERS	-0.17	0.28
21	GENERAL MILLS FIBER ONE HONEY CLUSTERS	0.31	0.2	71	MOM COCOA DYNOBITES	-0.33	0.28
22	GENERAL MILLS FROSTED CHEERIOS	-1.88*	0.31	72	MOM FROSTED FLAKES	-0.34	0.31
23	GENERAL MILLS GOLDEN GRAHAMS	-0.17*	0.08	73	MOM FROSTED MINI SPOONERS	0.64*	0.27
24	GENERAL MILLS HONEY NUT CHEERIOS	0.16*	0.05	74	MOM FRUITY DYNOBITES	-0.13	0.27
25	GENERAL MILLS HONEY NUT CHEX	-1.06*	0.17	75	MOM GOLDEN PUFFS	0.27	0.18
26	GENERAL MILLS LUCKY CHARMS	0.12*	0.06	76	MOM TOOTIE FRUITIES	-0.4	0.26
27	GENERAL MILLS MPL CHEERIOS CLC DSS FNDTN	-0.74*	0.11	77	POST COCOA PEBBLES	-0.54*	0.15
28	GENERAL MILLS MULTIGRAIN CHEERIOS	-0.01	0.07	78	POST FRUITY PEBBLES	-0.36*	0.09
29	GENERAL MILLS NATURE VALLEY GRN PROTEIN	-0.54	0.36	79	POST GOLDEN CRISP	-1.45*	0.19
30	GENERAL MILLS RAISIN NUT BRAN	0.15	0.16	80	POST GRAPENUTS	-0.34	0.21
31	GENERAL MILLS REESE'S PUFFS	0.21*	0.07	81	POST GRAPENUTS FLAKES	0.16	0.29
32	GENERAL MILLS RICE CHEX	-0.44*	0.09	82	POST HONEY BUNCHES OF OATS	0.27*	0.07
33	GENERAL MILLS VANILLA CHEX	-0.85*	0.16	83	POST HONEY BUNCHES OF OATS GRN	-2.75*	0.76
34	GENERAL MILLS VERY BERRY CHEERIOS	-0.99*	0.15	84	POST HONEYCOMB	-1.08*	0.14
35	GENERAL MILLS WHEAT CHEX	-0.67*	0.16	85	POST OREO OS	-1.8*	0.36
36	GENERAL MILLS WHEATIES	0.06	0.16	86	POST RAISIN BRAN	-0.64*	0.22
37	KASHI CINNAMON HARVEST	-0.69*	0.3	87	POST SELECTS GREAT GRAINS	-0.23*	0.12
38	KASHI GO LEAN	-0.98*	0.16	88	POST SHRD WHT 'N BRN SP SZ	0.35*	0.16
39	KASHI GO LEAN CRUNCH!	-1.5*	0.28	89	POST SHREDDED WHEAT	-1.12*	0.36
40	KASHI ORGANIC BLUEBERRY CLST	-1.56*	0.28	90	QUAKER	-0.15*	0.05
41	KELLOGGS AL JS CN PS FRFL FTLP CKSP	-3.28*	0.49	91	QUAKER CAP'N CRN	-0.87*	0.13
42	KELLOGGS ALLBRAN	-0.64*	0.31	92	QUAKER CAP'N CRN CRN BRY	-0.98*	0.12
43	KELLOGGS ALLBRAN COMPLETE WHT FLK	0.29	0.48	93	QUAKER CINNAMON LIFE	-0.58*	0.09
44	KELLOGGS APPLE JACKS	-0.35*	0.09	94	QUAKER GRN	-1.38*	0.67
45	KELLOGGS CHOCOLT FRS FLKS TN TH TGR	-1.81*	0.25	95	QUAKER LIFE	-0.69*	0.1
46	KELLOGGS COCOA KRISPIES	-0.75*	0.12	96	QUAKER OATMEAL SQUARES	-0.2*	0.11
47	KELLOGGS CORN FLAKES	-0.47*	0.1	97	QUAKER OVERNIGHT OATS	-2.5*	0.25
48	KELLOGGS CORN POPS	-0.61*	0.11	98	QUAKER PROTEIN	-0.61*	0.15
49	KELLOGGS CRACKLIN' OAT BRAN	0.33	0.24	99	QUAKER REAL MEDLEYS	-1.2*	0.21
50	KELLOGGS CRISPIX	-0.18	0.1	100	QUAKER SELECT STARTS	-0.98*	0.14

The stars indicate that the corresponding 95% credible interval does not include 0. The "other" option -specific fixed-effect is normalized to 0 for identification. The results are obtained under the MNL_RC model.

References

J. Abaluck and A. Adams-Prassl. What do consumers consider before they choose? *The Quarterly Journal of Economics*, 136(3):1611–1663, 2021.

N. Agarwal and P. J. Somaini. Demand analysis under latent choice constraints. Technical report, National Bureau of Economic Research, 2022.

V. H. Aguiar and N. Kashaev. Identification and estimation of discrete choice models with unobserved choice sets. *Journal of Business & Economic Statistics*, pages 1–25, 2024.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

T. Amano, A. Rhodes, and S. Seiler. *Large-scale demand estimation with search data.* Harvard Business School, 2018.

L. Barseghyan, M. Coughlin, F. Molinari, and J. C. Teitelbaum. Heterogeneous choice sets and preferences. *Econometrica*, 89(5):2015–2048, 2021a.

L. Barseghyan, F. Molinari, and M. Thirkettle. Discrete choice under risk with limited consideration. *American Economic Review*, 111(6):1972–2006, 2021b.

M. Ben-Akiva and B. Boccara. Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1):9–24, 1995.

M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

B. Bronnenberg and W. Vanhonacker. Limited choice sets, local price response, and implied measures of price competition. *J. of Marketing Research*, 33(2):163–173, 1996.

Y. Cao, A. Zhang, and H. Li. Multisample estimation of bacterial composition matrices in metagenomics data. *Biometrika*, 107(1):75–92, 2020.

M. D. Cattaneo, X. Ma, Y. Masatlioglu, and E. Suleymanov. A random attention model. *Journal of Political Economy*, 128(7):2796–2836, 2020.

E. Z. Chen and H. Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617, 2016.

J. Chiang, S. Chib, and C. Narasimhan. Markov chain monte carlo and models of consideration set and parameter heterogeneity. *J. of Econometrics*, 89(1):223–248, 1998.

S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2): 347–361, 1998.

A. Ching, T. Erdem, and M. Keane. The price consideration model of brand choice. *Journal of Applied Econometrics*, 24(3):393–420, 2009.

A. T. Ching, T. Erdem, and M. P. Keane. A simple method to estimate the roles of learning, inventories and category consideration in consumer choice. *Journal of choice modelling*, 13:60–72, 2014.

K. X. Chiong and M. Shum. Random projection estimation of discrete-choice models with large choice sets. *Management Science*, 65(1):256–271, 2019.

G. S. Crawford, R. Griffith, and A. Iaria. A survey of preference estimation with unobserved choice set heterogeneity. *J. of Econometrics*, 222(1):4–43, 2021.

V. Dardanoni, P. Manzini, M. Mariotti, and C. J. Tyson. Inferring cognitive heterogeneity from aggregate choices. *Econometrica*, 88(3):1269–1296, 2020.

B. De los Santos. Consumer search on the internet. *International Journal of Industrial Organization*, 58:66–105, 2018.

L. Devroye, A. Mehrabian, and T. Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

M. Draganska and D. Klapper. Choice set heterogeneity and the role of advertising: An analysis with micro and macro data. *J. of Marketing Research*, 48(4):653–669, 2011.

D. B. Dunson and C. Xing. Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

D. Ershov, J.-W. Laliberté, M. Marcoux, and S. Orr. Estimating complementarity with large choice sets: An application to mergers. *RAND J. of Economics (accepted)*, 2024.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

S. Ghosal and A. W. van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

M. S. Goeree. Limited information and advertising in the us personal computer industry. *Econometrica*, 76(5):1017–1074, 2008.

E. Honka, A. Hortaçsu, and M. Wildenbeest. Empirical search and consideration sets. In *Handbook of the Economics of Marketing*, volume 1, pages 193–257. Elsevier, 2019.

A. Iaria and A. Wang. An empirical model of quantity discounts with large choice sets. *Available at SSRN 3946475*, 2024.

Z. Jiang, J. Li, and D. Zhang. A high-dimensional choice model for online retailing. *Management Science*, 2024.

N. Kashaev, N. Lazzati, and R. Xiao. Peer effects in random consideration sets. *arXiv preprint arXiv:1904.06742*, 2019.

K. Kawaguchi, K. Uetake, and Y. Watanabe. Designing context-based marketing: Product recommendations under time pressure. *Management Science*, 2021.

C. Kwon and E. Mbakop. Estimation of the number of components of nonparametric multivariate finite mixture models. *The Annals of Statistics*, 49(4):2178–2205, 2021.

T. Liu, H. Zhao, and T. Wang. An empirical bayes approach to normalization and differential abundance testing for microbiome data. *BMC bioinformatics*, 21:1–18, 2020.

R. Loaiza-Maya and D. Nibbering. Scalable bayesian estimation in the multinomial probit model. *Journal of Business & Economic Statistics*, 40(4):1678–1690, 2022.

Z. Lu. Estimating multinomial choice models with unobserved choice sets. *J. of Econometrics*, 226(2):368–398, 2022.

C. F. Manski. The structure of random utility models. *Theory and decision*, 8(3):229, 1977.

P. Manzini and M. Mariotti. Stochastic choice and consideration sets. *Econometrica*, 82 (3):1153–1176, 2014.

J.-A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2015.

Y. Masatlioglu, D. Nakajima, and E. Y. Ozbay. Revealed attention. *American Economic Review*, 102(5):2183–2205, 2012.

I. Morozov, S. Seiler, X. Dong, and L. Hou. Estimation of preference heterogeneity in markets with costly search. *Marketing Science*, 40(5):871–899, 2021.

A. Norets and K. Shimizu. Semiparametric bayesian estimation of dynamic discrete choice models. *J. of Econometrics*, 238(2):105642, 2024.

R. Paleti, S. Mishra, K. Haque, and M. M. Golias. Latent class analysis of residential and work location choices. *Transportation Letters*, 13(10):696–706, 2021.

