Understanding Choice Intensity: A Poisson Mixture Model with Logit-based Random Utility Selective Mixing *

Martin Burda,† Matthew Harding,‡ Jerry Hausman§

February 28, 2011

Abstract

In this paper we introduce a new Poisson mixture model for count panel data where the underlying Poisson process intensity is determined endogenously by consumer latent utility maximization over a set of choice alternatives. This formulation accommodates the choice and count in a single random utility framework with desirable theoretical properties. Individual heterogeneity is introduced through a random coefficient scheme with a flexible semiparametric distribution. We deal with the analytical intractability of the resulting mixture by recasting the model as an embedding of infinite sequences of scaled moments of the mixing distribution, and newly derive their cumulant representations along with bounds on their rate of numerical convergence. We further develop an efficient recursive algorithm for fast evaluation of the model likelihood within a Bayesian Gibbs sampling scheme, and show posterior consistency. We apply our model to a recent household panel of supermarket visit counts. We estimate the nonparametric density of three key variables of interest – price, driving distance, and their interaction – while controlling for a range of consumer demographic characteristics. We use this econometric framework to assess the opportunity cost of time and analyze the interaction between store choice, trip frequency, search intensity, and household and store characteristics. We also conduct a counterfactual welfare experiment and compute the compensating variation for a 10% to 30% increase in Walmart prices.

JEL: C11, C13, C14, C15, C23, C25
Keywords: Bayesian nonparametric analysis, Markov chain Monte Carlo, Dirichlet process prior

*We would like to thank the participants of seminars at Harvard-MIT, Princeton, Stanford, USC, Yale, and audiences of the 2009 CEA, MEG, and SITE meetings for their insightful comments and suggestions. This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca).

†Department of Economics, University of Toronto, 150 St. George St., Toronto, ON M5S 3G7, Canada; Phone: (416) 978-4479; Email: `martin.burda@utoronto.ca`

‡Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305; Phone: (650) 723-4116; Fax: (650) 725-5702; Email: `mch@stanford.edu`

§Corresponding author. Department of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02142; Email: `jhausman@mit.edu`

1. **Introduction**

Count data arise naturally in a wide range of economic applications. Frequently, the observed event counts are realized in connection with an underlying individual choice from a number of various event alternatives. Examples include household patronization of a set of alternative shopping destinations, utilization rates for various recreational sites, transportation mode frequencies, household urban alternative trip frequencies, or patent counts obtained by different groups within a company, among others. Despite their broad applicability, count data models remain relatively scarce in applications compared to binary or multinomial choice models. For example, in consumer choice analysis of ready-to-eat cereals, instead of assuming independent choices of one product unit that yields highest utility (Nevo, 2001), it is more realistic to allow for multiple purchases over time taking into account the choices among a number of various alternatives that consumers enjoy. In this spirit, a parametric three-level model of demand in the cereal industry addressing variation in quantities and brand choice was analyzed in Hausman (1997).

However, specification and estimation of utility-consistent joint count and multinomial choice models remains a challenge if one wishes to abstain from imposing a number of potentially restrictive simplifying assumptions that may be violated in practice. In this paper we introduce a new flexible random coefficient mixed Poisson model for panel data that seamlessly merges the event count process with the alternative choice selection process under a very weak set of assumptions. Specifically: (*i*) both count and choice processes are embedded in a single random utility framework establishing a direct mapping between the Poisson count intensity λ and the *selected* choice utility; (*ii*) both processes are influenced by unobserved individual heterogeneity; (*iii*) the model framework allows for identification and estimation of coefficients on characteristics that are individual-specific, individual-alternative-specific, and alternative-specific.

The first feature is novel in the literature. Previous studies that link count intensity with choice utility (e.g. Mannering and Hamed, 1990) leave a simplifying dichotomy between these two quantities by specifying the Poisson count intensity parameter λ as a function of expected utility given by an index function of the observables. A key element of the actual choice utility – the idiosyncratic error term ε – never maps into λ. We believe that this link should be preserved since the event of making a trip is intrinsically endogenous to *where* the trip is being taken which in turn is influenced by the numerous factors included in the idiosyncratic term. Indeed, trips are taken because they are taken to their destinations; not to their expected destinations or due to other processes unrelated to choice utility maximization, as implied in the previous literature lacking the first feature. In principle, ε can be included in λ using Bayesian data augmentation. However, such an approach suffers from the curse of dimensionality with increasing number of choices and growing sample size – for example in our application this initial approach proved unfeasible, resulting in failure of convergence of the parameters of interest. As a remedy, we propose an analytical approach that does not rely on data augmentation.

The second feature of individual heterogeneity that enters the model via random coefficients on covariates is rare in the literature on count data. Random effects for count panel data models were introduced by Hausman, Hall, and Griliches (1984) (HHG) in the form of an additive individual-specific stochastic term whose exponential transformation follows the gamma distribution. Further generalizations of HHG regarding the distribution of the additive term are put forward in Greene (2007) and references therein. We take HHG as our natural point of departure. In our model, we specify two types of random coefficient distributions: a flexible nonparametric one on a subset of key coefficients of interest and a parametric one on other control variables, as introduced in Burda, Harding, and Hausman (2008). This feature allows us to uncover clustering structures and other features such as multimodalities in the joint distribution of select variables while preserving model parsimony in controlling for a potentially large number of other relevant variables. At the same time, the number of parameters to be estimated increases much slower in our random coefficient framework than in a possible alternative fixed coefficient framework as N and T grow large. Moreover, the use of choice specific coefficients drawn from a multivariate distribution eliminates the independence of irrelevant alternatives (IIA) at the individual level. Due to its flexibility, our model generalizes a number of popular models such as the Negative Binomial regression model which is obtained as a special case under restrictive parametric assumptions.

Finally, the Poisson panel count level of our model framework allows also the inclusion and identification of individual-specific variables that are constant across choice alternatives and are not identified from the multinomial choice level alone, such as demographic characteristics. However, for identification purposes the coefficients on these variables are restricted to be drawn from the same population across individuals as the Bayesian counterpart of fixed effects[5].

To our knowledge this is the first paper to allow for the nonparametric estimation of preferences in a combined discrete choice and count model. It provides a very natural extension of the discrete choice literature by allowing us to capture the intensity of the choice in addition to the choices made and relate both of these to the same underlying preference structures. At the same time it eliminates undesirable features of older modeling strategies such as the independence of irrelevant alternatives. This approach provides a very intuitive modeling framework within the context of our empirical application where consumers make repeated grocery purchases over several shopping cycles. In this paper we do not aim to capture strategic inter-temporal decision making through the use of dynamic programming techniques which would be relevant in a context with durable goods or strategic interactions between agents. Our focus is on the use of panel data from repeated choice occasions to estimate heterogeneous multimodal preferences. It is our aim to show that real life economic agents

[5]In the Bayesian framework adopted here both fixed and random effects are treated as random parameters. While the Bayesian counterpart of fixed effects estimation updates the posterior distribution of the parameters, the Bayesian counterpart of random effects estimation also updates the posterior distribution of hyperparameters at higher levels of the model hierarchy. For an in-depth discussion on the fixed vs random effects distinction in the Bayesian setting see Rendon (2002).

have complex multimodal preference structures reflecting the underlying heterogeneity of consumer preferences. It is important to account for these irregular features of consumer preferences in policy analysis as they drive different responses at the margin. The resulting enhanced degree of realism is highlighted in our empirical application.

A large body of literature on count data models focus specifically on excess zero counts. Hurdle models and zero-inflated models are two leading examples (Winkelmann, 2008). In hurdle models, the process determining zeros is generally different from the process determining positive counts. In zero-inflated models, there are in general two different types of regimes yielding two different types of zeros. Neither of these features apply to our situation where zero counts are conceptually treated the same way as positive counts; both are assumed to be realizations of the same underlying stochastic process based on the magnitude of the individual-specific Poisson process intensity. Moreover, our model does not fall into the sample selection category since all consumer choices are observed. Instead, we treat such choices as endogenous to the underlying utility maximization process.

Our link of Poisson count intensity to the random utility of choice is driven by flexible individual heterogeneity and the idiosyncratic logit-type error term. As a result, our model formulation leads to a new Poisson mixture model that has not been analyzed in the economic or statistical literature. Various special cases of mixed Poisson distributions have been studied previously, with the leading example of the parametric Negative Binomial model (for a comprehensive literature overview on Poisson mixtures see Karlis and Xekalaki (2005), Table 1). Flexible economic models based on the Poisson probability mass function were analyzed in Terza (1998), Gurmu, Rilstone, and Stern (1999), Munkin and Trivedi (2003), Romeu and Vera-Hernández (2005), and Jochmann and León-González (2004), among others.

Due to the origin of our mixing distribution arising from a latent utility maximization problem of an economic agent, our mixing distribution is a novel convolution of a stochastic count of order statistics of extreme value type 1 distributions. Convolutions of order statistics take a very complicated form and are in general analytically intractable, except for very few special cases. We deal with this complication by recasting the Poisson mixed model as an embedding of infinite convergent sequences of scaled moments of the conditional mixing distribution. We newly derive their form via their cumulant representations and determine the bounds on their rates of numerical convergence. The subsequent analysis is based on Bayesian Markov chain Monte Carlo methodology that partitions the complicated joint model likelihood into a sequence of simple conditional ones with analytically appealing properties utilized in a Gibbs sampling scheme. The nonparametric component of individual heterogeneity is modeled via a Dirichlet process prior specified for a subset of key parameters of interest.

We apply our model to the supermarket trip count data for groceries in a panel of Houston households whose shopping behavior was observed over a 24-month period in years 2004-2005. The detailed AC Nielsen scanner dataset that we utilize contains nearly one million individual entries. In the application, we estimate the nonparametric density of three key variables of interest – price, driving distance, and their interaction – while controlling for a range of consumer demographic characteristics such as age, income, household size, marital and employment status.

The remainder of the paper is organized as follows. Section 2 introduces the mixed Poisson model with its analyzed properties and Section 3 presents an efficient recursive estimation procedure. Section 4 elaborates on the tools of Bayesian analysis used in model implementation, and Section 5 on the issues on identification and posterior consistency. Section 6 discusses the application results and Section 7 concludes.

2. Model

2.1. Poisson Mixtures

In this Section we establish notation and briefly review several relevant concepts and definitions that will serve as the basis for subsequent analysis. In the Poisson regression model the probability of a non-negative integer-valued random variable Y is given by the probability mass function (p.m.f.)

$$P(Y = y) = \frac{\exp(-\lambda)\lambda^y}{y!} \tag{2.1}$$

where $y \in \mathbb{N}_0$ and $\lambda \in \mathbb{R}_+$. For count data models this p.m.f. can be derived from an underlying continuous-time stochastic count process $\{Y(\tau), \tau \geq 0\}$ where $Y(\tau)$ represents the total number of events that have occurred before the time τ. The Poisson assumption stipulates stationary and independent increments for $Y(\tau)$ whereby the occurrence of a random event at a particular instant is independent of time and the number of events that have already taken place. The probability of a unit addition to the count process $Y(\tau)$ within the interval Δ is given by

$$P\{Y(\tau + \Delta) - Y(\tau) = 1\} = \lambda\Delta + o(\Delta)$$

Hence the probability of an event occurring in an infinitesimal time interval $d\tau$ is $\lambda d\tau$ and the parameter λ is thus interpreted as the *intensity* of the count process per unit of time, with the property $E[Y] = \lambda$.

In the temporal context a useful generalization of the base-case Poisson model is to allow for evolution of λ over time by replacing the constant λ with a time-dependent variable $\widetilde{\lambda}(\tau)$. Then the probability of a unit addition to the count process $Y(\tau)$ within the interval Δ is given by

$$P\{Y(\tau + \Delta) - Y(\tau) = 1\} = \widetilde{\lambda}(\tau)\Delta + o(\Delta)$$

Due to the Poisson independence assumption on the evolution of counts, for the *integrated intensity*

$$\lambda_t = \int_{\underline{t}}^{\overline{t}} \widetilde{\lambda}(\tau) d\tau \tag{2.2}$$

it holds that the p.m.f. of the resulting Y on the time interval $t = [\underline{t}, \overline{t})$ is given again by the base-case $P(Y = y)$ in (2.1). In our model $\widetilde{\lambda}(\tau)$ will be assumed constant over small *discrete* equal-length time increments $t_s \equiv [\underline{t_s}, \overline{t_s}) \subset t$, $\underline{t_s} < \overline{t_s}$, $\cup_{s=1}^{S} t_s = t$, with $\widetilde{\lambda}(\tau) = \widetilde{\lambda}_{t_a}$ for $\tau \in t_s$, which will allow us to obtain a convenient form for the integral (2.2) in terms of a summation.

The base-case model further generalizes to a Poisson *mixture* model by turning the parameter λ into a stochastic variable. Thus, a random variable Y follows a *mixed* Poisson distribution, with the mixing density function $g(\lambda)$, if its probability mass function is given by

$$P(Y = y) = \int_0^\infty \frac{\exp(-\lambda)\,\lambda^y}{y!} g(\lambda) d\lambda \tag{2.3}$$

for $y \in \mathbb{N}_0$. Mixing over λ with $g(\lambda)$ provides the model with the flexibility to account for overdispersion typically present in count data. Parametrizing $g(\lambda)$ in (2.3) as the gamma density yields the Negative Binomial model as a special case of (2.3). For a number of existing mixed Poisson specifications applied in other model contexts, see Karlis and Xekalaki (2005), Table 1.