J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–1202, 2013.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

A. N. Smith and G. M. Allenby. Demand models with random partitions. *Journal of the American Statistical Association*, 2019.

P. R. Stopher. Captivity and choice in travel-behavior models. *Transportation engineering journal of ASCE*, 106(4):427–435, 1980.

J. Swait and M. Ben-Akiva. Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological*, 21(2):91–102, 1987.

J. D. Swait and M. Ben-Akiva. Analysis of the effects of captivity on travel time and cost elasticities. In *Behavioural Research for Transport Policy*, 1986.

Z.-Z. Tang and G. Chen. Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713, 2019.

F. Traets, M. Meulders, and M. Vandebroek. Modelling consideration heterogeneity in a two-stage conjunctive model. *Journal of Mathematical Psychology*, 109:102687, 2022.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

E. Van Nierop, B. Bronnenberg, R. Paap, M. Wedel, and P. Franses. Retrieving unobserved consideration sets from household panel data. *J.ofMarketingResearch*, 47(1):63–74, 2010.

S. G. Walker. Sampling the dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54, 2007.

Y. Zeng, D. Pang, H. Zhao, and T. Wang. A zero-inflated logistic normal multinomial model for extracting microbial compositions. *Journal of the American Statistical Association*, 118(544):2356–2369, 2023.

Acknowledgement

We benefited from useful comments from Paola Manzini, Francesca Molinari, Nail Kashaev, Alessandro Iaria, Ao Wang, Elena Manresa, Dimitris Korobilis, Martin Weidner, Laura Liu, Asim Ansari, Andriy Norets, Toru Kitagawa, Frank Schorfheide, Zhentong Lu, Victor Aguiar, Yao Luo, Victor Aguirregabiria, Andrew Ching, Kosuke Uetake, Matthijs Wildenbeest, Mitsuru Igami, and the audience at the 2023 NBER-NSF Seminar in Bayesian Econometrics and Statistics (SBIES) at the Federal Reserve Bank of Philadelphia, the 2023 INFORMS Annual Meeting in Phoenix, and the 2024 International Industrial Organization Conference in Boston.

SUPPLEMENTARY MATERIAL

Section SA provides intermediate results used to prove Theorems 1 and 2 of the main paper, and their proofs. Section SB establishes a "sparsity property" of the proposed Metropolis-Hastings algorithm. Section SC presents the conditional posterior distributions of the parameters other than the consideration sets. Section SD illustrates the impact of the prior choice for the attention probabilities on the prior on the distribution of consideration sets. Section SE shows additional simulation results under $J = 4$. Section SF presents a simulation study under a large number of alternatives, $J = 100$. Section SG provides additional results from the empirical application.

SA Intermediate theoretical results and proofs

SA.1 Intermediate results in the proof of Lemma 2

The following two lemmas are used to prove Lemma 2. They state that the marginal response probability is continuous with respect to the mixture parameters as well as the parameters in the response model. We prove the intermediate results for the case of $T = 1$. The extensions to the $T > 1$ case can be done similarly but at the expense of proof simplicity.

Lemma SA.1 (Continuity of response probabilities wrt mixture parameters)**.** *Let* $\boldsymbol{\theta}$ *and* $\boldsymbol{w} \in \mathcal{W}$*. Then for each* $j \in \mathcal{J}$*,* $\forall \varepsilon > 0$ *and* $\boldsymbol{\phi}^{(1)}_{1:K}$*,* $\exists \delta > 0$ *such that for any* $\boldsymbol{\phi}^{(2)}_{1:K}$ *satisfying* $\sum_{j=1}^{J} |q^{(1)}_{hj} - q^{(2)}_{hj}| < \delta$ *and* $|\omega^{(1)}_h - \omega^{(2)}_h| < \delta$*, for* $h = 1, \ldots, K$*, we have*

$$\left| p(j|\boldsymbol{w}; \boldsymbol{\theta}, K, \boldsymbol{\phi}^{(1)}_{1:K}) - p(j|\boldsymbol{w}; \boldsymbol{\theta}, K, \boldsymbol{\phi}^{(2)}_{1:K}) \right| < \varepsilon.$$

Proof of Lemma SA.1. We have

$$\left|p(j|\boldsymbol{w};\boldsymbol{\theta},K,\boldsymbol{\phi}^{(1)}_{1:K}) - p(j|\boldsymbol{w};\boldsymbol{\theta},K,\boldsymbol{\phi}^{(2)}_{1:K})\right| \leq \sum_{c\in\mathcal{C}} \left|\pi(c|K,\boldsymbol{\phi}^{(1)}_{1:K}) - \pi(c|K,\boldsymbol{\phi}^{(2)}_{1:K})\right| \Pr(Y_{it}=j|\boldsymbol{\theta},\boldsymbol{w}_t,c)$$

where $\Pr(Y_{it}=j|\boldsymbol{\theta},\boldsymbol{w}_t,c) \leq 1$. The term in the absolute value is

$$\begin{aligned}
&\left|\sum_{h=1}^{K}\omega_h^{(1)}\prod_{j\in c} q_{hj}^{(1)}\prod_{j\notin c}(1-q_{hj}^{(1)}) - \sum_{h=1}^{K}\omega_h^{(2)}\prod_{j\in c} q_{hj}^{(2)}\prod_{j\notin c}(1-q_{hj}^{(2)}) \pm \sum_{h=1}^{K}\omega_h^{(1)}\prod_{j\in c} q_{hj}^{(2)}\prod_{j\notin c}(1-q_{hj}^{(2)})\right| \\
&\leq \sum_{h=1}^{K}\omega_h^{(1)}\left|\underbrace{\prod_{j\in c} q_{hj}^{(1)}\prod_{j\notin c}(1-q_{hj}^{(1)}) - \prod_{j\in c} q_{hj}^{(2)}\prod_{j\notin c}(1-q_{hj}^{(2)})}_{I}\right| + \sum_{h=1}^{K}\left|\omega_h^{(1)}-\omega_h^{(2)}\right|
\end{aligned}$$

The term I equals to

$$\begin{aligned}
&\prod_{j\in c} q_{hj}^{(1)}\prod_{j\notin c}(1-q_{hj}^{(1)}) - \prod_{j\in c} q_{hj}^{(2)}\prod_{j\notin c}(1-q_{hj}^{(2)}) \pm \prod_{j\in c} q_{hj}^{(2)}\prod_{j\notin c}(1-q_{hj}^{(1)}) \\
&= \left(\prod_{j\in c} q_{hj}^{(1)} - \prod_{j\in c} q_{hj}^{(2)}\right)\prod_{j\notin c}(1-q_{hj}^{(1)}) + \prod_{j\in c} q_{hj}^{(2)}\left(\prod_{j\notin c}(1-q_{hj}^{(1)}) - \prod_{j\notin c}(1-q_{hj}^{(2)})\right),
\end{aligned}$$

and hence the absolute value of I is bounded by the sum of the two terms: $\left|\prod_{j\in c} q_{hj}^{(1)} - \prod_{j\in c} q_{hj}^{(2)}\right|$ and $\left|\prod_{j\notin c}(1-q_{hj}^{(1)}) - \prod_{j\notin c}(1-q_{hj}^{(2)})\right|$. It is easy to show that the former is bounded by $c_1\sum_{j\in c}|q_{hj}^{(1)}-q_{hj}^{(2)}|$ and the latter is bounded by $c_2\sum_{j\notin c}|q_{hj}^{(1)}-q_{hj}^{(2)}|$ for some $c_1, c_2 > 0$. So, $|I| \leq c_3\sum_{j=1}^{J}|q_{hj}^{(1)}-q_{hj}^{(2)}|$ for some $c_3 > 0$. □

Lemma SA.2 (Continuity of response probabilities wrt θ)**.** *Suppose $\mathcal{W}$ is compact. Let $(K,\boldsymbol{\phi}_{1:K})$ and $\boldsymbol{w}\in\mathcal{W}$. Then for each $j\in\mathcal{J}$, $\forall\varepsilon>0$ and $\boldsymbol{\theta}^{(1)}=\{\boldsymbol{\beta}^{(1)},\boldsymbol{D}^{(1)}\}$, $\exists\delta>0$ such that for any $\boldsymbol{\theta}^{(2)}=\{\boldsymbol{\beta}^{(2)},\boldsymbol{D}^{(2)}\}$ satisfying $||\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}^{(2)}||<\delta$ and*

$$\sqrt{tr(\boldsymbol{D}^{(1)^{-1}}\boldsymbol{D}^{(2)} - \boldsymbol{I}) - \log\det(\boldsymbol{D}^{(1)}\boldsymbol{D}^{(2)^{-1}})} < \delta,$$

$$\left| p(j|\boldsymbol{w};\boldsymbol{\theta}^{(1)}, K, \boldsymbol{\phi}_{1:K}) - p(j|\boldsymbol{w};\boldsymbol{\theta}^{(2)}, K, \boldsymbol{\phi}_{1:K}) \right| < \varepsilon.$$

Proof of Lemma SA.2. Recall that for $j \in c$,

$$Pr(Y = j|\boldsymbol{\theta}, \boldsymbol{w}, c) = \int k_j(\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{b})\phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D})d\boldsymbol{b},$$

where we introduced the shorthand notation for the kernel

$$k_j(\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{b}) = \frac{e^{\boldsymbol{x}_j'\boldsymbol{\beta} + \boldsymbol{z}_j'\boldsymbol{b}}}{\sum_{\ell \in c} e^{\boldsymbol{x}_\ell'\boldsymbol{\beta} + \boldsymbol{z}_\ell'\boldsymbol{b}}},$$

where we suppressed the subscripts with respect to the units i for simplicity of notation.

We have

$$\begin{aligned}
\left| p(j|\boldsymbol{w};\boldsymbol{\theta}^{(1)}, K, \boldsymbol{\phi}_{1:K}) - p(j|\boldsymbol{w};\boldsymbol{\theta}^{(2)}, K, \boldsymbol{\phi}_{1:K}) \right| &\leq \sum_{c \in \mathcal{C}} \pi(c|K, \boldsymbol{\phi}_{1:K}) \left| \Pr(j|\boldsymbol{\theta}^{(1)}, \boldsymbol{w}, c) - \Pr(j|\boldsymbol{\theta}^{(2)}, \boldsymbol{w}, c) \right| \\
&= \sum_{c : j \in c} \pi(c|K, \boldsymbol{\phi}_{1:K}) \left| \Pr(j|\boldsymbol{\theta}^{(1)}, \boldsymbol{w}, c) - \Pr(j|\boldsymbol{\theta}^{(2)}, \boldsymbol{w}, c) \right|,
\end{aligned}$$

where if there is no $c \in \mathcal{C}$ such that $j \in c$ and $\pi(c|K, \boldsymbol{\phi}_{1:K}) > 0$, the claim is trivially true.

Now,

$$\left| \Pr(j|\boldsymbol{\theta}^{(1)}, \boldsymbol{w}, c)) - \Pr(j|\boldsymbol{\theta}^{(2)}, \boldsymbol{w}, c) \right| \leq \left| \Pr(j|\{\boldsymbol{\beta}^{(1)}, \boldsymbol{D}^{(1)}\}, \boldsymbol{w}, c) - \Pr(j|\{\boldsymbol{\beta}^{(2)}, \boldsymbol{D}^{(1)}\}, \boldsymbol{w}, c) \right| \tag{SA.1}$$

$$+ \left| \Pr(j|\{\boldsymbol{\beta}^{(2)}, \boldsymbol{D}^{(1)}\}, \boldsymbol{w}, c) - \Pr(j|\{\boldsymbol{\beta}^{(2)}, \boldsymbol{D}^{(2)}\}, \boldsymbol{w}, c) \right|. \tag{SA.2}$$

To bound (SA.1), note that for any $\rho > 0$, one can find $M_\rho > 0$ such that $\int 1\{||\boldsymbol{b}|| > M_\rho\}\phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(1)})d\boldsymbol{b} < \rho$. The term (SA.1) equals to

$$
\begin{aligned}
&\left|\int \left(k_j(\boldsymbol{w}, \boldsymbol{\beta}^{(1)}, \boldsymbol{b}) - k_j(\boldsymbol{w}, \boldsymbol{\beta}^{(2)}, \boldsymbol{b})\right)\phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(1)})d\boldsymbol{b}\right| \\
&\leq \int_{||\boldsymbol{b}|| \leq M_\rho} \left|k_j(\boldsymbol{w}, \boldsymbol{\beta}^{(1)}, \boldsymbol{b}) - k_j(\boldsymbol{w}, \boldsymbol{\beta}^{(2)}, \boldsymbol{b})\right| \phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(1)})d\boldsymbol{b} \\
&+ \int_{||\boldsymbol{b}|| > M_\rho} \left|k_j(\boldsymbol{w}, \boldsymbol{\beta}^{(1)}, \boldsymbol{b}) - k_j(\boldsymbol{w}, \boldsymbol{\beta}^{(2)}, \boldsymbol{b})\right| \phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(1)})d\boldsymbol{b}.
\end{aligned}
$$

Since $k_j(\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{b})$ has a bounded first derivative with respect to $\boldsymbol{\beta}$ for $||\boldsymbol{b}|| \leq M_\rho$ and under a compact $\mathcal{W}$, there is some $c_1 > 0$ such that the first term above is bounded by $c_1||\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}||$. The second term is bounded by $2\int 1\{||\boldsymbol{b}|| > M_\rho\}\phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(1)})d\boldsymbol{b} < 2\rho$, which can be made smaller than $||\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}||$. Hence, (SA.1) is bounded by $c_2||\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}||$ for some constant $c_2 > 0$.

The term (SA.2) equals to

$$
\begin{aligned}
&\left|\int k_j(\boldsymbol{w}, \boldsymbol{\beta}^{(2)}, \boldsymbol{b})\left(\phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(1)}) - \phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(2)})\right)d\boldsymbol{b}\right| \\
&\leq \int \left|\phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(1)}) - \phi(\boldsymbol{b}|\boldsymbol{0}, \boldsymbol{D}^{(2)})\right| d\boldsymbol{b} \\
&\leq \sqrt{\mathrm{tr}(\boldsymbol{D}^{(1)^{-1}}\boldsymbol{D}^{(2)} - \boldsymbol{I}) - \log\det(\boldsymbol{D}^{(1)}\boldsymbol{D}^{(2)^{-1}})},
\end{aligned}
$$

where the last inequality is due to a known bound on the total variation distance between normal distributions with a same mean vector but different covariance matrices (Devroye et al., 2018).

□

SA.2 Intermediate results in the proof of Theorem 2

The next lemma shows that the response probabilities are continuous for the total variation distance defined as

$$d_{TV}(\boldsymbol{p}_{\boldsymbol{\beta},\boldsymbol{\pi}}, \boldsymbol{p}_{\boldsymbol{\beta}',\boldsymbol{\pi}'}) = \int \sum_{\boldsymbol{y}\in\mathcal{J}^T} |p_{\boldsymbol{\beta},\boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{\omega})g^*(\boldsymbol{\omega}) - p_{\boldsymbol{\beta}',\boldsymbol{\pi}'}(\boldsymbol{y}|\boldsymbol{\omega})g^*(\boldsymbol{\omega})|d\boldsymbol{\omega}.$$

Lemma SA.3 (Continuity of response probabilities)**.** *Let $\varepsilon > 0$. Then there is $\delta > 0$ such that $d((\boldsymbol{\beta}, \boldsymbol{\pi}), (\boldsymbol{\beta}', \boldsymbol{\pi}')) < \delta$ implies that $d_{TV}(\boldsymbol{p}_{\boldsymbol{\beta},\boldsymbol{\pi}}, \boldsymbol{p}_{\boldsymbol{\beta}',\boldsymbol{\pi}'}) < \varepsilon$.*

Proof of Lemma SA.3.

$$\begin{aligned}
&|p_{\boldsymbol{\beta},\boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{\omega}) - p_{\boldsymbol{\beta}',\boldsymbol{\pi}'}(\boldsymbol{y}|\boldsymbol{\omega})| \leq |p_{\boldsymbol{\beta},\boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{\omega}) - p_{\boldsymbol{\beta}',\boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{\omega})| + |p_{\boldsymbol{\beta}',\boldsymbol{\pi}}(\boldsymbol{y}|\boldsymbol{\omega}) - p_{\boldsymbol{\beta}',\boldsymbol{\pi}'}(\boldsymbol{y}|\boldsymbol{\omega})| \\
&\leq \sum_c \pi_c \left| \prod_{t=1}^T \Pr\left(Y_{it} = y_t|\boldsymbol{\beta}, \boldsymbol{w}_t, c\right) - \prod_{t=1}^T \Pr\left(Y_{it} = y_t|\boldsymbol{\beta}', \boldsymbol{w}_t, c\right) \right| \\
&+ \sum_c |\pi_c - \pi_c'| \prod_{t=1}^T \Pr\left(Y_{it} = y_t|\boldsymbol{\beta}', \boldsymbol{w}_t, c\right) \\
&\leq \gamma_1||\boldsymbol{\beta} - \boldsymbol{\beta}'||_2 + \gamma_2||\boldsymbol{\pi} - \boldsymbol{\pi}'||_1,
\end{aligned}$$

for some positive constants γ_1 and γ_2. □

SB Sparsity property of the proposed M-H step

The proposed M-H step exhibits a "sparsity property" which we describe below. Suppose that an alternative j was not chosen by the subject i in any period (otherwise, it must be in the consideration set for i and $C_{ij} = 1$). Depending on the current $C_{ij}^{(g)}$, and the proposed $\tilde{C}_{ij}$, there are four possible moves in the M-H step. First, if $\tilde{C}_{ij} = C_{ij}^{(g)} = 1$ or $\tilde{C}_{ij} = C_{ij}^{(g)} = 0$, then the proposed value is accepted with probability one. Second, if

$\tilde{C}_{ij} = 0$ and $C_{ij}^{(g)} = 1$, then the proposed value is also accepted with probability one. In other words, the algorithm "prefers" a smaller consideration set. This sparsity-inducing property is proven below. Lastly, when the proposed consideration set adds an alternative j that is not in the current consideration set, that is, $\tilde{C}_{ij} = 1$ and $C_{ij}^{(g)} = 0$, the acceptance probability is between 0 and 1 and is determined by the likelihood ratio.

Proposition 1 (Sparsity-inducing property)**.** *Consider the M-H step described in Algorithm 1. Let j be an alternative that is not observed to be chosen by the subject i. If the step proposes to exclude j from the consideration set of i, it is accepted with probability 1.*

Proof. Let the consideration set for the ith subject at iteration g be $\mathcal{C}_i^{(g)}$. Suppose that a category $j \in \mathcal{C}_i^{(g)}$ is proposed to be removed so that $\tilde{\mathcal{C}}_i = \mathcal{C}_i^{(g)} \setminus \{j\}$. The acceptance probability is

$$\min\left\{\frac{p\big(\boldsymbol{Y}_i = \boldsymbol{y}_i \big| \boldsymbol{\beta}^{(g)}, \boldsymbol{b}_i^{(g)}, \boldsymbol{\omega}_i, \tilde{\mathcal{C}}_i\big)}{p\big(\boldsymbol{Y}_i = \boldsymbol{y}_i \big| \boldsymbol{\beta}^{(g)}, \boldsymbol{b}_i^{(g)}, \boldsymbol{\omega}_i, \mathcal{C}_i^{(g)}\big)}, 1\right\} = \min\left\{\frac{\prod_t \sum_{\ell \in \mathcal{C}_i^{(g)}} \exp\left(V_{i\ell t}\right)}{\prod_t \sum_{\ell \in \tilde{\mathcal{C}}_i} \exp\left(V_{i\ell t}\right)}, 1\right\} = 1,$$

where $V_{ijt} = \boldsymbol{x}_{ijt}'\boldsymbol{\beta}^{(g)} + \boldsymbol{z}_{ijt}'\boldsymbol{b}_i^{(g)}$, and the last equality is due to the fact that the ratio is larger than 1. Hence, $\tilde{\mathcal{C}}_i$ is accepted with probability 1. □