An additional convenient feature of the Poisson process is proportional divisibility of its p.m.f. with respect to subintervals over the interval of observation: the p.m.f. of a count variable Y arising from a Poisson process whose counts y_s are observed on time intervals $[a_s, b_s)$ for $s = 1, \ldots, T$ with $a_s < b_s \leq a_{s+1} < b_{s+1}$ is given by

$$P(\{Y_s = y_s\}_{s=1}^{T}) = \prod_{s=1}^{T} \frac{\exp(-\lambda(b_s - a_s))\,[\lambda(b_s - a_s)]^{y_s}}{y_s!} \tag{2.4}$$

2.2. Model Structure

We develop our model as a two-level mixture. Throughout, we will motivate the model features by referring to our application on grocery store choice and monthly trip count of a panel of households even though the model is quite general. We conceptualize the observed shopping behavior as realizations of a continuous joint decision process on store selection and trip count intensity made by a household representative individual. We will first describe the structure of the model and then lay out the specific technical assumptions on its various components.

An individual i faces various time constraints on the number of trips they can devote for the purpose of shopping. We do not attempt to model such constraints explicitly as households' shopping patterns can be highly irregular – people can make unplanned spontaneous visits of grocery stores or cancel pre-planned trips on a moment's notice due to external factors. Instead, we treat the actual occurrences of shopping trips as realizations of an underlying continuous-time non-homogenous Poisson process whereby the *probability* of taking the trip to store j in the next instant $d\tau$ is a function of the continuous-time shopping intensity $\widetilde{\lambda}_{itj}(\tau)$ which in turn is a function of the maximum of the underlying alternative-specific utility $\max_{j \in \mathcal{J}} \widetilde{U}_{itj}(\tau)$, including its idiosyncratic component. We believe this structure is well suited for our application where each time period t of one month spans a number of potential shopping cycles. The individual is then viewed as making a joint decision on the store choice and the shopping intensity, both driven by the same alternative-specific utility $\widetilde{U}_{itj}(\tau)$. Certain technical aspects of our analysis are simplified by assuming that $\widetilde{U}_{itj}(\tau)$ stays constant within small discrete time intervals, which we make precise further below.

The bottom level of the individual decision process is formed by the utility-maximizing choice among the various store alternatives or the outside option of no shopping at any given instant τ. Here the economic agent i continuously forms their preference ranking of the choice alternatives j in terms of the latent continuous-time *potential* utility $\widetilde{U}_{itj}(\tau)$ at the time instant $\tau \in [\underline{t}, \overline{t})$. At any given τ, $\widetilde{U}_{itj}(\tau) \geq 0$ may or may not result in an actual trip; the maximum $\widetilde{U}_{itj}(\tau)$ determines the probability of the next trip incidence. The outside option of no trip is always taken if $\widetilde{U}_{itj}(\tau) < 0$ in which case $\lambda_{itc} \equiv 0$.

The top level of the individual decision process then models the trip count during the time period $t \equiv [\underline{t}, \overline{t})$ as a realization of a non-homogenous Poisson process with intensity parameter λ_{itc} that is a function of

$$\widetilde{U}_{itc}(\tau) \equiv \max_{j \in \mathcal{J}} \widetilde{U}_{itj}(\tau)$$

formed at the bottom level. The index c denotes the alternative j that maximizes $\widetilde{U}_{itj}(\tau)$. The Poisson intensity parameter λ_{itc} is in our model governed by $\widetilde{U}_{itc}(\tau)$ and hence the trip counts are endogenous to all utility components, *including* the idiosyncratic part of the utility. Following the model description, we will now lay out the mathematical model structure and impose explicit assumptions on all stochastic terms.

2.3. **Utility**

Let $\widetilde{U}_{itj}(\tau)$ denote individual i's latent potential utility of alternatives $j = 1, \ldots, J$ at time $\tau \in t$, given by the following assumption:

ASSUMPTION 1. *$\widetilde{U}_{itj}(\tau)$ takes the linear additively separable form*

$$\widetilde{U}_{itj}(\tau) = \widetilde{\beta}_i' X_{itj}(\tau) + \widetilde{\theta}_i' D_{itj}(\tau) + \widetilde{\varepsilon}_{itj}(\tau)$$

where X_{itj} are key variables of interest, D_{itj} are other relevant (individual-)alternative-specific variables, and $\widetilde{\varepsilon}_{itj}$ is the idiosyncratic term.

In our application of supermarket trip choice and count, X_{itck} is composed of price, driving distance, and their interaction, while the D_{itck} are formed by store indicator variables. The $\widetilde{U}_{itj}(\tau)$ is rationalized as providing the individual's subjective utility evaluation of the choice alternatives at the instant τ as a function of individual-choice characteristics, choice attributes and an idiosyncratic component. As in the logit model, the parameters $\widetilde{\beta}_i$ and $\widetilde{\theta}_i$ are only identified up to a common scale.

We impose the following assumptions on the utility components in the model:

ASSUMPTION 2. *The values of the variables $X_{itj}(\tau)$ and $D_{itj}(\tau)$ are constant on small equal-length time intervals $t_s \equiv [\underline{t_s}, \overline{t_s}) \ni \tau$, with $\underline{t_s} < \overline{t_s}$ and $t = \cup_{s=1}^{S} t_s$, for each $i, t,$ and j.*

ASSUMPTION 3. *The idiosyncratic term $\widetilde{\varepsilon}_{itj}(\tau)$ is drawn at every $\underline{t_s}$ for each i, j, from the extreme value type 1 distribution with density*

$$f_{\widetilde{\varepsilon}}(\widetilde{\varepsilon}) = \exp(-\widetilde{\varepsilon}) \exp(-\exp(-\widetilde{\varepsilon}))$$

and stays constant for the remainder of t_s. The distribution of $\widetilde{\varepsilon}_{itj}(\tau)$ is independent over time.

Assumptions 2 and 3 discretize the evolution of $\widetilde{U}_{itj}(\tau)$ over time which leads to a convenient expression for the ensuing integrated count intensity in terms of summation. Assumption 3 further yields convenient analytical expression for the shares of utility maximizing alternatives.

2.4. **Count Intensity**

We parametrize the link between $\widetilde{\lambda}_{itc}(\tau)$ and $\widetilde{U}_{itc}(\tau)$ as follows:

ASSUMPTION 4. *Let $h : \mathbb{R}_+ \to \mathbb{R}_+$ be a monotonic invertible mapping that takes the form*

$$\begin{aligned} \widetilde{\lambda}_{itc}(\tau) &= h(\widetilde{U}_{itc}(\tau)) \\ &= \gamma' Z_{it}(\tau) + \omega_{1i}\widetilde{\beta}_i' X_{itc}(\tau) + \omega_{2i}\widetilde{\theta}_i' D_{itc}(\tau) + \omega_{3i}\widetilde{\varepsilon}_{itc}(\tau) \\ &= \gamma' Z_{it}(\tau) + \beta_i' X_{itc}(\tau) + \theta_i' D_{itc}(\tau) + \varepsilon_{itc}(\tau) \end{aligned} \tag{2.5}$$

where ω_{1i}, ω_{2i}, and ω_{3i} are unknown factors of proportionality.

The distribution of $\beta_i, \theta_i,$ and γ along with the nature of their independence is given by the following assumption:

ASSUMPTION 5. *The parameter β_i is distributed according to the Dirichlet Process Mixture (DPM) model*

$$\begin{aligned} \beta_i | \psi_i &\sim F(\psi_i) \\ \psi_i | G &\sim G \\ G &\sim DP(\alpha, G_0) \end{aligned}$$

where $F(\psi_i)$ is the distribution of β_i conditional on the hyperparameters ψ_i drawn from a random measure G distributed according to the Dirichlet Process $DP(\alpha, G_0)$ with intensity parameter α and base measure G_0. The parameters θ_i and γ are distributed according to

$$\begin{aligned} \theta_i &\sim N(\underline{\mu}_\theta, \underline{\Sigma}_\theta) \\ \gamma &\sim N(\underline{\mu}_\gamma, \underline{\Sigma}_\gamma) \end{aligned}$$

where $\underline{\mu}_\theta, \underline{\Sigma}_\theta, \underline{\mu}_\gamma$, and $\underline{\Sigma}_\gamma$ are model hyperparameters. The distributions of β_i and θ_i are mutually independent for each i.

For treatment of the Dirichlet Process Mixture model and its statistical properties, see e.g. Neal (2000). In our application, Z_{it} includes various demographic characteristics, while X_{itc} and D_{itc} were described above. Higher utility derived from the *most preferred* alternative thus corresponds to higher count probabilities for that alternative. Conversely, higher count intensity implies higher utility derived from the alternative of choice through the invertibility of h. This isotonic model constraint is motivated as a stylized fact of a choice-count shopping behavior, providing a utility-theoretic interpretation of the count process. We postulate the specific linearly additive functional form of h for ease of implementation. In principle, h only needs to be monotonic for a utility-consistent model framework. Note that we do not need to separately identify ω_{1i}, ω_{2i}, and ω_{3i} from $\widetilde{\beta}_i$, $\widetilde{\theta}_i$, and the variance of $\widetilde{\varepsilon}_{itc}$ in (2.5) for a predictive model of the counts Y_{itc}. In cases where the former are of special interest, one could run a mixed logit model on (2.6), and then use these in our mixed Poisson model for a separate identification of these parameters. Without loss of generality, the scale parameter of the density of $\varepsilon_{itc}(\tau)$ is normalized to unity.

2.5. **Count Probability Function**

The top level of our model is formed by the trip count mechanism based on a non-homogenous Poisson process with the intensity parameter $\lambda_{itc}(\tau)$. We impose the following assumption on the p.m.f. of the trip count stochastic variable $Y_{itc}(\tau)$ as a function of $\lambda_{itc}(\tau)$:

ASSUMPTION 6. *The count variable $Y_{itc}(\tau)$ is distributed according to the Poisson probability mass function*

$$P(Y_{itc}(\tau) = y_{itc}(\tau)) = \frac{\exp(-\widetilde{\lambda}_{itc}(\tau))\widetilde{\lambda}_{itc}(\tau)^{y_{itc}}}{y_{itc}!}$$

This assumption enables us to stochastically complete the model by relating the observed trip counts to the underlying alternative-specific utility via the intensity parameter. The independence of Poisson increments also facilitates evaluation of the integrated probability mass function of the observed counts for each time period t. Let $k = 1, \ldots, Y_{itc}$ denote the index over the observed trips for the individual i during the time period t and let

$$\widetilde{U}_{itck} = \widetilde{\beta}_i' X_{itck} + \widetilde{\theta}_i' D_{itck} + \widetilde{\varepsilon}_{itck} \tag{2.6}$$

denote the associated realizations of $\widetilde{U}_{itc}(\tau)$ for $\tau \in [\underline{t}, \overline{t})$. From the independence of the Poisson increments in the count process evolution of Assumption 6 and Assumption 2, the integrated count intensity (2.3) for the period t becomes

$$\lambda_{itc} = y_{itc}^{-1} \sum_{k=1}^{y_{itc}} \lambda_{itck}$$

with

$$\lambda_{itc} = \gamma' Z_{it} + \beta_i' \overline{X}_{itc} + \theta_i \overline{D}_{itc} + \overline{\varepsilon}_{itc} = \overline{V}_{itc} + \overline{\varepsilon}_{itc} \tag{2.7}$$

for $\lambda_{itck} \geq 0$ where y_{itc} is a given realization of Y_{itc}, $\overline{X}_{itc} = \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} X_{itck}$, $\overline{D}_{itc} = \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} D_{itck}$, and $\overline{\varepsilon}_{itc} = \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} \varepsilon_{itck}$. The individuals in our model are fully rational with respect to the store choice by utility maximization. The possible deviations of the counts y_{it} from the count intensity λ_{it} are stochastic in nature and reflect the various constraints the consumers face regarding the realized shopping frequency.

For alternatives whose selection was not observed in a given period it is possible that their latent utility could have exceeded the latent utilities of other alternatives and been strictly positive for a small fraction of the time period, but the corresponding count intensity was not sufficiently high to result in a unit increase of its count process. Capturing this effect necessitates the inclusion of a latent measurement of the probability of selection associated with each alternative, δ_{itc}. This effect allows us to conduct counterfactual experiments based on the micro-foundations that alter the observables (e.g. price) even for alternatives whose selection is rarely observed in a given sample, and trace the impact of the counterfactual through the latent preference selection process to predictions about expected counts.