SC Conditional posterior distributions

For the mixture model on the latent consideration sets $\boldsymbol{C} = (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_n)$, let $S_i \in \{1, 2, \ldots\}$ be the latent cluster assignment such that $C_{ij} | S_i = h \sim \text{Bernoulli}(q_{hj})$, independently $j = 1, \ldots, J$, for $i = 1, \ldots, n$. We have the latent consideration sets $\boldsymbol{C}$, the common fixed-effects $\boldsymbol{\beta}$, the random effects $\boldsymbol{b}$, the corresponding covariance matrix $\boldsymbol{D}$, the DP parameters $\boldsymbol{V} = (V_1, V_2, \ldots)$ as well as $\boldsymbol{Q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots)$, the DP cluster assignment variables $\boldsymbol{S} = (S_1, \ldots, S_n)$, and the DP concentration parameter α. Then, from the Bayes theorem, we define the posterior density of interest to be

$$
\begin{aligned}
p\big(\boldsymbol{C},\, \boldsymbol{S}, \boldsymbol{V}, \boldsymbol{Q}, \alpha, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{D}\big|\boldsymbol{y}, \boldsymbol{\omega}\big) &\propto p\big(\boldsymbol{y}\big|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{C}\big) \cdot p(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{D}) \cdot p\big(\boldsymbol{C}, \boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{V}, \alpha\big) \\
&= p\big(\boldsymbol{y}\big|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{C}\big) \cdot \pi(\boldsymbol{\beta}) p(\boldsymbol{b}|\boldsymbol{D}) \pi(\boldsymbol{D}) \cdot p\big(\boldsymbol{C}, \boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{V}, \alpha\big), \qquad \text{(SC.1)}
\end{aligned}
$$

where the first term is the likelihood function and $\pi(\cdot)$ denotes the prior density. Only the last term in (SC.1) is associated with the DP model and

$$
\begin{aligned}
p\big(\boldsymbol{C}, \boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{V}, \alpha\big) &\propto p\big(\boldsymbol{C}|\boldsymbol{Q}, \boldsymbol{S}\big) p\big(\boldsymbol{Q}, \boldsymbol{V}, \boldsymbol{S}, \alpha\big) \\
&\propto \left[\prod_{i=1}^{n} p\big(\boldsymbol{C}_i|\boldsymbol{q}_{S_i}\big) p\big(S_i|\boldsymbol{V}\big)\right] \cdot \left[\prod_{h=1}^{\infty} p\big(V_h|\alpha\big) p\big(\boldsymbol{q}_h|\underline{\boldsymbol{\phi}}_q\big)\right] \cdot \pi(\alpha), \qquad \text{(SC.2)}
\end{aligned}
$$

where $p\big(\boldsymbol{C}_i|\boldsymbol{q}_{S_i}\big)$ is the product of densities for the independent Bernoulli distributions Bernoulli$(q_{S_i j})$ $j = 1, \ldots, J$, $p\big(S_i|\boldsymbol{V}\big) = \omega_{S_i}$, $p\big(V_h|\alpha\big)$ is the density of Beta$(1, \alpha)$, $p(\boldsymbol{q}_h|\underline{\boldsymbol{\phi}}_q)$ is the product of densities for the independent Beta distributions Beta$(\underline{a}_{q_j}, \underline{b}_{q_j})$ $j = 1, \ldots, J$, and $\pi(\alpha)$ is the prior density for α. We apply the slice sampling approach (Walker, 2007) by augmenting the joint distribution with a sequence of auxiliary random variables $\boldsymbol{u} = (u_1, \ldots, u_n)$ that follow the uniform distribution on $(0, 1)$, $u_i \sim \mathcal{U}(0, 1)$, $i = 1, \ldots, n$:

$$
p(\boldsymbol{C}, \boldsymbol{S}, \boldsymbol{Q}, \boldsymbol{V}, \boldsymbol{u}, \alpha) \propto \left[\prod_{i=1}^{n} p\big(\boldsymbol{C}_i|\boldsymbol{q}_{S_i}\big) I(u_i \leq \omega_{S_i})\right] \cdot \left[\prod_{h=1}^{\infty} p\big(V_h|\alpha\big) p\big(\boldsymbol{q}_h|\underline{\boldsymbol{\phi}}_q\big)\right] \cdot \pi(\alpha). \quad \text{(SC.3)}
$$

It is easy to show that we can recover (SC.2) by integrating out $\boldsymbol{u}$ from (SC.3). However, by introducing $\boldsymbol{u}$, one only has to choose labels S_i in the finite set $\{h : \omega_h \geq u_i\}$. See the Supplementary Material for discussion on hyperparameter selections.

Our MCMC algorithm proceeds by cycling through various conditional distributions, where these distributions are conditioned on the most recent values of the remaining unknowns. Specifically, given the current draw at the gth iteration $\{u_i^{(g)}\}, \{V_h^{(g)}\}, \{\boldsymbol{q}_h^{(g)}\}, \{S_i^{(g)}\}$, $\alpha^{(g)}$, $\{\boldsymbol{C}_i^{(g)}\}, \boldsymbol{\delta}^{(g)}$, $\boldsymbol{\beta}^{(g)}$, $\{\boldsymbol{b}_i^{(g)}\}$, and $\boldsymbol{D}^{(g)}$, the next draw in the sequence is obtained by simulating

$$\begin{aligned}
\boldsymbol{\beta}^{(g+1)} &\text{ from } \boldsymbol{\beta}\big|\{y_{it}\}, \{\boldsymbol{C}_i^{(g)}\}, \boldsymbol{\delta}^{(g)}, \{\boldsymbol{b}_i^{(g)}\}, \\
\{\boldsymbol{b}_i^{(g+1)}\} &\text{ from } \{\boldsymbol{b}_i\}\big|\{y_{it}\}, \{\boldsymbol{C}_i^{(g)}\}, \boldsymbol{\delta}^{(g)}, \boldsymbol{\beta}^{(g+1)}, \boldsymbol{D}^{(g)}, \\
\boldsymbol{D}^{(g+1)} &\text{ from } \boldsymbol{D}\big|\{\boldsymbol{b}_i^{(g+1)}\}, \\
\boldsymbol{\delta}^{(g+1)} &\text{ from } \boldsymbol{\delta}\big|\{y_{it}\}, \{\boldsymbol{C}_i^{(g)}\}, \boldsymbol{\beta}^{(g+1)}, \{\boldsymbol{b}_i^{(g+1)}\}, \\
\{\boldsymbol{C}_i^{(g+1)}\} &\text{ from } \{\boldsymbol{C}_i\}\big|\{y_{it}\}, \boldsymbol{\delta}^{(g+1)}, \boldsymbol{\beta}^{(g+1)}, \{\boldsymbol{b}_i^{(g+1)}\}, \{\boldsymbol{q}_h^{(g)}\}, \{S_i^{(g)}\}, \\
\{V_h^{(g+1)}\} &\text{ from } \{V_h\}\big|\{S_i^{(g)}\}, \alpha^{(g)}, \\
\{\boldsymbol{q}_h^{(g+1)}\} &\text{ from } \{\boldsymbol{q}_h\}\big|\{\boldsymbol{C}_i^{(g+1)}\}, \{S_i^{(g)}\}, \\
\{u_i^{(g+1)}\} &\text{ from } \{u_i\}\big|\{S_i^{(g)}\}, \{V_h^{(g+1)}\}, \\
\{S_i^{(g+1)}\} &\text{ from } \{S_i\}\big|\{u_i^{(g+1)}\}, \{\boldsymbol{q}_h^{(g+1)}\}, \{V_h^{(g+1)}\}, \{\boldsymbol{C}_i^{(g+1)}\}, \\
\alpha^{(g+1)} &\text{ from } \alpha\big|\{V_h^{(g+1)}\}, \{S_i^{(g+1)}\}.
\end{aligned}$$

Repeating this procedure G times (beyond a suitable burn-in) produces a sample from the posterior distribution.

The main paper illustrates how the consideration sets are simulated. In this section, we show the conditional posterior distributions of the remaining parameters. Let $K^* = \min\{h : \sum_{\ell=1}^{h} \omega_h > 1 - u^*\}$, where $u^* = \min(u_1, \ldots, u_n)$. Define $n_h = \sum_{i=1}^{n} I(S_i = h)$. Let the dot $\bullet$ denote all other parameters and the data.

SC.1 Simulation of q_h

From (SC.2), we have that

$$p(\boldsymbol{q}_h|\bullet) \propto p(\boldsymbol{q}_h|\underline{\boldsymbol{\phi}}_q) \cdot \prod_{i:S_i=h} \prod_{j=1}^{J} q_{hj}^{C_{ij}} \left(1 - q_{hj}\right)^{1-C_{ij}},$$

where $p(\boldsymbol{q}_h|\underline{\boldsymbol{\phi}}_q)$ is the product of densities for Beta distributions $Beta(\underline{a}_{q_j}, \underline{b}_{q_j})$, independently over $j = 1, \ldots, J$. Then

$$q_{hj}|\bullet \sim \text{Beta}\left(\underline{a}_{q_j} + \sum_{i:S_i=h} C_{ij}, \underline{b}_{q_j} + \sum_{i:S_i=h} (1 - C_{ij})\right),$$

independently over $j = 1, \ldots, J$ for $h = 1, 2, \ldots, K^*$. If component $h \leq K^*$ does not contain any observations, then the corresponding $\boldsymbol{q}_h$ is drawn from the prior.

SC.2 Simulation of V_h

From (SC.2), the conditional distribution of $\boldsymbol{V}$ is independent and the marginal conditional distributions are

$$V_h|\bullet \sim \text{Beta}\left(1 + n_h, \alpha + \sum_{\ell>h} n_\ell\right),$$

for $h = 1, 2, \ldots, K^*$. If component $h \leq K^*$ is empty, then the corresponding V_h is drawn from the prior.