For each time period t, denote by δ_{itc} the fraction of that time period over which the alternative c was maximizing the latent utility $\widetilde{U}_{itc}(\tau)$ among other alternatives. By Assumption 3, δ_{itc} is the standard market share of c for the period t given by

$$\delta_{itc} = \frac{\exp\left(\widetilde{V}_{itc}\right)}{\sum_{j=1}^{J} \exp\left(\widetilde{V}_{itj}\right)}$$

where

$$\widetilde{V}_{itc} = \widetilde{\beta}_i' \overline{X}_{itc} + \widetilde{\theta}_i' \overline{D}_{itc}$$

is the deterministic part of the utility function (2.6). With δ_{itc} representing the fractions of the time interval t of unit length, the conditional count probability function is a special case of the proportional Poisson pmf (2.4)

$$P(Y_{itc} = y_{itc} | \lambda_{itc}) = \frac{\exp\left(-\delta_{itc}\lambda_{itc}\right)\left(\delta_{itc}\lambda_{itc}\right)^{y_{itc}}}{y_{itc}!}$$

Note that the count intensity λ_{itc} given by (2.7) is *stochastic* due to the inclusion of the idiosyncratic $\overline{\varepsilon}_{itc}$ and the stochastic specification of β_i and θ_i. Hence, the unconditional count probability mass function is given by

$$P(Y_{itc} = y_{itc}) = \int \frac{\exp\left(-\delta_{itc}\lambda_{itc}\right)\left(\delta_{itc}\lambda_{itc}\right)^{y_{itc}}}{y_{itc}!} g(\lambda_{itc}) d(\lambda_{itc}) \tag{2.8}$$

which is a special case of the generic Poisson mixture model (2.3) with the mixing distribution $g(\lambda_{itc})$ that arises from the underlying individual utility maximization problem. However, $g(\lambda_{itc})$

takes on a very complicated form. From (2.7), each ε_{itck} entering λ_{itc} represents a J-order statistic (i.e. maximum) of the random variables ε_{itjk} with means $V_{itjk} \equiv \gamma' Z_{it} + \beta_i' X_{itjk} + \theta_i D_{itjk}$. The conditional density $g(\overline{\varepsilon}_{itc}|\overline{V}_{itc})$ is thus the convolution of y_{itc} densities of $J-$order statistics which is in general analytically intractable except for some special cases such as for the uniform and the exponential distributions (David and Nagaraja, 2003). The product of $g(\overline{\varepsilon}_{itc}|\overline{V}_{itc})$ and $g(\overline{V}_{itc})$ then yields $g(\lambda_{itc})$.

The stochastic nature of $\lambda_{itc} = \overline{V}_{itc} + \overline{\varepsilon}_{itc}$ as defined in (2.7) is driven by the randomness inherent in the coefficients $\gamma, \theta_i, \beta_i$ and the idiosyncratic component ε_{itck}. Due to the high dimensionality of the latter, we perform integration with respect to ε_{itck} analytically[6] while $\gamma, \theta_i, \beta_i$ are sampled by Bayesian data augmentation. In particular, the algorithm used for nonparametric density estimation of β_i is built on explicitly sampling β_i.

Using the boundedness properties of a probability function and applying Fubini's theorem,

$$
\begin{aligned}
P(Y_{itc} &= y_{itc}) = \int_{\Lambda} f(y_{itc}|\overline{\varepsilon}_{itc}, \overline{V}_{itc}) g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) g(\overline{V}_{itc}) d(\overline{\varepsilon}_{itc}, \overline{V}_{itc}) \\
&= \int_{\mathcal{V}} \int_{\varepsilon} f(y_{itc}|\overline{\varepsilon}_{itc}, \overline{V}_{itc}) g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} g(\overline{V}_{itc}) d\overline{V}_{itc} \qquad (2.9) \\
&= \int_{\mathcal{V}} E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) g(\overline{V}_{itc}) d\overline{V}_{itc}
\end{aligned}
$$

where

$$
E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) = \int_{\varepsilon} f(y_{itc}|\overline{\varepsilon}_{itc}, \overline{V}_{itc}) g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} \qquad (2.10)
$$

Using (2.4), the joint count probability of the observed sample $y = \{y_{itc}\}$ is given by

$$
P(Y = y) = \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{c=1}^{C_{it}} P(Y_{itc} = y_{itc})
$$

3. Analytical Expressions for High Dimensional Integrals

In this Section we derive a new approach for analytical evaluation of $E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc})$ in (2.10). Bayesian data augmentation on $\gamma, \theta_i, \beta_i, \delta$ will be treated in the following Section.

As described above, the conditional mixing distribution $g(\overline{\varepsilon}_{itc}|\overline{V}_{itc})$ takes on a very complicated form. Nonetheless, using a series expansion of the exponential function, the Poisson mixture in (2.8) admits a representation in terms of an infinite sequence of *moments* of the mixing distribution

$$
E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) = \sum_{r=0}^{\infty} \frac{(-1)^r}{y_{itc}! r!} \delta_{itc}^{r+y_{itc}} \eta_w'(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) \qquad (3.1)
$$

[6] In an earlier version of the paper we tried to data-augment also with respect to ε_{itjk} but due to its high dimensionality in the panel this led to very poor convergence properties of the sampler for the resulting posterior.

with $w = y_{itc} + r$, where $\eta'_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ is the w^{th} generalized moment of $\overline{\varepsilon}_{itc}$ about value $\overline{V}_{itc}$ [see the Technical Appendix for a detailed derivation of this result]. Since the subsequent weights in the series expansion (3.1) decrease quite rapidly with r, one only needs to use a truncated sequence of moments with $r \leq R$ such that the last increment to the sum in (3.1) is smaller than some numerical tolerance level δ local to zero in the implementation.

3.1. Recursive Closed-Form Evaluation of Conditional Mixed Poisson Intensity

Evaluation of $\eta'_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ as the conventional probability integrals of powers of $\overline{\varepsilon}_{itc}$ is precluded by the complicated form of the conditional density of $\overline{\varepsilon}_{itc}$.[7] In theory, (3.1) could be evaluated directly in terms of scaled moments derived from a Moment Generating Function (MGF) $M_{\overline{\varepsilon}_{itc}}(s)$ of $\overline{\varepsilon}_{itc}$ constructed as a composite mapping of the individual MGFs $M_{\varepsilon_{itck}}(s)$ of ε_{itck}. However, this approach turns out to be computationally prohibitive [see the Technical Appendix for details][8].

We transform $M_{\varepsilon_{itck}}(s)$ to the the Cumulant Generating Function (CGF) $K_{\varepsilon_{itck}}(s)$ of ε_{itck} and derive the *cumulants* of the composite random variable $\overline{\varepsilon}_{itc}$. We then obtain a new analytical expression for the expected conditional mixed Poisson density in (3.1) based on a highly efficient recursive updating scheme detailed in Theorem 1. Our approach to the cumulant-based recursive evaluation of a moment expansion for a likelihood function may find further applications beyond our model specification.

In our derivation we benefit from the fact that for some distributions, such as the one of $\overline{\varepsilon}_{itc}$, cumulants and the CGF are easier to analyze than moments and the MGF. In particular, a useful feature of cumulants is their linear additivity which is not shared by moments [see the Technical Appendix for a brief summary of the properties of cumulants compared to moments]. Due to their desirable analytical properties, cumulants are used in a variety of settings that necessitate factorization of probability measures. For example, cumulants form the coefficient series in the derivation of higher-order terms in the Edgeworth and saddle-point expansions for densities.

In theory it is possible to express any uncentered moment η' in terms of the related cumulants κ in a closed form via the Faà di Bruno formula (Lukacs (1970), p. 27). However, as a typical attribute of non-Gaussian densities, unscaled moments and cumulants tend to behave in a numerically explosive

[7]We note that Nadarajah (2008) provides a result on the exact distribution of a sum of Gumbel distributed random variables along with the first two moments but the distribution is extremely complicated to be used in direct evaluation of all moments and their functionals given the setup of our problem. This follows from the fact that Gumbel random variables are closed under maximization, i.e. the maximum of Gumbel random variables is also Gumbel, but not under summation which is our case, unlike many other distributions. At the same time, the Gumbel assumption on ε_{itjk} facilitates the result of Lemma 1 in the same spirit as in the logit model.

[8]The evaluation of each additional scaled moment $\eta'_{y_{itc}+r}(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ requires summation over all multi-indices $w_1 + \cdots + w_{y_{itc}} = y_{itc} + r$ for each MC iteration with high run-time costs for a Bayesian nonparametric algorithm.

manner. The same holds when the uncentered moments η' are first converted to the central moments η which are in turn expressed in terms centered expression involving cumulants. In our recursive updating scheme, the explosive terms in the series expansion are canceled out due to the form of the distribution of $\overline{\varepsilon}_{itc}$ which stems from Assumption 3 of extreme value type 1 distribution on the stochastic disturbances $\varepsilon_{itj}(\tau)$ in the underlying individual choice model (2.5). The details are given in the proof of Theorem 1 below.

Recall that the ε_{itck} is an $J-$order statistic of the utility-maximizing choice. As a building block in the derivation of $K_{\varepsilon_{itck}}(s)$ we present the following Lemma regarding the form of the distribution $f_{\max}(\varepsilon_{itck})$ of ε_{itck} that is of interest in its own right.

LEMMA 1. *Under Assumptions 1 and 3, $f_{\max}(\varepsilon_{itck})$ is a Gumbel distribution with mean $\log(\nu_{itck})$ where*

$$\nu_{itck} = \sum_{j=1}^{J} \exp\left[-\left(V_{itck} - V_{itjk}\right)\right]$$

The proof of Lemma 1 in the Appendix follows the approach used in derivation of closed-form choice probabilities of logit discrete choice models (McFadden, 1974). In fact, McFadden's choice probability is equivalent to the zero-th uncentered moment of the $J-$order statistic in our case. However, for our mixed Poisson model we need all the remaining moments except the zero-th one and hence we complement McFadden's result with these cases. We do not obtain closed-form moment expressions directly though. Instead, we derive the CGF $K_{\varepsilon_{itck}}(s)$ of ε_{itck} based on Lemma 1.

Before proceeding further it is worthwhile to take a look at the intuition behind the result in Lemma 1. Increasing the gap $(V_{itck} - V_{itjk})$ increases the probability of lower values of ε_{itck} to be utility-maximizing. As $(V_{itck} - V_{itjk}) \to 0$ the mean of $f_{\max}(\varepsilon_{itck})$ approaches zero. If $V_{itck} < V_{itjk}$ then the mean of $f_{\max}(\varepsilon_{itck})$ increases above 0 which implies that unusually high realizations of ε_{itck} maximized the utility, compensating for the previously relatively low V_{itck}.

We can now derive $K_{\overline{\varepsilon}_{itc}}(s)$ and the conditional mixed Poisson choice probabilities. Using the form of $f_{\max}(\varepsilon_{itck})$ obtained in Lemma 1, the CGF $K_{\varepsilon_{itck}}(s)$ of ε_{itck} is

$$K_{\varepsilon_{itck}}(s) = s \log\left(\nu_{itck}\right) - \log \Gamma(1-s) \tag{3.2}$$

where $\Gamma(\cdot)$ is the gamma function. Let $w \in \mathbb{N}$ denote the order of the moments for which $w = y_{itc} + r$ for $w \geq y_{itc}$. Let $\widetilde{\eta}'_{y_{itc},r-2} = (\widetilde{\eta}'_0, \ldots, \widetilde{\eta}'_{y_{itc}+r-2})^T$ denote a column vector of scaled moments. Let further $\mathbf{Q}_{y_{itc},r} = (Q_{y_{itc},r,q}, \ldots, Q_{y_{itc}r,r-2})^T$ denote a column vector of weights. The recursive scheme for analytical evaluation of (3.1) is given by the following Theorem.

THEOREM 1. *Under Assumptions 1-4 and 6,*

$$E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) = \sum_{r=0}^{\infty} \widetilde{\eta}'_{y_{itc}+r}$$

where

$$\widetilde{\eta}'_{y_{itc}+r} = \delta_{itc}^{y_{itc}+r}\left[\mathbf{Q}^T_{y_{itc},r}\widetilde{\eta}'_{y_{itc},r-2} + (-1)^r r^{-1}\kappa_1\left(\nu_{itc}\right)\widetilde{\eta}'_{y_{itc}+r-1}\right]$$

is obtained recursively for all $r = 0, ..., R$ *with* $\widetilde{\eta}'_0 = y_{itc}!^{-1}$. *Let* $q = 0, ..., y_{itc} + r - 2$. *Then, for* $r = 0$

$$Q_{y_{itc},r,q} = \frac{(y_{itc}+r-1)!}{q!}\left(\frac{1}{y_{itc}}\right)^{y_{itc}+r-q-1}\zeta(y_{itc}+r-q)$$

and for $r > 0$

$$\begin{aligned}
Q_{y_{itc},r,q} &= \frac{1}{r!}B_{y_{itc},r,q} \quad \text{for } 0 \leq q \leq y_{itc} \\
Q_{y_{itc},r,q} &= \frac{1}{r!^{(q-y_{itc})}}B_{y_{itc},r,q} \quad \text{for } y_{itc}+1 \leq q \leq y_{itc}+r-2 \\
B_{y_{itc},r,q} &= (-1)^r\frac{(y_{itc}+r-1)!}{q!}\left(\frac{1}{y_{itc}}\right)^{y_{itc}+r-q-1}\zeta(y_{itc}+r-q) \\
r!^{(q-y_{itc})} &\equiv \prod_{p=q-y_{itc}}^{r} p
\end{aligned}$$

where $\zeta(j)$ *is the Riemann zeta function.*

The proof is provided in the Appendix along with an illustrative example of the recursion for the case where $y_{itc} = 4$. The Riemann zeta function is a well-behaved term bounded with $|\widetilde{\zeta}(j)| < \frac{\pi^2}{6}$ for $j > 1$ and $\widetilde{\zeta}(j) \to 1$ as $j \to \infty$. The following Lemma verifies the desirable properties of the series representation for $E_{\overline{\varepsilon}}f(y_{itc}|\overline{V}_{itc})$ and derives bounds on the numerical convergence rates of the expansion.

LEMMA 2. *Under Assumptions 1-4 and 6, the series representation of* $E_{\overline{\varepsilon}}f(y_{itc}|\overline{V}_{itc})$ *in Theorem 1 is absolutely summable, with bounds on numerical convergence given by* $O(y_{itc}^{-r})$ *as* r *grows large.*

All weight terms in $\mathbf{Q}_{y_{itc},r}$ that enter the expression for $\widetilde{\eta}'_{y_{itc}+r}$ can be computed before the MCMC run by only using the observed data sample since none of these weights is a function of the model parameters. Moreover, only the first cumulant κ_1 of $\overline{\varepsilon}_{itc}$ needs to be updated with MCMC parameter updates as higher-order cumulants are independent of ν_{itck} in Lemma 1 and hence also enter $\mathbf{Q}_{y_{itc},r}$. This feature follows from fact that the constituent higher-order cumulants of the underlying ε_{itck} for $w > 1$ depend purely on the *shape* of the Gumbel distribution $f_{\max}(\varepsilon_{itck})$ which does not change with the MCMC parameter updates in ν_{itck}. It is only the mean $\eta'_1(\varepsilon_{itck}) = \kappa_1(\varepsilon_{itck})$ of $f_{\max}(\varepsilon_{itck})$ which is updated with ν_{itck} shifting the distribution while leaving its shape unaltered. In contrast, all higher-order moments of ε_{itck} and $\overline{\varepsilon}_{itc}$ are functions of the parameters updated in the MCMC run. Hence, our recursive scheme based on cumulants results in significant gains in terms of computational speed relative to any potential moment-based alternatives.

4. Bayesian Analysis

4.1. Semiparametric Random Coefficient Environment

In this Section we briefly discuss the background and rationale for our semiparametric approach to modeling of our random coefficient distributions. Consider an econometric models (or its part) specified by a distribution $F(\cdot;\psi)$, with associated density $f(\cdot;\psi)$, known up to a set of parameters $\psi \in \Psi \subset \mathbb{R}^d$. Under the Bayesian paradigm, the parameters ψ are treated as random variables which necessitates further specification of their probability distribution. Consider further an exchangeable sequence $z = \{z_i\}_{i=1}^n$ of realizations of a set of random variables $Z = \{Z_i\}_{i=1}^n$ defined over a measurable space $(\Phi, \mathcal{D})$ where $\mathcal{D}$ is a σ-field of subsets of Φ. In a parametric Bayesian model, the joint distribution of z and the parameters is defined as

$$Q(\cdot;\psi, G_0) \propto \; F(\cdot;\psi)G_0$$

where G_0 is the (so-called prior) distribution of the parameters over a measurable space $(\Psi, \mathcal{B})$ with $\mathcal{B}$ being a σ-field of subsets of Ψ. Conditioning on the data turns $F(\cdot;\psi)$ into the likelihood function $L(\psi|\cdot)$ and $Q(\cdot;\psi, G_0)$ into the posterior density $K(\psi|G_0, \cdot)$.

In the class of nonparametric Bayesian models[9] considered here, the joint distribution of data and parameters is defined as a mixture

$$Q(\cdot;\psi, G) \propto \int F(\cdot;\psi)G(d\psi)$$

where G is the mixing distribution over ψ. It is useful to think of $G(d\psi)$ as the conditional distribution of ψ given G. The distribution of the parameters, G, is now random which leads to a complete flexibility of the resulting mixture. The model parameters ψ are no longer restricted to follow any given pre-specified distribution as was stipulated by G_0 in the parametric case.

The parameter space now also includes the random infinite-dimensional G with the additional need for a prior distribution for G. The Dirichlet Process (DP) prior (Ferguson, 1973; Antoniak, 1974) is a popular alternative due to its numerous desirable properties. A DP prior for G is determined by two parameters: a distribution G_0 that defines the "location" of the DP prior, and a positive scalar precision parameter α. The distribution G_0 may be viewed as a baseline prior that would be used in a typical parametric analysis. The flexibility of the DP prior model environment stems from allowing G – the actual prior on the model parameters – to stochastically deviate from G_0. The precision parameter α determines the concentration of the prior for G around the DP prior location G_0 and thus measures the strength of belief in G_0. For large values of α, a sampled G is very likely to be close to G_0, and vice versa.

In our model, $\beta = (\beta_1, ..., \beta_N)'$, $\theta = (\theta_1, ..., \theta_N)'$ are vectors of unknown coefficients. The distribution of β_i is modeled nonparametrically in accordance with the model for the random vector z described

[9] A commonly used technical definition of nonparametric Bayesian models are probability models with infinitely many parameters (Bernardo and Smith, 1994).

above. The coefficients on choice specific indicator variables θ_i are assumed to follow a parametric multivariate normal distribution. This formulation for the distribution of β and θ was introduced for a multinomial logit in Burda, Harding, and Hausman (2008) as the "logit-probit" model. The choice specific random normal variables θ form the "probit" element of the model. We retain this specification in order to eliminate the IIA assumption at the individual level. In typical random coefficients logit models used to date, for a given individual the IIA property still holds since the error term is independent extreme value. With the inclusion of choice specific correlated random variables the IIA property no longer holds since a given individual who has a positive realization for one choice is more likely to have a positive realization for another positively correlated choice specific variable. Choices are no longer independent conditional on attributes and hence the IIA property no longer binds. Thus, the "probit" part of the model allows an unrestricted covariance matrix of the stochastic terms in the choice specification.

4.2. **Prior Structure**

Denote the model hyperparameters by W and their joint prior by $k(W)$. From (2.9),

$$P(Y_{itc} = y_{itc}) = \int_{\mathcal{V}} E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) g(\overline{V}_{itc}) d\overline{V}_{itc} \tag{4.1}$$

where $E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc})$ is evaluated analytically in Lemma 1 and Theorem 1. Using an approach analogous to Train's (2003, ch 12) treatment of the Bayesian mixed logit, we data-augment (4.1) with respect to $\gamma, \beta_i, \theta_i$ for all i and t. Thus, the joint posterior takes the form

$$K(W, \overline{V}_{itc}\ \forall i, t) \propto \prod_i \prod_t E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) g(\overline{V}_{itc}|W) k(W)$$

The structure of prior distributions is given in Assumption 5. Denote the respective priors by $k(\beta_i)$, $k(\theta_i)$, $k(\gamma)$. The model hyperparameters W are thus formed by $\{\psi_i\}_{i=1}^N$, G, α, G_0, $\underline{\mu}_\theta$, $\underline{\Sigma}_\theta$, $\underline{\mu}_\gamma$, and $\underline{\Sigma}_\gamma$. Following Escobar and West (1995), inference for α is performed under the prior $\alpha \sim gamma(a, b)$.

4.3. **Sampling**

The Gibbs blocks sampled are specified as follows:

- Draw $\beta_i|\tau, \gamma, \theta$ for each i from $K(\beta_i|\gamma, \theta, Z, X, D) \propto \prod_{t=1}^T E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) k(\beta)$
- Draw θ_i analogously to β_i.
- Draw $\gamma|\beta, \theta, \sigma^2$ from the joint posterior $K(\gamma|\beta, \theta, \sigma^2, Z, X, D) \propto \prod_{i=1}^N \prod_{t=1}^T E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) k(\gamma)$
- Update the DP prior hyperparameters, with the Escobar and West (1995) update for α.
- Update the parameter δ_{itc} as in Burda, Harding, and Hausman (2008).
- Update the remaining hyperparameters based on the identified θ_{ij} (Train (2003), ch 12).

5. Identification and Posterior Consistency

5.1. Identification Issues

Parameter identifiability is generally based on the properties of the likelihood function as hence rests on the same fundamentals in both classical and Bayesian analysis (Kadane, 1974; Aldrich, 2002). Identification of nonparametric random utility models of multinomial choice has recently been analyzed by Berry and Haile (2010). Related aspects of identification of discrete choice models have been treated in Bajari, Fox, Kim, and Ryan (2009), Chiappori and Komunjer (2009), Lewbel (2000) Briesch, Chintagunta, and Matzkin (2010), and Fox and Gandhi (2010). In our model likelihood context, a proof of the identifiability of infinite mixtures of Poisson distributions is derived from the uniqueness of the Laplace transform (Teicher, 1960; Sapatinas, 1995).

With the use of informative priors the Bayesian framework can address situations where certain parameters are empirically partially identified or unidentified. Our data exhibits a certain degree of customer loyalty: many i never visit certain types of stores j (denote the subset of θ_{ij} on these by θ_{ij}^n). In such cases θ_{ij}^n is not identified. Two different low values of θ_{ij}^n can yield the same observation whereby the corresponding store j is not selected by i. In the context of a random coefficient model, such cases are routinely treated by a common informative prior $\theta_i \sim N(\mu, \Sigma)$ that shrinks θ_{ij}^n to the origin. In our model, the informativeness of the common prior is never effectively invoked since θ_i are coefficients on store indicator variables. The sampled values of θ_{ij}^n are inconsequential since they multiply the zero indicators of the non-selected stores, thus dropping out of the likelihood function evaluation. Hence b_θ and Σ_θ are computed only on the basis of the identified θ_{ij}. This result precludes any potential influence of the unidentified dimensions of θ_{ij} on the model likelihood via b_θ and Σ_θ. The unidentified dimensions of θ_{ij} are shrunk to zero with the prior $k(b_\theta, \Sigma_\theta)$. As the time dimension T grows, all dimensions of θ_{ij} become eventually empirically identified, diminishing the influence of the prior in the model.

5.2. Posterior Consistency

The importance of posterior consistency stems from the desire to be able to correctly identify the data generating mechanism with an increasing sample size. Even though consistency is purely a large sample property, an inconsistent posterior is often an indication of invalid inference even for moderate sample sizes. Moreover, consistency can be shown to be equivalent with agreement among Bayesians with different sets of priors (Diaconis and Freedman, 1986b). If posterior consistency holds, then for convex parameter spaces such as the space of densities which induces convex neighborhoods, the posterior mean gives another consistent estimator.

In a seminal paper, Doob (1949) showed that under i.i.d. observations and identifiability conditions, the posterior is consistent everywhere except possibly on a null set with respect to the prior, almost surely. Almost sure posterior consistency in various models, including examples of inconsistency, has been extensively discussed by Diaconis and Freedman (1986b,a, 1990). These authors note that in the nonparametric context such null set may be topologically very large and include cases of interest. Consequently, they warn against careless use of priors. We show consistency of the posterior density of β, which forms the nonparametric component of our model under the Dirichlet process prior, by verifying the conditions necessary for invoking an extension of a consistency result by Schwartz (1965) based on Ghosh and Ramamoorthi (2003) and Ghosal (2010). The extension applies Schwartz's result to a sieve constructed on the parameter space. We verify that that the prior probability mass assigned to a complement of the sieve space is exponentially small, and that the model sieve satisfies an entropy condition binding the rate of growth of the sieve space. The result is summarized in the following Theorem.

THEOREM 2. *Under the Assumptions 1-6 and two additional regularity conditions given in the Appendix, the marginal posterior density of β is:*
(a) weakly consistent and
(b) strongly or L_1-consistent
at the true distribution of the observables as the sample size tends to infinity.

The relevant definitions of weak and strong consistency as well as the proof the Theorem are given in the Appendix.

6. Application

In this section we introduce a stylized yet realistic empirical application of our method to consumers' joint decision process over the number of shopping trips to a grocery story and the choice of the grocery stores where purchases are made. Shopping behavior has recently been analyzed by economists in order to better understand the process through which consumers search for their preferred options and the interaction between consumer choices and demographics responsible for various search frictions. Thus, Aguiar and Hurst (2007) and Harding and Lovenheim (2010) focus on demographics limiting search behavior, while Broda, Leibtag, and Weinstein (2009) measure inequality in consumption.

6.1. Data description

The data used in this study is similar to that used by Burda, Harding, and Hausman (2008) and is a subsample of the 2004-2005 Nielsen Homescan panel for the Houston area over 24 months. We use an unbalanced panel of consumer purchases augmented by a rich set of demographic characteristics for the households. The data is collected from a sample of individuals who joined the Nielsen panel and identified at Universal Product Code (UPC) level for each product.

The data is obtained through a combination of store scanners and home scanners which were provided to individual households. Households are required to upload a detailed list of their purchases with identifying information weekly and are rewarded through points which can be used to purchase merchandise in an online store. The uploaded data is merged with data obtained directly from store scanners in participating stores. For each household, Nielsen records a rich set of demographics as well as the declared place of residence. Note that while the stated aim of the Nielsen panel is to obtain a nationally representative sample, certain sampling distortions remain. For example, over 30% of the Nielsen sample is collected from individuals who are registered as not employed i.e. unemployed or not in the labor force.