SC.3 Simulation of u_i

From (SC.3), it is easy to see that

$$u_i|\bullet \overset{ind}{\sim} \mathcal{U}[0, \omega_{S_i}], \quad i = 1, \ldots, n.$$

SC.4 Simulation of S_i

From (SC.3), we can see that for $h = 1, 2, \ldots, K^*$,

$$Pr\left(S_i = h|\boldsymbol{C}, \boldsymbol{u}, \boldsymbol{V}, \boldsymbol{Q}\right) = \frac{I\left(u_i \leq \omega_h\right) \prod_{j=1}^{J} q_{hj}^{C_{ij}} \left(1 - q_{hj}\right)^{1-C_{ij}}}{\sum_{\ell} I\left(u_i \leq \omega_\ell\right) \prod_{j=1}^{J} q_{\ell j}^{C_{ij}} \left(1 - q_{\ell j}\right)^{1-C_{ij}}}.$$

Note that $Pr(S_i = h|\bullet) = 0$ for $h > K^*$.

SC.5 Simulation of α

The conditional posterior of α is

$$p(\alpha|\bullet) \propto p(\boldsymbol{S}|\alpha)\pi(\alpha).$$

Following Escobar and West (1995), this distribution is sampled by first generating η conditional on α from the Beta distribution

$$\eta|\alpha, \boldsymbol{S} \sim \text{Beta}(\alpha + 1, n),$$

and then sampling α conditional on η from the Gamma mixture

$$\begin{aligned} p(\alpha|\eta, \boldsymbol{S}) = & \frac{\underline{a}_\alpha + G - 1}{\underline{a}_\alpha + G - 1 + n(\underline{b}_\alpha - \log(\eta))} \text{Gamma}(\underline{a}_\alpha + G, \underline{b}_\alpha - \log(\eta)) \\ & + \frac{n(\underline{b}_\alpha - \log(\eta))}{\underline{a}_\alpha + G - 1 + n(\underline{b}_\alpha - \log(\eta))} \text{Gamma}(\underline{a}_\alpha + G - 1, \underline{b}_\alpha - \log(\eta)), \end{aligned}$$

where G is the total number of existing clusters.

SC.6 Simulation of β

From Bayes theorem,

$$\pi(\boldsymbol{\beta}|\bullet) \propto \pi(\boldsymbol{\beta}) \cdot \prod_{i=1}^{n} \prod_{t=1}^{T_i} \Pr(Y_{it} = y_{it}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i),$$

where $\Pr(Y_{it} = y_{it}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i) = \frac{\exp\left(V_{iy_{it}t}\right)}{\sum_{\ell \in \mathcal{C}_i} \exp(V_{i\ell t})}$ and $V_{ijt} = \delta_j + \boldsymbol{x}'_{ijt}\boldsymbol{\beta} + \boldsymbol{z}'_{ijt}\boldsymbol{b}_i$.

We use a tailored Metropolis–Hastings (M-H) algorithm to sample $\boldsymbol{\beta}$ (Chib and Greenberg, 1995). Define the conditional log-likelihood of $\boldsymbol{\beta}$ given $\boldsymbol{\delta}$, $\{\boldsymbol{b}_i\}$, and $\{\boldsymbol{C}_i\}$: $\log L(\boldsymbol{\beta}|\bullet) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \log \Pr(Y_{it} = y_{it}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \mathcal{C}_i)$. At iteration g, let $\boldsymbol{\beta}^{(g)}$ be the value of $\boldsymbol{\beta}$. A candidate value is drawn as

$$\tilde{\boldsymbol{\beta}} \sim N_{d_x}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{V}}_\beta\right),$$

where

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}|\bullet)\pi(\boldsymbol{\beta}), \qquad \hat{\boldsymbol{V}}_\beta^{-1} = -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \log L(\boldsymbol{\beta}|\bullet)\pi(\boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},$$

which is accepted with probability

$$\min\left\{\frac{\pi(\tilde{\boldsymbol{\beta}}|\bullet)\phi(\boldsymbol{\beta}^{(g)}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{V}}_\beta)}{\pi(\boldsymbol{\beta}^{(g)}|\bullet)\phi(\tilde{\boldsymbol{\beta}}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{V}}_\beta)}, 1\right\},$$

where $\phi(\quad)$ denotes the density of normal distribution. The conditional posterior mode $\hat{\boldsymbol{\beta}}$ is computed using the Newton-Raphson method. The likelihood is known to be concave with respect to $\boldsymbol{\beta}$ under the Gumbel error distribution, so the convergence to $\hat{\boldsymbol{\beta}}$ is fast and only requires a few iterations in many cases. In the empirical application, we multiply the variance of the proposal distribution by 10^{-2} in order to achieve desirable acceptance rates.

SC.7 Simulation of b_i

The full conditional of $\boldsymbol{b}_i$ (for each i) is proportional to

$$\pi(\boldsymbol{b}_i|\bullet) \propto \phi(\boldsymbol{b}_i|\boldsymbol{0}, \boldsymbol{D}) \cdot \prod_{t=1}^{T_i} \Pr(Y_{it} = y_{it}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i).$$

We use a symmetric random-walk M-H to draw from the conditional distribution. Define the conditional log-likelihood of $\boldsymbol{b}_i$ given $\boldsymbol{\delta}$, $\boldsymbol{\beta}$, and $\{\boldsymbol{C}_i\}$: $\log L(\boldsymbol{b}_i|\bullet) = \sum_{t=1}^{T_i} \log \Pr(Y_{it} = y_{it}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i)$. At iteration g, let $\boldsymbol{b}_i^{(g)}$ be the value of $\boldsymbol{b}_i$. A candidate value is drawn as

$$\tilde{\boldsymbol{b}}_i \sim N_{d_z}\left(\boldsymbol{b}_i^{(g)}, \boldsymbol{D}^{(g)}\right),$$

which is accepted with probability

$$\min\left\{\frac{\pi(\tilde{\boldsymbol{b}}_i|\bullet)}{\pi(\boldsymbol{b}_i^{(g)}|\bullet)}, 1\right\}.$$

The updating step for b_i is independent over i, so it can be easily parallelized in a modern computer.

SC.8 Simulation of D

We simulate $\boldsymbol{D}$ by first simulating $\boldsymbol{D}^{-1}$ and then taking the inverse of the simulated draw. This is because it can be shown that

$$\boldsymbol{D}^{-1}|\bullet \sim \text{Wishart}\left(\underline{v} + n, \left[\underline{\boldsymbol{R}}^{-1} + \sum_{i=1}^{n} \boldsymbol{b}_i \boldsymbol{b}_i'\right]^{-1}\right).$$

SC.9 Simulation of δ

In princple, we could treat $\boldsymbol{\delta}$ as a part of $\boldsymbol{\beta}$ and sample from the conditional distribution altogether using a tailored M-H algorithm. However, the involved optimization step could be slow when J is large, which is exactly our focus of the current paper. Hence, we sample $\boldsymbol{\delta}$ separately from $\boldsymbol{\beta}$. Specifically, we use a tailored Metropolis–Hastings (M-H) algorithm to sample δ_k for $k = 1, \ldots, J-1$, one after another.

From Bayes theorem,

$$\pi(\delta_k|\boldsymbol{\delta}_{\setminus k}, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\omega}, \boldsymbol{C}) \propto \pi(\delta_k) \cdot \prod_{i=1}^{n} \prod_{t=1}^{T_i} \Pr(Y_{it} = y_{it}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i),$$

where $\boldsymbol{\delta}_{\setminus k}$ denotes $\boldsymbol{\delta}$ except for the kth element. Define the conditional log-likelihood of δ_k given $\boldsymbol{\delta}_{\setminus k}$, $\boldsymbol{\beta}$, $\{\boldsymbol{b}_i\}$, and $\{\boldsymbol{C}_i\}$: $\log L(\delta_k|\bullet) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \log \Pr(Y_{it} = y_{it}|\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{b}_i, \boldsymbol{\omega}_{it}, \mathcal{C}_i)$. At iteration g, let $\delta_k^{(g)}$ be the value of δ_k. A candidate value is drawn as

$$\tilde{\delta}_k \sim N_1\left(\hat{\delta}_k, \hat{\sigma}^2_{\delta_k}\right),$$

where

$$\hat{\delta}_k = \arg\max_{\delta_k} \log L(\delta_k|\bullet)\pi(\delta_k), \qquad \hat{\sigma}^{-2}_{\delta_k} = -\frac{\partial^2}{\partial \delta_k^2} \log L(\delta_k|\bullet)\pi(\delta_k)\bigg|_{\delta_k=\hat{\delta}_k},$$

which is accepted with probability

$$\min\left\{\frac{\pi(\tilde{\delta}_k|\bullet)\phi(\delta_k^{(g)}|\hat{\delta}_k, \hat{\sigma}^2_{\delta_k})}{\pi(\delta_k^{(g)}|\bullet)\phi(\tilde{\delta}_k|\hat{\delta}_k, \hat{\sigma}^2_{\delta_k})}, 1\right\}.$$

We randomize the order of updating δ_k, $k = 1, \ldots, J-1$.

SD Prior on the distribution of attention probabilities

SD.1 Remarks on hyperparameters

In the fitting, we set the parameters of the prior as follows: for the product-specific fixed-effects, $\delta_j \sim N(0,2)$ independently for $j = 1, \ldots, J$, for the common fixed-effect, $\beta_k \sim N(0,3)$ independently for $k = 1, \ldots, d_x$, for the variance of the random effects, $\boldsymbol{D}^{-1} \sim \text{Wishart}(9, (1/9)\boldsymbol{I}_{d_z})$, and for the DP concentration parameter, $\alpha \sim \text{Gamma}(\underline{a}_\alpha, \underline{b}_\alpha) = (1/4, 1/4)$ to produce $\Pr(H_0 : \omega^* > 1 - \varepsilon) \approx 0.5$. The prior of the attention probabilities is $q_{hj} \sim \text{Beta}(\underline{a}_{q_j}, \underline{b}_{q_j})$, independently over $j = 1, \ldots, J$ for $h = 1, \ldots, \infty$. The choice of hyperparameters, $(\underline{a}_{q_j}, \underline{b}_{q_j})$, is important, as it controls the sparsity of the consideration sets. We set $(\underline{a}_{q_j}, \underline{b}_{q_j}) = (s \cdot r, s \cdot (1-r))$, where $s > 0$ and r is a small prior expectation of q_{hj} (that is, $r < 0.5$), for example, $r = \frac{r_0}{J}$, where r_0 is a positive integer. We call this a sparsity-supporting prior because the prior probability is smaller for consideration sets with larger cardinality.