The shopping trips are recorded weekly and we decided to aggregate them to monthly counts. This avoids excessive sparsity and provides a natural recurring cycle over which consumers purchase groceries. We only observe information about the products purchased and do not observe information about the order in which they were purchased or route traveled by the consumer. We excluded from the sample a very small number of outliers such as households who appeared to live more than 200 miles away from the stores at which they shopped. We also dropped from the sample households with fewer than 4 months of observations, and households that shop every month only at one store type in order to discard cases of degenerate variation. The total number of individual data entries use for estimation was thus 491,706 for a total 660 households.

We consider each household as having a choice among 6 different stores (H.E.B., Kroger, Randall's, Walmart, PantryFoods[10] and "Other"). The last category includes any remaining stores adhering to the standard grocery store format (excluding club stores and convenience stores) that the households visit. Most consumers shop in at least two different stores in any given month. The mean number of trips per month conditional on shopping at a given store for the stores in the sample is: H.E.B. (3.10), Kroger (3.61), Randall's (2.78), Walmart (3.49), PantryFoods (3.08), Other (3.34). The histogram in Figure 1 summarizes the frequency of each trip count for the households in the sample.

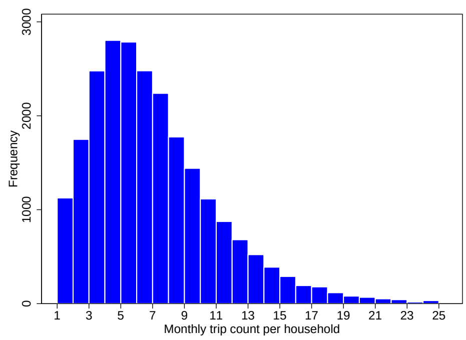

FIGURE 1. Histogram of the monthly total number of trips to a store per month for the households in the sample.

[10]PantryFoods stores are owned by H.E.B. and are typically limited-assortment stores with reduced surface area and facilities.

We employ three key variables: log *price*, which corresponds to the price of a basket of goods in a given store-month; log *distance,* which corresponds to the estimated driving distance for each household to the corresponding supermarket; and their interaction.

Product Category	Weight
Bread	0.0804
Butter and Margarine	0.0405
Canned Soup	0.0533
Cereal	0.0960
Chips	0.0741
Coffee	0.0450
Cookies	0.0528
Eggs	0.0323
Ice Cream	0.0663
Milk	0.1437
Orange Juice	0.0339
Salad Mix	0.0387
Soda	0.1724
Water	0.0326
Yogurt	0.0379

TABLE 1. Product categories and the weights used in the construction of the price index.

In order to construct the *price* variable we first normalize observations from the price paid to a dollars/unit measure, where unit corresponds to the unit in which the idem was sold. Typically, this is ounces or grams. For bread, butter and margarine, coffee, cookies and ice cream we drop all observations where the transaction is reported in terms of the number of unit instead of a volume or mass measure. Fortunately, few observations are affected by this alternative reporting practice. We also verify that only one unit of measurement was used for a given item. Furthermore, for each produce we drop observations for which the price is reported as being outside two standard deviations of the standard deviations of the average price in the market and store over the periods in the sample.

We also compute the average price for each product in each store and month in addition to the total amount spent on each produce. Each product's weight in the basket is computed as the total amount spent on that product across all stores and months divided by the total amount spent across all stores and months. We look at a subset of the total product universe and focus on the following product categories: bread, butter and margarine, canned soup cereal, chips, coffee, cookies, eggs, ice cream, milk, orange juice, salad mix, soda, water, yogurt. The estimated weights are given in Table 1.

For a subset of the products we also have available directly comparable product weights as reported in the CPI. As shows in Table 2 the scaled CPI weights match well with the scaled produce weights derived from the data. The price of a basket for a given store and month is thus the sum across product of the average price per unit of the product in that store and month multiplied by the product weight.

Product Category	2006 CPI Weight	Scaled CPI Weight	Scaled Product Weight
Bread	0.2210	0.1442	0.1102
Butter and Margarine	0.0680	0.0444	0.0555
Canned Soup	0.0860	0.0561	0.0730
Cereal	0.1990	0.1298	0.1315
Coffee	0.1000	0.0652	0.0617
Eggs	0.0990	0.0646	0.0443
Ice Cream	0.1420	0.0926	0.0909
Milk	0.2930	0.1911	0.1969
Soda	0.3250	0.2120	0.2362

TABLE 2. Comparison of estimated and CPI weights for matching product categories.

In order to construct the *distance* variable we employ GPS software to measure the arc distance from the centroid of the census tract in which a household lives to the centroid of the zip code in which a store is located.[11] For stores in which a household does not shop in the sense that we don't observe a trip to this store in the sample, we take the store at which they would have shopped to be the store that has the smallest arc distance from the centroid of the census tract in which the household lives out of the set of stores at which people in the same market shopped. If a household shops at a store only intermittently, we take the store location at which they would have shopped in a given month to be the store location where we most frequently observe the household shopping when we do observe them shopping at that store. The store location they would have gone to is the mode location of the observed trips to that store. Additionally, we drop households that shop at a store more than 200 miles from their reported home census tract.

6.2. **Results**

First we consider the estimated densities of our key parameters of interest on log price, log distance and their interaction. Plots of these marginal densities are presented in Figure 2 with summary statistics in Table 3. Plots of joint densities of pairs of these parameters (log price vs log distance, log price vs interaction, log distance vs interaction) are given in Figure 3. All plots attest to the existence of several sizeable preference clusters of consumers. This finding of multi-modality is potentially quite important for policy analysis as it allows for a more complex reaction to changes in

[11]Our data does not capture occasional grocery store trips along the way from a location other than one's home.

prices, say. The nonparametric estimation procedure developed in this paper is particularly potent at uncovering clustering in the preference space of the consumers thus highlighting the extent to which consumers make trade-offs between desirable characteristics in the process of choosing where to make their desired purchase.

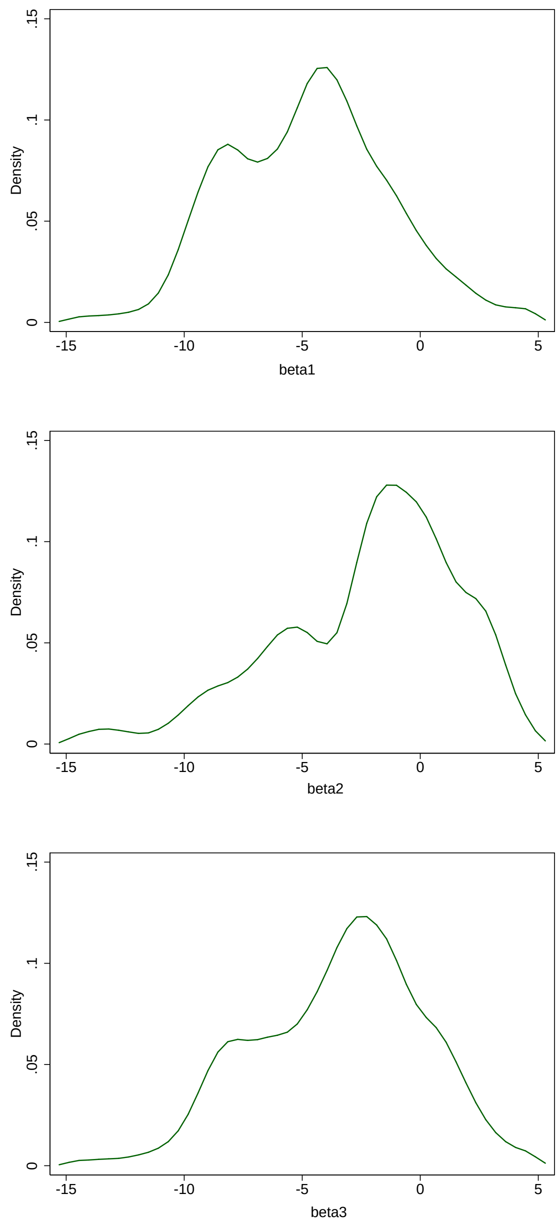

FIGURE 2. Density of MC draws of the individual-specific coefficients on price β_{i1} (top), distance β_{i2} (middle), and their interaction β_{i3} (bottom) variables.

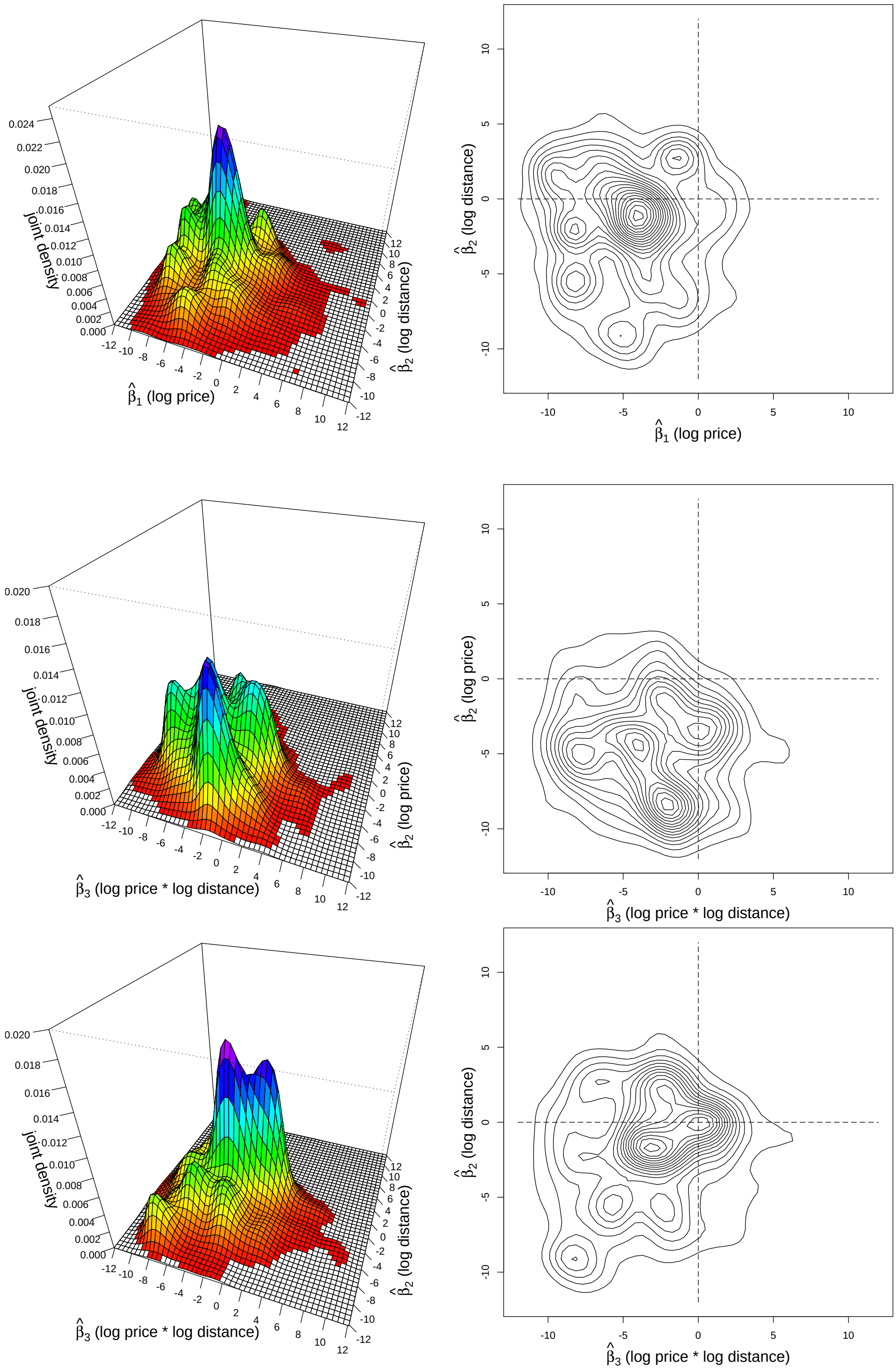

FIGURE 3. Joint density of MC draws of β_1 vs β_2 β_1 vs β_3, and β_2 vs β_3.

While most consumers react negatively in terms of shopping intensity to higher price and increased travel distance, they nevertheless do appear to be making trade-offs in their responsiveness to the key variables. The top graph pair in Figure 3 shows several distinct preference clusters in the price-distance preference space. Moreover, consumers become even more price sensitive with increased travel distance (bottom graph).[12]

Two animations capturing the evolution of the joint density of individual-specific coefficients on log price β_{1i} and log distance β_{2i} in a window sliding over the domain of the interaction coefficient β_{3i}. A 3D animation is available at http://dl.getdropbox.com/u/716158/pde867b.wmv while a 2D contour animation is at http://dl.getdropbox.com/u/716158/pde867bct.wmv. The trend in the movement of the joint density along the diagonal confirms that aversion to higher prices enters both coefficients for the whole range of aversion to higher distances.

For comparison purposes, we also ran a parametric benchmark model where the parameters of interest β_i were distributed according to a multivariate Normal density, with common alternative-specific indicator variables. This specification is by far the most widely used specification for the "mixed logit model" which allows for random preference distributions. The means of the estimated densities of β_i were statistically not different from zero and the unimodal parametric density precluded the discovery of interesting clusters of preferences found in Figure 3. The Hausman test applied to the means of the estimated densities strongly rejected the null of mean equivalence (p-value less than 0.001), suggesting that imposing the Normal density on the model for β_i distorts the central tendency of the estimates.

Now let us turn our attention to the coefficients on the demographic variables which are identified in the model through the variation in trip counts for different consumers and stores. These coefficients relate directly to common economic intuitions on the importance of household demographics in driving search costs (Harding and Lovenheim, 2010). The posterior mean, median, standard deviation and 90% Bayesian Credible Sets (BCS, corresponding to the 5^{th} and 95^{th} quantiles) for coefficients γ on demographic variables are presented in Table 4 with their marginal counterparts incorporating the price interaction effects in Table 5, under the heading Selective Flexible Poisson Mixture.

Faced with higher prices, households decrease their volumes of goods purchased in their stores of choice more than proportionately (the price elasticity of demand was estimated as -1.389 for our sample). This phenomenon is characteristic of all households, albeit differing in its extent over price

[12] In our previous work (Burda, Harding, and Hausman, 2008) we estimated a relevant parametric benchmark case for the price vs distance trade-off for each individual household separately. Even though such benchmark estimates on short panels contained a small bias, the multimodality in preferences and price vs distance tradeoff became apparent once these individual estimates were brought together and smoothed over using conventional kernels. These features qualitatively confirm the nature of consumer preferences uncovered in this paper.

levels over household types with the high income households exhibiting the lowest propensity to reduce quantity.

The base category for which all demographic indicator variables were equal to zero is formed by households with more than one member but without children, with low income, low level of education, employed, and white. Following a price increase, virtually all other household attributes increase the shopping count intensity for the stores of choice relative to the base category (Table 5), with one exception being the demographic attribute non-white whose coefficient was not statistically significant from the base category. This phenomenon reflects the higher search intensity exhibited by households shopping around various store alternatives selecting the favorably-priced items and switching away from food categories with relatively higher price tags. Equivalently, households are able to take advantage of sales on individual items across different store types. The extent to which this happens differs across various demographic groups (Tables 4 and 5). Households that feature the high age (65+) and high total household income attribute (50K+) intensify their search most when faced with higher prices. The search effect further increases at higher price levels for high age while abating for high income. The opportunity cost of time relative to other household types is a likely factor at play. Middle age (more than 40 but less than 65 average for the household head) and middle income (25K to 50K) attributes substantially increase search intensity at the same rate regardless of the absolute price level. Households with children, Hispanic, and unemployed, attributes exhibit similar behavior albeit to a lower degree. The higher education (college and higher) and singleton (one-member households) categories do not exhibit any additional reaction to higher prices beyond the effects their other demographic attributes.

Table 6 shows the posterior means, medians, standard deviations and 90% Bayesian Credible Sets for the means of b_θ and Table 7 for the variances Σ_θ of the store indicator variable coefficients θ_{ij}. In the absence of an overall fixed model intercept while including all store indicator variables, these coefficients play the role of random intercepts for each household. Hence, interpretation of their estimated distributions needs to be conducted in the context of other model variables. Kroger, Walmart, and Other have the lowest store effect means but also relatively large variances of the means, reflecting the diversity of preferences regarding the shopping intensity at these store types on the part of the pool of households. Pantry Foods and Randalls exhibits the highest store effects which likely stems from their business concept featuring an array of specialty departments, once their price levels – the highest among all the store types – have been controlled for. H.E.B. belongs to the mid-range category in terms of store shopping intensity preference. The store effects also exhibit various interesting covariance patterns (Table 7). While H.E.B. and Pantry Foods exhibit a low covariance, Randalls and Pantry Foods exhibit relatively high covariance, which is explained by the fact that their marketing approach targets similar customer segments.

Figure 4 shows the kernel density estimate of the MC draws of the Dirichlet process latent class model hyperparameter α. The sharp curvature on the posterior density of α against a diffuse prior suggests that the large data sample exerts a high degree of influence in the parameter updating process. When we restricted α to unity in a trial run, the number of latent classes in the Dirichlet mixture fell to about a third of its unrestricted count, yielding a lower resolution on the estimated clusters of β_i; the demographic parameters γ were smaller on the base demographic variables while larger on the interactions with price, suggesting an estimation bias under the restriction. Hence, sampling α under a diffuse prior does play a role for the accuracy of estimation results.

Figure 5 features the density of the number of latent classes obtained at each MC step in the Dirichlet process latent class sampling algorithm (left) and their ordered average membership counts (right). Thus, the density of β estimate is on average composed of about 74 mixture components, while this number oscillates roughly in the range of 65 to 85 components with the exact data-driven count determined by the algorithm in each MC step. However, only about 20 latent class components contain a substantial number of individuals associated with them at any given MC step while the remainder is composed of low-membership or myopic classes. This flexible mixture can be contrasted with the parametric benchmark Normal model which is by construction composed of one component lacking any adaptability properties. In earlier work we have also conducted a sensitivity analysis by restricting the number of mixture components and found little variation in the results once the number of components exceeds 20.

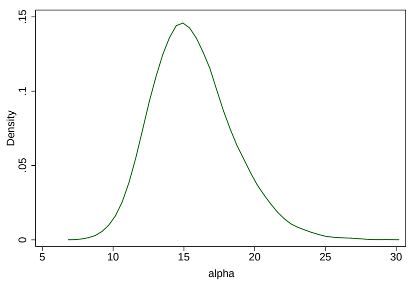

FIGURE 4. Density of MC draws of α.

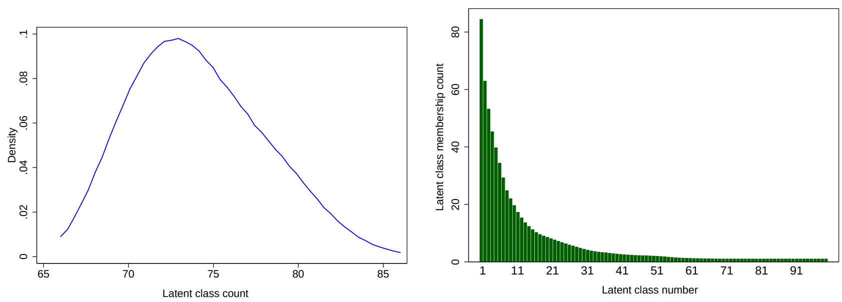

FIGURE 5. Density of the number of latent classes (left) and ordered average latent class membership counts (right).

7. Counterfactual Welfare Experiment

In order to illustrate the usefulness of our model in applications, we conducted a welfare experiment that seeks to evaluate the amount of compensating variation resulting from a price increase of a particular choice alternative. We chose Walmart for the price change since its large share of the market will affect a wide spectrum of consumers across all demographic categories. In the welfare experiment, we ask the following question: after a price increase for a given choice alternative (Walmart), how much additional funding do we need to provide to each person each month in order to achieve the same level of utility regarding both the choice and count intensity as they exhibited before the price increase? In 2006 the state of Maryland passed such a tax for Walmart, but the tax was not implemented on US constitutional grounds.

For every i, t the expected count intensity is

$$\begin{aligned} E[\lambda_{it}] &= \sum_{c=1}^{J} \delta_{itc} E[\lambda_{itc}] \\ &= \sum_{c=1}^{J} \delta_{itc} \int \lambda_{itc} g(\lambda_{itc}) d\lambda_{itc} \end{aligned}$$

and conditionally on $\overline{V}_{it} = (\overline{V}_{it1}, \ldots, \overline{V}_{itJ})'$ we have

$$\begin{aligned} E[\lambda_{it}|\overline{V}_{it}] &= \sum_{c=1}^{J} \delta_{itc} \int (\overline{V}_{itc} + \overline{\varepsilon}_{itc}) g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} \\ &= \sum_{c=1}^{J} \delta_{itc} \eta_1(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) \end{aligned}$$

where $\eta_1(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ is the first uncentered moment of $\overline{\varepsilon}_{itc}$, i.e.

$$\begin{aligned} E[\lambda_{it}|\overline{V}_{itc}] &= \sum_{c=1}^{J} \delta_{itc} \left(\overline{V}_{itc} + \frac{1}{y_{itck}} \sum_{k=1}^{y_{itck}} \log(\nu_{itck}) + \gamma_e \right) \\ \nu_{itck} &= \sum_{j=1}^{J} \exp(-\overline{V}_{itck} + \overline{V}_{itjk}) \end{aligned}$$

Since in the counterfactual experiment we do not directly observe the new hypothetical y_{itck} and the corresponding changes in $\overline{V}_{itck}$ within the time period t, instead of $\frac{1}{y_{itck}} \sum_{k=1}^{y_{itck}} \log(\nu_{itck})$ we use $\log(\nu_{itc})$ with $\nu_{itc} = \sum_{j=1}^{J} \exp(-\overline{V}_{itc} + \overline{V}_{itj})$ where $\overline{V}_{itc}$ and $\overline{V}_{itj}$ are quantities set in the counterfactual experiment to be constant throughout the given time period t. We also assume that following the price increase the demand for the affected alternative will initially fall at the same rate than the price hike. Thus,

$$E[\lambda_{it}|\overline{V}_{itc}] = \sum_{c=1}^{J} \delta_{itc} \left(\overline{V}_{itc} + \log\left(\sum_{j=1}^{J} \exp(-\overline{V}_{itc} + \overline{V}_{itj}) \right) + \gamma_e \right)$$

where

$$
\begin{aligned}
-\overline{V}_{itc} + \overline{V}_{itj} &= \beta_{i1}\left(-\ln Price_{itc} + \ln Price_{itj}\right) + \beta_{i2}\left(-\ln Dist_{itc} + \ln Dist_{itj}\right) \\
&\quad + \beta_{i3}\left(-\ln Price_{itc} \times \ln Dist_{itc} + \ln Price_{itj} \times \ln Dist_{itj}\right) + \theta_{ic} - \theta_{ij}
\end{aligned}
$$

The difference in count intensities after the price increase is

$$
\begin{aligned}
E[\lambda_{it}^{new}|\overline{V}_{it}^{new}] - E[\lambda_{it}^{old}|\overline{V}_{it}^{old}] &= \sum_{c=1}^{J} \delta_{itc}^{new} E[\lambda_{itc}^{new}|\overline{V}_{itc}^{new}] - \sum_{c=1}^{J} \delta_{itc}^{old} E[\lambda_{itc}^{old}|\overline{V}_{itc}^{old}] \\
&= \sum_{c=1}^{J} \delta_{itc}^{new}\left(\overline{V}_{itc}^{new} + \log\left(\sum_{j=1}^{J} \exp(-\overline{V}_{itc}^{new} + \overline{V}_{itj}^{new})\right) + \gamma_e\right) \\
&\quad - \sum_{c=1}^{J} \delta_{itc}^{old}\left(\overline{V}_{itc}^{old} + \log\left(\sum_{j=1}^{J} \exp(-\overline{V}_{itc}^{old} + \overline{V}_{itj}^{old})\right) + \gamma_e\right) \\
&= \Delta_{it}
\end{aligned}
\tag{7.1}
$$

The answer to our welfare question is then obtained from the solution to the equation

$$
-\Delta_{it} = \sum_{c=1}^{J} \delta_{itc}^{new*} E[\lambda_{itc}^{new*}|\overline{V}_{itc}^{new*}] - \sum_{c=1}^{J} \delta_{itc}^{new} E[\lambda_{itc}^{new}|\overline{V}_{itc}^{new}] \tag{7.2}
$$

where $new*$ denotes the state with additional funding that compensates for the change in prices and brings individual's shopping intensity on the original level. We evaluate Δ_{it} in (7.1) and then, using a univariate fixed point search, we solve (7.2) for the additional funds, split proportionately by δ_{itj} among the choice alternatives, that are required to compensate for the price increase, yielding the required compensating variation.

The results (Table 8) reveal that on average consumers require about six dollars a month, or just under a hundred dollars a year, to compensate for the change of their shopping habits after a 10% Walmart price increase, nine dollars a month (or just over one hundred dollars a year) following a 20% increase, and eleven dollars a month (or hundred and thirty dollars a year) following a 30% increase. The average sample household monthly expenditure on grocery store food is \$170 of which \$84 is spent in Walmart. A 10% Walmart price increase thus translates to about 7% of Walmart (or 4% overall) increased grocery cost to consumers in terms of compensating variation, reflecting the fact that individuals are able to switch to other store alternatives. For higher Walmart price increases the relative cost to consumers rises less than proportionately since the elevated Walmart prices approach and exceed the prices in competing stores and store switching becomes relatively cheaper. In contrast, the parametric benchmark Normal model predicts much higher welfare costs, reaching to about three times the amounts of the semiparametric Poisson model. The Hausman test applied to the means of the estimated compensating variations rejected the null of mean equivalence (p-value less than 0.001). We find that the benchmark Normal model finds this unrealistic policy response because of its use of a Normal distribution and imposition of the IIA property, at the individual level.