SD.2 Illustration

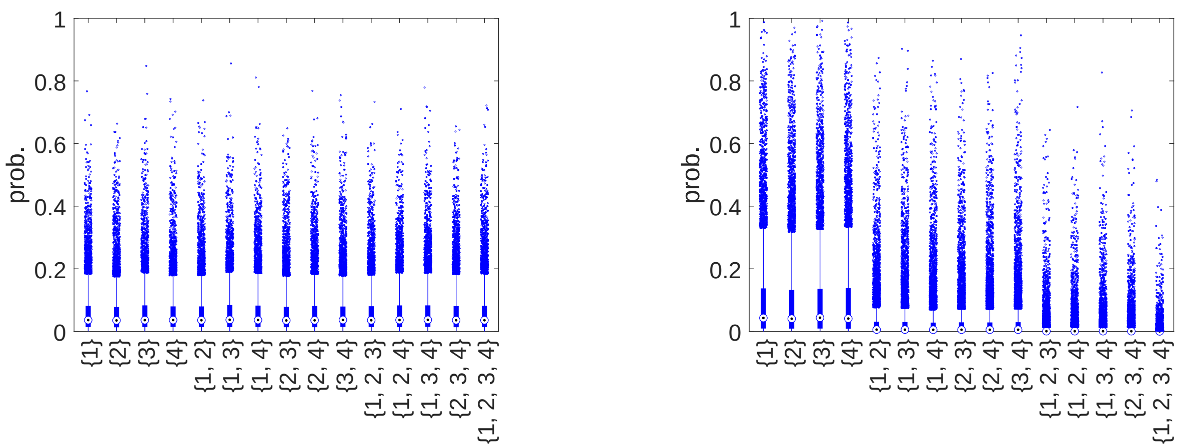
When J is small, we can examine the impact of the hyperparameters on the implied prior probability distribution on consideration sets by simulating from the prior. First, fix a large positive integer K. Second, generate draws from the prior by drawing

$$
\begin{aligned}
\alpha &\sim \text{Gamma}(\underline{a}_\alpha, \underline{b}_\alpha), \\
V_h|\alpha &\stackrel{ind}{\sim} \text{Beta}(1, \alpha) \text{ for } h = 1, \ldots, K, \\
\omega_h &= V_h \prod_{\ell > h} (1 - V_\ell) \text{ for } h = 1, \ldots, K, \\
q_{hj} &\stackrel{ind}{\sim} \text{Beta}(\underline{a}_{q_j}, \underline{b}_{q_j}) \text{ for } j = 1, \ldots, J, \quad h = 1, \ldots, K.
\end{aligned}
$$

Finally, given these draws, calculate the probability of each possible consideration set using the representation in Lemma 1; that is,

$$\pi_c = \sum_{h=1}^{K} \omega_h \left\{ \prod_{j \in c} q_{hj} \prod_{j \notin c} (1 - q_{hj}) \right\}.$$

For example, when $J = 4$, $\Pr(\mathcal{C}_i = \{2,4\}) = \sum_{h=1}^{K} \omega_h \left\{ q_{h2} q_{h4} \big(1 - q_{h1}\big)\big(1 - q_{h3}\big) \right\}$.



(a) Uniform prior

(b) A sparsity-supporting prior

Figure SD.1: Implied prior distribution over consideration sets (box plots) for two different priors on q_{hj}. Uniform prior (a) with $(\underline{a}_{q_j}, \underline{b}_{q_j}) = (1, 1)$ and sparsity supporting prior (b) with $(\underline{a}_{q_j}, \underline{b}_{q_j}) = (sr, s(1-r))$ with $r = \frac{1}{J}$, $s = 1$. $K = 20$. $(\underline{a}_{\alpha}, \underline{b}_{\alpha}) = (1/4, 1/4)$. and 10,000 draws from the prior.

Panel (a) in Figure SD.1 shows the implied prior distribution over the consideration sets under the uniform prior on q_{hj} when $J = 4$. Under the uniform prior, the prior expectation of $q_{hj} = 0.5$, so the prior is the same across all consideration sets and is centered around $0.5^4 = 0.0625$. Panel (b) gives results under our sparsity-supporting prior. In this case, the prior distribution shrinks to 0 as the cardinality of the consideration set increases.

The preceding shows that the prior on the attention probabilities $\{q_{hj}\}$ induces quite different prior distributions on consideration sets. As the number of consideration sets increase exponentially in J, it is crucial to apply regularization to the parameter space.

Our sparsity-supporting prior promotes this regularization. It favors smaller consideration sets, while maintaining positive probabilities on larger sets.

SE Additional material for the simulation

SE.1 Simulation results with random effects

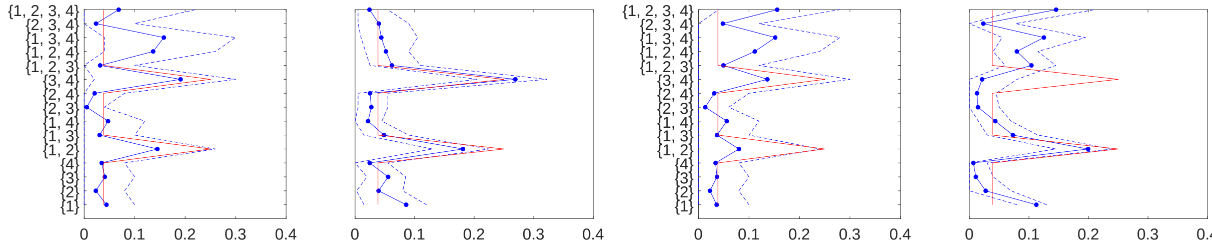
We repeat the simulation study now with preference heterogeneity. We generate the consideration sets as before, but now use the random effects logit with the specification $V_{ijt} = \delta_j^* + (\beta^* + b_i)x_{ijt}$, where $b_i \sim N(0, D^*)$ with $D^* = 1.1^2$. We fit the random effects logit with the proposed flexible approach for the distribution of consideration sets. The results are presented in Table SE.1. We see that as n increases, the MSEs/bias/standard deviation tend to decrease. However, this is not the case when the independent consideration is imposed, i.e. $K = 1$. Also, MSEs/biases are larger in general than for the proposed flexible approach. Figure SE.1 shows that for $K = \infty$, we see the posterior on $\boldsymbol{\pi}$ approaches to the truth although it does not under $K = 1$ due to the mis-specification.

Table SE.1: Simulation results for $J = 4$ with random effects

		β		δ_1		δ_2		δ_3		D		π	
$K = \infty$	$n = 50$	0.17	(0.183)	0.509	(0.674)	0.507	(0.696)	0.108	(0.437)	0.952	(0.036)	0.521	(0.04)
	$n = 200$	0.1	(0.114)	0.169	(0.368)	0.164	(0.377)	0.057	(0.245)	0.556	(0.124)	0.311	(0.02)
$K = 1$	$n = 50$	0.186	(0.172)	1.358	(0.672)	1.369	(0.702)	0.125	(0.427)	0.951	(0.034)	0.767	(0.04)
	$n = 200$	0.166	(0.093)	2.298	(0.258)	2.321	(0.265)	0.071	(0.214)	0.708	(0.101)	0.781	(0.02)

For β and δ, we show the averages of the mean squared errors, using the posterior means as point estimators. For π, we show the average of L_1 norm between the posterior mean and $\boldsymbol{\pi}^*$. The average posterior standard deviations are in parentheses.

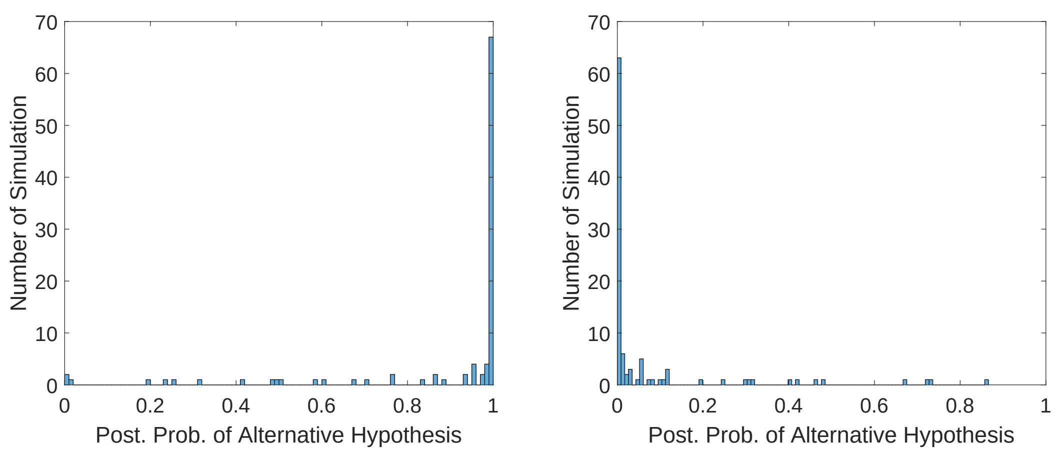
Figure SE.2 (a) shows a histogram of the estimated posterior probability of H_1 (dependent consideration) when H_1 is true. The method appropriately assigns values close to one for the majority of the simulations. Figure SE.2 (b) shows the result when H_0 is true. The posterior probability assigned to H_1 is close to zero for the majority of the simulations.



(a) $K = \infty, n = 50$ (b) $K = \infty, n = 200$ (c) $K = 1, n = 50$ (d) $K = 1, n = 200$

Figure SE.1: The true distribution over consideration sets (solid, red), posterior mean (solid with dots, blue), 95% equal-tailed credible interval (dashed, blue). Each plot is based on one realization of simulated data. $J = 4$, $T = 3$, with random effects.

In summary, even with random effects, our proposed method can deliver consistent estimates of the preference parameters i.e. $\boldsymbol{\beta}$ and $\boldsymbol{D}$ as well as the distribution of consideration sets $\boldsymbol{\pi}$. In addition, our method can be used to test whether latent consideration is independent or dependent.