8. Conclusion

In this paper we have introduced a new mixed Poisson model with a stochastic count intensity parameter that incorporates flexible individual heterogeneity via endogenous latent utility maximization among a range of alternative choices. Our model thus combines latent utility maximization of an alternative selection process within a count data generating process under relatively weak assumptions. The distribution of individual heterogeneity is modeled semiparametrically, relaxing the independence of irrelevant alternatives at the individual level. The coefficients on key variables of interest are assumed to be distributed according to an infinite mixture model while other individual-specific parameters are distributed parametrically, allowing for uncovering local details in the former while preserving parameter parsimony with respect to the latter. To overcome the curse of dimensionality in our model, we develop a closed-form analytical expression for a central conditional expectation term and implement it using an efficient recursive algorithm based on higher-order moment expansion of the Poisson conditional intensity function. We also include a proof of posterior consistency.

Our model is applied to the supermarket visit count data in a panel of Houston households. The results reveal an interesting mixture of consumer clusters in their preferences over the price-distance trade-off, and their joint density for diverse levels of the variable interaction. Various household demographic types exhibit differing patterns of search intensity adjustment when faced with higher prices. The opportunity cost of time and the income effect appear as plausible explanations behind the observed shopping patterns. The results of a counterfactual welfare experiment that subjects Walmart to 10% to 30% price increase suggest that consumers need to be compensated by one to two hundred dollars per year on average in order to achieve the original levels of utility.

	Selective Flexible Poisson Mixture				Normal Poisson			
Parameter	Mean	Median	S.D.	90% BCS	Mean	Median	S.D.	90% BCS
β_{i1} (log price)	-4.05	-4.16	3.67	(-10.13,0.75)	-1.06	-1.05	1.43	(-2.68,0.20)
β_{i2} (log distance)	-1.23	-0.31	3.82	(-9.48,2.68)	0.55	0.55	1.09	(-1.02,1.80)
β_{i3} (interaction)	-2.58	-2.16	3.86	(-9.58,2.25)	-0.87	-0.87	1.18	(-2.49,0.38)

TABLE 3. Summary statistics of β_i draws.

	Selective Flexible Poisson Mixture				Normal Poisson			
Variable	Mean	Median	S.D.	90% BCS	Mean	Median	S.D.	90% BCS
Singleton	0.90	0.69	0.20	(0.64, 1.30)	1.89	1.92	0.25	(1.41, 2.27)
Children	1.04	0.85	0.10	(0.88, 1.25)	0.24	0.23	0.36	(-0.35, 0.77)
Non-white	0.20	0.35	0.13	(-0.03, 0.41)	-0.58	-0.64	0.38	(-1.17, 0.09)
Hispanic	0.98	0.41	0.28	(0.43, 1.37)	1.33	1.32	0.31	(0.82, 1.82)
Unemployed	0.66	0.46	0.20	(0.32, 0.98)	-0.61	-0.63	0.43	(-1.32, 0.15)
Education	0.81	0.68	0.15	(0.59, 1.11)	0.79	0.77	0.23	(0.40, 1.18)
Middle Age	0.86	1.12	0.12	(0.68, 1.09)	1.56	1.62	0.30	(0.91, 1.98)
High Age	1.97	1.91	0.18	(1.67, 2.28)	2.67	2.63	0.46	(1.97, 3.42)
Middle Income	2.15	2.41	0.12	(1.95, 2.36)	1.08	1.06	0.25	(0.64, 1.46)
High Income	2.53	2.61	0.20	(2.20, 2.89)	1.33	1.36	0.19	(0.96, 1.62)
$logP\times$ Singleton	-1.63	-1.84	0.42	(-2.36,-0.95)	-3.01	-3.08	0.69	(-3.95,-1.91)
$logP\times$ Children	-0.66	-0.45	0.44	(-1.35,-0.07)	1.14	1.09	0.70	(-0.24, 2.12)
$logP\times$ Non-white	0.01	0.24	0.37	(-0.42, 0.86)	4.93	5.51	1.24	(2.55, 6.43)
$logP\times$ Hispanic	0.78	0.76	0.28	(0.34, 1.31)	0.97	1.06	0.51	(0.05, 1.69)
$logP\times$ Unemployed	1.92	1.36	0.44	(1.40, 2.67)	3.74	3.96	0.63	(2.39, 4.48)
$logP\times$ Education	-1.16	-0.75	0.39	(-1.72,-0.60)	-0.69	-0.86	0.61	(-1.58, 0.38)
$logP\times$ M Age	4.19	2.60	0.69	(3.10, 5.15)	-0.67	-0.97	0.92	(-1.77, 1.38)
$logP\times$ H Age	2.03	1.33	0.18	(1.68, 2.27)	-3.39	-2.96	1.16	(-5.22,-1.97)
$logP\times$ M Income	0.02	0.44	0.51	(-0.88, 0.84)	1.66	1.66	0.45	(0.82, 2.48)
$logP\times$ H Income	-0.30	-0.29	0.42	(-1.16, 0.34)	1.29	1.36	0.65	(0.09, 2.35)

TABLE 4. Coefficients γ on demographic variables. $logP$ denotes interaction term with price.

	Selective Flexible Poisson Mixture				Normal Poisson			
Variable	Mean	Median	S.D.	90% BCS	Mean	Median	S.D.	90% BCS
Singleton	0.33	0.31	0.13	(0.12,0.60)	0.85	0.85	0.18	(0.54,1.17)
Children	0.81	0.81	0.15	(0.55,1.05)	0.64	0.59	0.24	(0.27,1.06)
Non-white	0.20	0.20	0.12	(-0.02,0.43)	1.12	1.14	0.19	(0.76,1.39)
Hispanic	1.26	1.30	0.24	(0.74,1.58)	1.67	1.66	0.25	(1.25,2.08)
Unemployed	1.33	1.30	0.24	(0.97,1.79)	0.68	0.70	0.30	(0.14,1.14)
Education	0.41	0.39	0.17	(0.11,0.72)	0.55	0.54	0.17	(0.28,0.86)
Middle Age	2.31	2.30	0.20	(1.95,2.64)	1.32	1.33	0.16	(1.03,1.59)
High Age	2.67	2.66	0.17	(2.41,2.93)	1.50	1.50	0.19	(1.17,1.79)
Middle Income	2.16	2.16	0.17	(1.86,2.46)	1.65	1.66	0.20	(1.31,1.98)
High Income	2.42	2.44	0.15	(2.12,2.64)	1.78	1.85	0.23	(1.36,2.10)

TABLE 5. Marginal coefficients γ on demographic variables.

Parameter	Mean	Median	Std.Dev.	90% BCS
$b_{\theta 1}$ (HEB)	7.672	7.708	0.301	(7.093, 8.112)
$b_{\theta 2}$ (Kroger)	5.651	5.838	1.016	(3.931, 7.127)
$b_{\theta 3}$ (Randalls)	8.225	8.365	0.937	(6.607, 9.369)
$b_{\theta 4}$ (Walmart)	4.830	4.915	0.877	(3.380, 6.177)
$b_{\theta 5}$ (Pantry Foods)	11.79	11.681	0.486	(11.168,12.679)
$b_{\theta 6}$ (other)	4.689	4.897	0.808	(3.331, 5.739)

TABLE 6. Means b_θ of distributions of store indicator variable coefficients θ_i.

Parameter	Mean	Median	Std.Dev.	90% BCS
$\Sigma_{\theta 1 \theta 1}$ (HEB)	2.205	2.199	0.142	(1.983, 2.450)
$\Sigma_{\theta 1 \theta 2}$ (HEB & Kroger)	-0.008	-0.009	0.084	(-0.146, 0.130)
$\Sigma_{\theta 1 \theta 3}$ (HEB & Randalls)	0.594	0.594	0.101	(0.428, 0.763)
$\Sigma_{\theta 1 \theta 4}$ (HEB & Walmart)	0.211	0.210	0.079	(0.078, 0.345)
$\Sigma_{\theta 1 \theta 5}$ (HEB & Pantry Foods)	-1.105	-1.090	0.144	(-1.366,-0.889)
$\Sigma_{\theta 1 \theta 6}$ (HEB & other)	-0.877	-0.872	0.109	(-1.067,-0.710)
$\Sigma_{\theta 2 \theta 2}$ (Kroger)	1.992	1.988	0.134	(1.779, 2.224)
$\Sigma_{\theta 2 \theta 3}$ (Kroger & Randalls)	0.139	0.137	0.087	(-0.001, 0.283)
$\Sigma_{\theta 2 \theta 4}$ (Kroger & Walmart)	0.060	0.059	0.073	(-0.060, 0.180)
$\Sigma_{\theta 2 \theta 5}$ (Kroger & Pantry Foods)	-0.169	-0.168	0.087	(-0.312,-0.028)
$\Sigma_{\theta 2 \theta 6}$ (Kroger & other)	0.086	0.084	0.081	(-0.047, 0.221)
$\Sigma_{\theta 3 \theta 3}$ (Randalls)	2.209	2.200	0.178	(1.933, 2.516)
$\Sigma_{\theta 3 \theta 4}$ (Randalls & Walmart)	-0.002	-0.003	0.076	(-0.126, 0.125)
$\Sigma_{\theta 3 \theta 5}$ (Randalls & Pantry Foods)	0.559	0.541	0.154	(0.341, 0.862)
$\Sigma_{\theta 3 \theta 6}$ (Randalls & other)	0.392	0.391	0.096	(0.236, 0.555)
$\Sigma_{\theta 4 \theta 4}$ (Walmart)	1.747	1.743	0.113	(1.569, 1.941)
$\Sigma_{\theta 4 \theta 5}$ (Walmart & Pantry Foods)	0.331	0.331	0.087	(0.186, 0.472)
$\Sigma_{\theta 4 \theta 6}$ (Walmart & other)	0.038	0.037	0.076	(-0.084, 0.162)
$\Sigma_{\theta 5 \theta 5}$ (Pantry Foods)	2.311	2.303	0.154	(2.074, 2.585)
$\Sigma_{\theta 5 \theta 6}$ (Pantry Foods & other)	-0.410	-0.409	0.096	(-0.572,-0.256)
$\Sigma_{\theta 6 \theta 6}$ (other)	2.180	2.173	0.138	(1.967, 2.421)

TABLE 7. Covariances Σ_θ of distributions of store indicator variable coefficients θ_i

Walmart price increase	10%		20%		30%	
Variable	Mean	Normal Mean	Mean	Normal Mean	Mean	Normal Mean
Pooled sample	5.96	17.76	8.57	22.12	10.6	26.36
Singleton = 1	9.84	13.05	12.22	17.12	12.9	21.03
Singleton = 0	4.93	19.12	7.61	23.56	9.98	27.89
Children = 1	3.88	12.50	5.58	16.71	7.68	20.73
Children = 0	6.49	19.11	9.34	23.48	11.31	27.75
Non-white = 1	8.78	21.62	9.71	26.28	8.78	30.81
Non-white = 0	5.27	17.00	8.27	21.31	11.10	25.48
Hispanic = 1	3.70	12.76	7.35	16.33	12.49	20.16
Hispanic = 0	6.18	18.41	8.68	22.84	10.44	27.11
Unemployed = 1	8.22	14.80	7.76	19.21	3.86	23.25
Unemployed = 0	5.79	18.07	8.63	22.43	11.11	26.69
Education = 1	7.01	17.29	9.11	21.39	11.04	25.67
Education = 0	4.77	18.17	7.95	22.76	10.11	26.95
Med Age = 1	5.31	18.17	7.41	22.57	8.96	26.77
Med Age = 0	6.71	17.05	9.93	21.36	12.67	25.67
High Age = 1	9.37	15.40	13.0	19.98	16.35	24.72
High Age = 0	4.59	18.41	6.77	22.72	8.45	26.83
Med Income = 1	3.31	13.55	4.99	16.79	8.81	19.72
Med Income = 0	6.88	19.92	9.77	24.80	11.20	29.64
High Income = 1	5.40	19.26	7.71	23.39	8.19	27.63
High Income = 0	6.64	16.18	9.63	20.78	13.61	25.02

TABLE 8. Monthly compensating variation in dollar amounts of compensating variation for individuals in different demographic categories: comparison of the Selective Flexible Poisson Mixture model and a parametric benchmark Normal Poisson model.