(a) Case 1 (dependent consideration) (b) Case 2 (independent consideration)

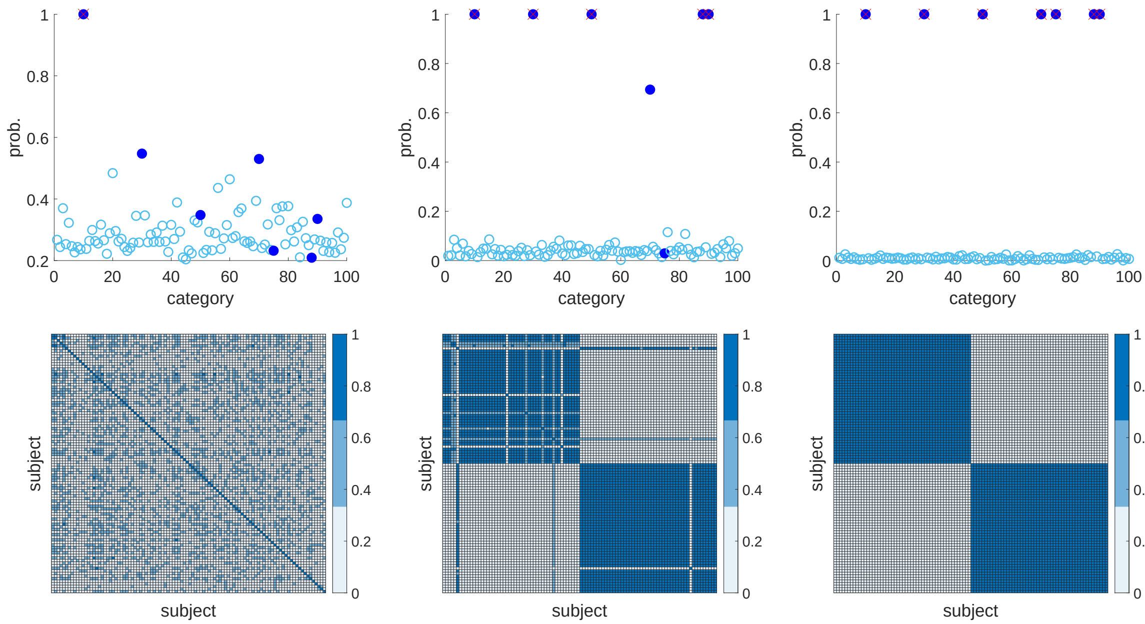
Figure SE.2: Histograms of estimated posterior probabilities of H_1 in each of the 100 simulations under (a) case 1 (dependent consideration - H_1 is true) and (b) case 2 (independent consideration - H_0 is true). $\varepsilon = 0.1, n = 50$.

SF Simulation with $J = 100$

SF.1 Large J

We now consider a high-dimensional scenario with $J = 100$. One mechanism by which the dependence of consideration among categories can be induced is through multiple latent subpopulations of subjects having different probabilities of consideration. Within a subpopulation, considerations are independent across categories. However, marginalizing out the latent subpopulation indicator, one obtains dependence in those category considerations. We generate the data with two subpopulations. To generate the true consideration set of a given subject, we used a Bernoulli distribution with attention probability 0.05 for each category except for categories 10, 30, 50, 70, and 90 for the first subpopulation ($i = 1, \ldots, 50$) where the Bernoulli attention probability was set to 0.8. For the remaining subjects in the second subpopulation ($i = 51, \ldots, 100$), the Bernoulli probability was set at 0.05 except for categories 20, 40, 60, 80, and 100 where the probability was set to 0.8. Conditional on the true consideration sets, we generated the responses as in the case with $J = 4$.

Because in this case there are $2^{100} - 1$ support points in the distribution of the consideration sets, it is not possible to show the entire distribution as in the case of $J = 4$ in the paper. We focus on the estimation results regarding subject 1 whose true consideration set contains 7 categories: $\mathcal{C}_1^* = \{10, 30, 50, 70, 75, 88, 90\}$. The upper panels of Figure SF.1 show the averages of the posterior probabilities that $\mathcal{C}_1$ include each of the categories (filled circle if the particular category is in the true consideration set and unfilled circle otherwise). The lower panels of Figure SF.1 are the $n \times n$ similarity matrices. These give the posterior probability that a given subject in a particular row i is in the same cluster



(a) $T = 1$ (b) $T = 10$ (c) $T = 20$

Figure SF.1: Upper panels: Posterior mean of $1\{j \in \mathcal{C}_1\}$ for each item $j = 1, \dots, J$. Filled circle if the category j is within the true consideration set $\mathcal{C}_1^* = \{10, 30, 50, 70, 75, 88, 90\}$. Unfilled circle if it is outside $\mathcal{C}_1^*$. A cross-sign indicates $\exists t \leq T : y_{it} = j$. Lower panels: similarity matrices. The results are based on the simulated data with $J = 100$, $T = \{1, 10, 20\}$, and $n = 100$.

as another subject at a specific column i' which is computed as the posterior probability of the event $\{S_i = S_{i'}\}$. This probability ranges from zero (light blue) to one (dark blue).

We again see the posterior concentration toward the true consideration set as T increases. Even with relatively small time periods such as $T = 1$ or $T = 10$, the categories that have not yet been observed as responses by the subject 1 have relatively large posterior probabilities of being in the consideration set $\mathcal{C}_1$. This is because subject 1 tends to be clustered together with other subjects whose observed responses include those categories. We see that the true clustering structure is recovered accurately at $T = 10$ in the similarity matrix. This leads to the estimated consideration sets to be similar within the cluster.

SG Additional material for the application

SG.1 Data description

We combine two sources of the data sets obtained from Nielsen, a store data and a purchase data, in order to prepare a panel data set. The preference and consideration patterns might have affected during the pandemic, so we chose the year 2019, which is the earliest year available before the pandemic. In the store data, we first choose a retailer, whose identity is not revealed in the Nielsen data, which consistently had over 100 cereal brands available at the majority of its stores. There are 239 stores under this retailer, operating mainly in the Midwest of the United States. See Figure 3 for the locations of the stores and percentages of the purchases. The store data contains product information at UPC (universal product code) level such as price and size (ounce). A "brand" can consist of multiple UPCs. Brand-level prices are defined as size-weighted averages of UPC prices. We first pick the top 135 cereal brands in terms of the availability at these stores, which are responsible for over 90 percentages of the purchases in the purchase data at these stores in 2019. In more than 95% of the store-week combinations, the price information of the 135 brands is available, but if it is missing, we impute the value with the average of the prices of the same brand at the other stores in the same week. We then defined the top 100 to be the inside options and the rest to be the "other" option. In the purchase data, we removed the households who made less than 3 units of cereal. When households purchased multiple units of cereal at one shopping trip, we treat them as separate purchases. This leaves us a sample with $J = 101$ brands (see Appendix for a complete list of the brand names). The data contains $n = 1880$ households, 25,849 purchases at 239 stores of the same retailer throughout 52 weeks.

SG.2 Hyperparameters

We set the hyper prior parameters as follows: a sparsity-supporting prior for the attention probabilities $q_{hj} \sim \text{Beta}(\underline{a}_{q_j}, \underline{b}_{q_j})$, independently over $j = 1, \ldots, J$ for $h = 1, \ldots, \infty$, with $(\underline{a}_{q_j}, \underline{b}_{q_j}) = (s \cdot r, s \cdot (1-r))$, $r = \frac{r_0}{J}$ with $s = 5$ and $r_0 = 30$, which implies that the prior mean of q_{hj} is about 0.44. For the DP concentration parameter, $\alpha \sim \text{Gamma}(1/4, 2)$. The priors for $\boldsymbol{\delta}$ and β are independent normal distributions with zero mean and variance 3. The prior for D is an inverse-Wishart distribution with hyper-parameters $(\underline{v}, \underline{R}) = (9, (1/9)I)$.

SG.3 Additional estimation results and discussion

SG.3.1 Estimated parameters in the response model

Brand-specific fixed-effects. The number of brand-specific fixed-effects whose 95% credible intervals do not include 0 is larger for MNL than MNL_C and for MNL_R than MNL_RC. This phenomenon was also observed by Chiang et al. (1998). To explain this, we note that under MNL_C and MNL_RC, the estimated consideration sets $\{\mathcal{C}_i\}$ are much smaller than the set with all brands. If, for example, there is a brand that is almost never chosen by any household, the estimated $\{\mathcal{C}_i\}$ tends to exclude such a brand. The standard logit model does not account for such nonconsideration and instead assumes that every household considers all brands. As a result, the magnitudes (absolute value) of brand-specific fixed effects tend to be overestimated. Under the full specification, for 70 out of 100 of them, the corresponding 95% credible interval does not include 0. Note that we fixed $\delta_J = 0$ for identification.

SG.3.2 Estimated parameters in the mixture model

In the following, we present additional estimation results concerning the parameters in the mixture model under the MNL_RC specification. Figure SG.1 compares the prior and posterior densities of the DP concentration parameter α. The vague prior density $\alpha \sim Gamma(1/4, 1/4)$, suggested by Dunson and Xing (2009), is shown as the dashed line and the posterior as the solid line.

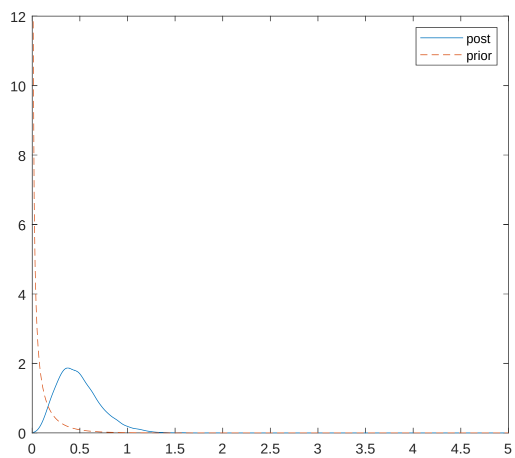


Figure SG.1: pdf's of α: prior (dashed) and posterior (solid) from the empirical application. MNL_RC.

Figure SG.2 shows the posterior probability mass function of the non-empty mixture components. The posterior mode of the number of non-empty components is six.

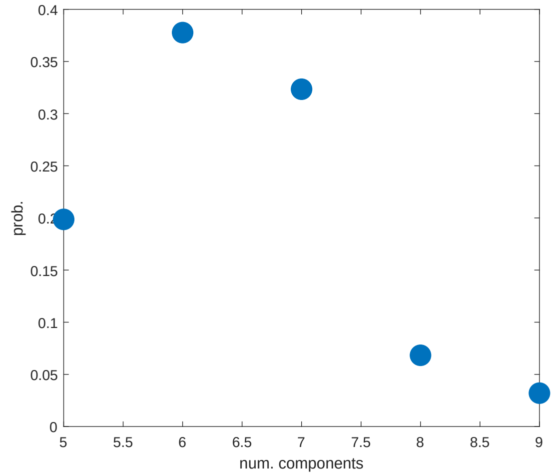


Figure SG.2: Posterior probability mass function of the number of nonempty components from the empirical application. MNL_RC.

The similarity matrix is shown in Figure SG.3. Each entry of the matrix shows the posterior probability that a given pair of households (i, i') are clustered together i.e. $S_i = S_{i'}$, ranging from zero (light blue) to one (dark blue).

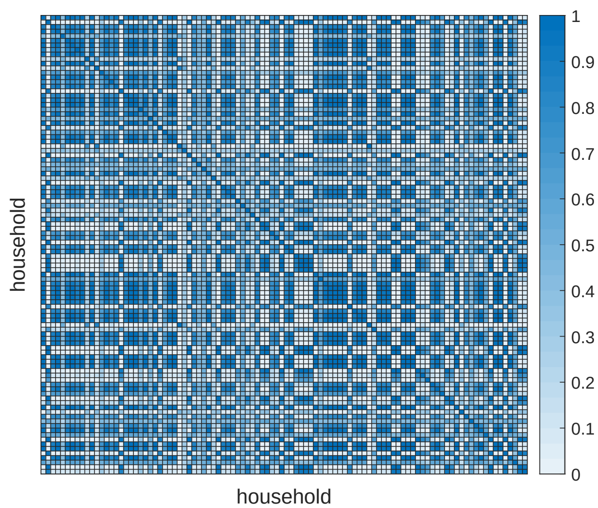


Figure SG.3: The similarity matrix of a sample of 100 households (out of 1880).

SG.3.3 Additional material on the prediction

To investigate why MNL_RC outperforms MNL_R in prediction, we compare the predictive response probabilities between the two models. For each $i \in \mathcal{O}$, we can compute the marginal posterior of $\Pr(Y_{iT_i+s} = j)$, for each alternative j and forecasting horizon $s = 1, \ldots, h_i$. Figure SG.4 presents the estimated response probabilities for the household $i = 3$ in the first out-of-sample period, $s = 1$. This household repeatedly purchased brands 13, 45, 57, and 101 in the estimation sample: $\{45, 13, 13, 101, 13, 13, 13, 57, 57, 57, 57\}$, in the order of the purchases. In the first out-of-sample week, the household purchased brand 13. The figure shows the 90% credible intervals (vertical bars) as well as the mean of the estimated response probabilities (circles). Clearly, the estimated response probabilities are much sparser for MNL_RC (lower panel) than MNL_R (upper). The traditional MNL approach necessarily implies a positive probability for every alternative. In contrast, the consideration set model allows many alternatives to actually receive zero predictive probabilities. Thus, incorporating consideration set heterogeneity can improve predictive performance due to the sparsity in the predictive response probabilities when the time-invariant consideration set assumption is appropriate, which seems to be the case in this data set.

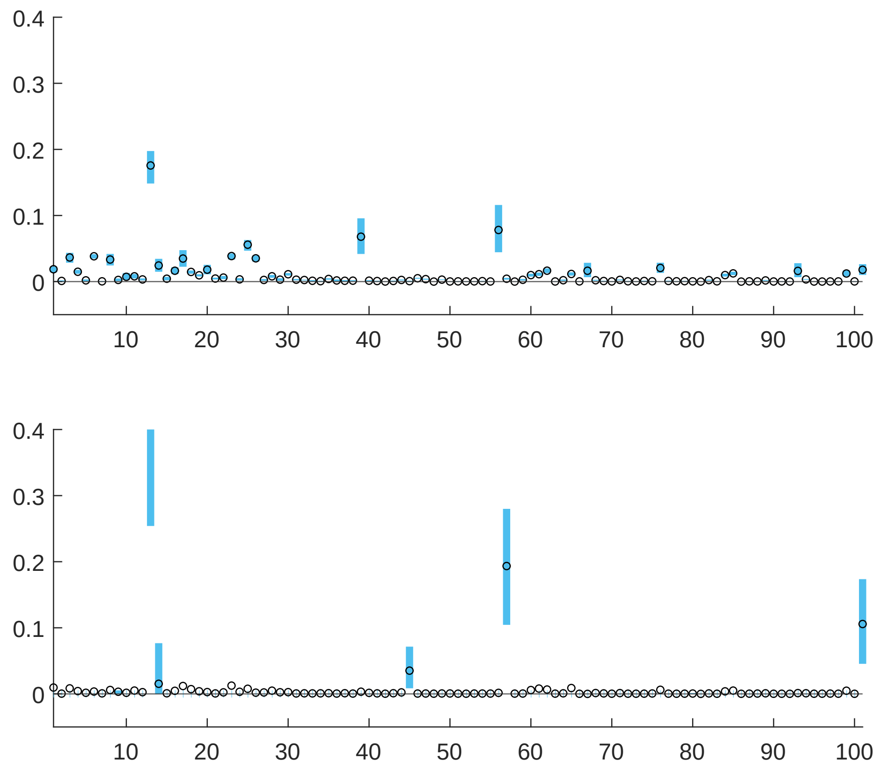
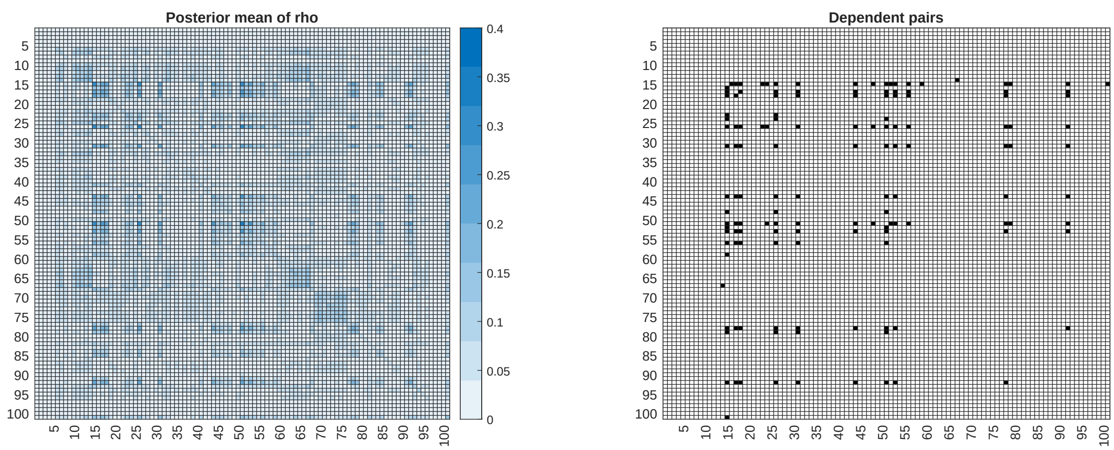


Figure SG.4: Estimated predictive response probabilities $\Pr(Y_{i,T_i+1} = j)$ for $i = 3$. 90% credible intervals (bars) and means (circles). The horizontal axis represents the brands $j \in \{1, \dots, 101\}$. The actual out-of-sample purchase was brand 13.

SG.3.4 Estimated consideration dependence

We conduct the test for independent consideration introduced in Section 6. The estimated posterior probability of the alternative hypothesis is very close to one i.e. $\Pr(H_1|\boldsymbol{D}^n) \approx 1$, and we conclude that the considerations of cereal products in this particular market are dependent. Furthermore, to investigate brand pair-level dependence, consider a hypothetical consumer i whose consideration set is drawn from the true unknown distribution. Define the marginal probability that brand j is (and not) considered: $\pi_1^{(j)} = \Pr(C_{ij} = 1)$ and $\pi_0^{(j)} = \Pr(C_{ij} = 0)$. Also define the probability that a pair of brands (j, ℓ) is considered jointly as $\pi_{11}^{(j,\ell)} = \Pr(C_{ij} = 1 \text{ and } C_{i\ell} = 1)$, and similarly define the probabilities for the remaining three cases: $\pi_{01}^{(j,\ell)} = \Pr(C_{ij} = 0 \text{ and } C_{i\ell} = 1)$, $\pi_{10}^{(j,\ell)} = \Pr(C_{ij} =$

1 and $C_{i\ell} = 0$), and $\pi_{00}^{(j,\ell)} = \Pr(C_{ij} = 0 \text{ and } C_{i\ell} = 0)$. We employ the model-based Cramer's V statistics as a measure of consideration dependence between brands j and ℓ as: $\rho_{j,\ell}^2 = \sum_{s=0}^{1}\sum_{m=0}^{1}\left(\pi_{(s,m)}^{(j,\ell)} - \pi_{(s)}^{(j)}\pi_{(m)}^{(\ell)}\right)^2 / \pi_{(s)}^{(j)}\pi_{(m)}^{(\ell)}$, which ranges from 0 to 1, and $\rho_{j,\ell}^2 \approx 0$ indicates that the consideration of the two brands (j,ℓ) is nearly independent. These probabilities are approximated as functions of the model parameters, for example, $\pi_1^{(j)} = \sum_{h=1}^{k^*}\omega_h q_{hj}$, $\pi_0^{(j)} = \sum_{h=1}^{k^*}\omega_h(1-q_{hj})$, and $\pi_{10}^{(j,\ell)} = \sum_{h=1}^{k^*}\omega_h q_{hj}(1-q_{h\ell})$, and so on. Figure SG.5a shows the posterior means of $\{\rho_{j,\ell}\}$. Figure SG.5b shows the brand pairs (j,ℓ) for which the posterior probability that $\rho_{j,\ell} > 0.1$ is greater than 0.95. Based on this criteria, we identified 69 brand pairs (shown in black).



(a) Posterior means of $\rho_{j,\ell}$

(b) Dependent brands

Figure SG.5: Consideration dependence in the 2019 Midwest cereal consumption data.