9. Appendix

9.1. Implementation Notes

The estimation results along with auxiliary output are presented below. All parameters were sampled by running 30,000 MCMC iterations, saving every fifth parameter draw, with a 10,000 burn-in phase. The entire run took about 24 hours of wall clock time on a 2.2 GHz AMD Opteron unix machine using the fortran 90 Intel compiler version 11.0. In applying Theorem 1, the Riemann zeta function $\zeta(j)$ was evaluated using a fortran 90 module Riemann_zeta. [13]

In the application, we used $F(\psi_i) = N(\mu_\beta^{\phi_i}, \Sigma_\beta^{\phi_i})$ with hyperparameters $\mu_\beta^{\phi_i}$ and $\Sigma_\beta^{\phi_i}$, with ϕ_i denoting a latent class label, drawn as $\mu_\beta^{\phi_i} \sim N(\underline{\mu}_\beta, \underline{\Sigma}_\beta)$, $\Sigma_\beta^{\phi_i} \sim IW(\underline{\Sigma}_\beta^{\phi_i}, v_{0\Sigma_\beta})$, $\underline{\mu}_\beta = 0$, $\underline{\Sigma}_\beta = diag(100)$, $\underline{\Sigma}_\beta^{\phi_i} = diag(1/2)$, and $v_{0\Sigma_\beta} = \dim\beta + 10$. Since the resulting density estimate should be capable of differentiating sufficient degree of local variation, we imposed a flexible upper bound on the variance of each latent class: if any such variance exceeded double the prior on $\Sigma_\beta^{\phi_i}$, the strength of the prior belief expressed as $v_{0\Sigma_\beta}$ was raised until the constraint was satisfied. This left the size of the latent classes to vary freely up to double the prior variance. This structure gives the means of individual latent classes of β_i sufficient room to explore the parameter space via the diffuse $\underline{\Sigma}_\beta$ while ensuring that each latent class can be well defined from its neighbor via the (potentially) informative $\underline{\Sigma}_\beta^{\phi_i}$ and $v_{0\Sigma_\beta}$ which enforce a minimum degree of local resolution in the nonparametrically estimated density of β_i. The priors on the hyperparameters μ_θ and Σ_θ of $\theta_i \sim N(\mu_\theta, \Sigma_\theta)$ were set to be informative due to partial identification of θ_i, as discussed above, with $\mu_\theta \sim N(\underline{\mu}_\theta, \underline{\Sigma}_\theta)$, $\underline{\mu}_\theta = 0$, $\underline{\Sigma}_\theta = diag(5)$, $\Sigma_\theta \sim IW(\underline{\Sigma}_\theta, v_{0\Sigma_\theta})$, and $v_{0\Sigma_\theta} = \dim(\theta) + 10$. Such prior could guide the θ_is that were empirically unidentified while leaving the overall dominating weight to the parameters themselves. We left the prior on γ completely diffuse without any hyperparameter updates since γ enters as a "fixed effect" parameter. The curvature on the likelihood of γ is very sharp as γ is identified and sampled for the entire panel.

The starting parameter values for γ, β and θ were obtained from the base-case parametric Poisson model estimated in Stata, with a $N(0, 0.1)$ random disturbance applied to β_i and θ_i. Initially, each individual was assigned their own class in the DPM algorithm. The RW-MH updates were automatically tuned using scale parameters to achieve the desired acceptance rates of approximately 0.3 (for a discussion, see e.g. p. 306 in Train, 2003). All chains appear to be mixing well and having converged. In contrast to frequentist methods, the draws from the Markov chain converge in distribution to the true posterior distribution, not to point estimates. For assessing convergence, we use the criterion given in Allenby, Rossi, and McCulloch (2005) characterizing draws as having

[13] The module is available in file r_zeta.f90 at http://users.bigpond.net.au/amiller/ converted to f90 by Alan Miller. The module was adapted from from DRIZET in the MATHLIB library from CERNLIB, K.S. Kolbig, Revision 1.1.1.1 1996/04/01, based on Cody, W.J., Hillstrom, K.E. & Thather, H.C., 'Chebyshev approximations for the Riemann zeta function', Math. Comp., vol.25 (1971), 537-547.

the same mean value and variability over iterations. Plots of individual chains are not reported here due to space limitations but can be provided on request.

9.2. Proof of Lemma 1: Derivation of $f_{\max}(\varepsilon_{itck})$

We have

$$\begin{aligned}
F_j(\varepsilon_{itck}) &= \exp\left\{-\exp\left[-\left(\varepsilon_{itck}+V_{itck}-V_{itjk}\right)\right]\right\}\\
f_c(\varepsilon_{itck}) &= \exp\left[-\left(\varepsilon_{itck}+V_{itck}-V_{itck}\right)\right]\exp\left\{-\exp\left[-\left(\varepsilon_{itck}+V_{itck}-V_{itck}\right)\right]\right\}
\end{aligned}$$

Therefore

$$\begin{aligned}
f_{\max}(\varepsilon_{itck}) &\propto \prod_{j\neq c}\exp\left\{-\exp\left[-\left(\varepsilon_{itck}+V_{itck}-V_{itjk}\right)\right]\right\}\\
&\quad \times\exp\left(-\varepsilon_{itck}\right)\exp\left\{-\exp\left(-\varepsilon_{itck}\right)\right\}\\
&= \exp\left\{-\sum_{j=1}^{J}\exp\left[-\left(\varepsilon_{itck}+V_{itck}-V_{itjk}\right)\right]\right\}\exp\left(-\varepsilon_{itck}\right)\\
&= \exp\left\{-\exp\left(-\varepsilon_{itck}\right)\sum_{j=1}^{J}\exp\left[-\left(V_{itck}-V_{itjk}\right)\right]\right\}\exp\left(-\varepsilon_{itck}\right)\\
&\equiv \widetilde{f}_{\max}(\varepsilon_{itck})
\end{aligned}$$

Defining $z_{itck} = \exp\left(-\varepsilon_{itck}\right)$ for a transformation of variables in $f_{\max}(\varepsilon_{itck})$, we note that the resulting $\widetilde{f}^{e}_{\max}(z_{itck})$ is an exponential density kernel with the rate parameter

$$\nu_{itck} = \sum_{j=1}^{J}\exp\left[-\left(V_{itck}-V_{itjk}\right)\right]$$

and hence ν_{itck} is the factor of proportionality for both probability kernels $\widetilde{f}^{e}_{\max}(z_{itck})$ and $\widetilde{f}_{\max}(\varepsilon_{itck})$ which can be shown as follows:

$$\begin{aligned}
\int_{-\infty}^{\infty}\nu_{itck}\widetilde{f}_{\max}(\varepsilon_{itck})d\varepsilon_{itck} &= \nu_{itck}\int_{-\infty}^{\infty}\exp\left\{-\exp\left(-\varepsilon_{itck}\right)\nu_{itck}\right\}\exp\left(-\varepsilon_{itck}\right)d\varepsilon_{itck}\\
&= \nu_{itck}\int_{-\infty}^{0}\exp\left\{-z_{itck}\nu_{itck}\right\}d(-z_{itck})\\
&= \nu_{itck}\int_{0}^{\infty}\exp\left\{-z_{itck}\nu_{itck}\right\}d(z_{itck})\\
&= \frac{\nu_{itck}}{\nu_{itck}}\exp\left\{-z_{itck}\nu_{itck}\right\}\Big|_{0}^{\infty}\\
&= 1
\end{aligned}$$

Hence

$$\begin{aligned}
f_{\max}(\varepsilon_{itck}) &= \exp\left(\log(\nu_{itck})\right)\widetilde{f}_{\max}(\varepsilon_{itck})\\
&= \exp\left\{-\exp\left(-(\varepsilon_{itck}-\log(\nu_{itck}))\right\}\exp\left(-(\varepsilon_{itck}-\log(\nu_{itck})\right)\right.
\end{aligned}$$

which is Gumbel with mean $\log(\nu_{itck})$ (as opposed to 0 for the constituent $f(\varepsilon_{ttjk})$) or exponential with rate ν_{itck} (as opposed to rate 1 for the constituent $f(z_{itck})$).

Note that the derivation of $f_{\max}(\varepsilon_{itck})$ is only concerns the distribution of ε_{itjk} and is independent of the form of λ_{it}.

9.3. **Proof of Theorem 1: Derivation of Conditional Choice Probabilities**

The proof proceeds by first deriving an analytical expression for the generalized w-th moment $\eta'_w(\overline{\varepsilon}_{itc};\overline{V}_{itc})$ in (3.1) via its composite cumulant representation, and then uses its structure to arrive at a closed-form expression for the desired full integral term $E_{\overline{\varepsilon}}f(y_{itc}|\overline{V}_{itc})$ in (3.1).

Let $\kappa(\overline{\varepsilon}_{itc};\overline{V}_{itc})$ denote the uncentered cumulant of $\overline{\varepsilon}_{itc}$ with mean $\overline{V}_{itc}$ while $\kappa(\overline{\varepsilon}_{itc})$ denotes the centered cumulant of $\overline{\varepsilon}_{itc}$ around its mean. Uncentered moments η'_w and cumulants κ_w of order w are related by the following formula:

$$\eta'_w = \sum_{q=0}^{w-1} \binom{w-1}{q} \kappa_{w-q}\eta'_q$$

where $\eta'_0 = 1$ (Smith, 1995). We adopt it by separating the first cumulant $\kappa_1(\overline{\varepsilon}_{itc};\overline{V}_{itc})$ in the form

$$\begin{aligned}\eta'_w(\overline{\varepsilon}_{itc};\overline{V}_{itc}) = \sum_{q=0}^{w-2} \frac{(w-1)!}{q!(w-1-q)!}\kappa_{w-q}(\overline{\varepsilon}_{itc};\overline{V}_{itc})\eta'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \\ +\kappa_1(\overline{\varepsilon}_{itc};\overline{V}_{itc})\eta'_{w-1}(\overline{\varepsilon}_{itc};\overline{V}_{itc})\end{aligned} \tag{9.1}$$

since only the first cumulant is updated during the MCMC run, as detailed below. Using the definition of $\overline{\varepsilon}_{itc}$ as

$$\overline{\varepsilon}_{itc} = \frac{1}{y_{itc}}\sum_{k=1}^{y_{itc}} \varepsilon_{itck}$$

by the linear additivity property of cumulants, conditionally on $\overline{V}_{itc}$, the centered cumulant $\kappa_w(\overline{\varepsilon}_{itc})$ of order w can be obtained by

$$\begin{aligned}\kappa_w(\overline{\varepsilon}_{itc}) &= \kappa_w\left(\frac{1}{y_{itc}}\sum_{k=1}^{y_{itc}}\varepsilon_{itck}\right) \\ &= \left(\frac{1}{y_{itc}}\right)^w \kappa_w\left(\sum_{k=1}^{y_{itc}}\varepsilon_{itck}\right) \\ &= \left(\frac{1}{y_{itc}}\right)^w \sum_{k=1}^{y_{itc}}\kappa_w\left(\varepsilon_{itck}\right)\end{aligned} \tag{9.2}$$

[see the Technical Appendix for a brief overview of properties of cumulants].

From Lemma 1, ε_{itck} is distributed Gumbel with mean $\log(\nu_{itck})$. The cumulant generating function of Gumbel distribution is given by

$$K_{\varepsilon_{itck}}(s) = \mu s - \log\Gamma(1-\sigma s)$$

and hence the centered cumulants $\kappa_w(\varepsilon_{itck})$ of ε_{itck} take the form

$$\begin{aligned}\kappa_w(\varepsilon_{itck}) &= \frac{d^w}{ds^w}K_{\varepsilon_{itck}}(s)\Big|_{s=0} \\ &= \frac{d^w}{ds^w}\left(\mu s - \log\Gamma(1-s)\right)\Big|_{s=0}\end{aligned}$$

yielding for $w = 1$

$$\kappa_1(\varepsilon_{itck}) = \log(\nu_{itck}) + \gamma_e \tag{9.3}$$

where $\gamma_e = 0.577...$ is the Euler's constant, and for $w > 1$

$$\begin{aligned}\kappa_w(\varepsilon_{itck}) &= -\frac{d^w}{ds^w}\log\Gamma(1-s)\Big|_{s=0} \\ &= (-1)^w\psi^{(w-1)}(1) \\ &= (w-1)!\zeta(w) \qquad (9.4)\end{aligned}$$

where $\psi^{(w-1)}$ is the polygamma function of order $w-1$ given by

$$\psi^{(w-1)}(1) = (-1)^w(w-1)!\zeta(w)$$

where $\zeta(w)$ is the Riemann zeta function

$$\zeta(w) = \sum_{p=0}^{\infty}\frac{1}{(1+p)^w} \tag{9.5}$$

(for properties of the zeta function see e.g. Abramowitz and Stegun (1964)).

Note that the higher-order cumulants for $w > 1$ are not functions of the model parameters $(\gamma, \beta_i, \theta_i)$ contained in ν_{itck}. Thus only the first cumulant $\kappa_1(\varepsilon_{itck})$ is subject to updates during the MCMC run. We exploit this fact in our recursive updating scheme by pre-computing all higher-order scaled cumulant terms, conditional on the data, before the MCMC iterations, resulting in significant run-time gains.

Substituting for $\kappa_w(\varepsilon_{itck})$ from (9.3) and (9.4) in (9.2) yields

$$\begin{aligned}\kappa_1(\overline{\varepsilon}_{itc}) &= \frac{1}{y_{itc}}\sum_{k=1}^{y_{itc}}\kappa_1(\varepsilon_{itck}) \\ &= \frac{1}{y_{itc}}\sum_{k=1}^{y_{itc}}\log(\nu_{itck}) + \gamma_e\end{aligned}$$

and for $w > 1$

$$\begin{aligned}\kappa_w(\overline{\varepsilon}_{itc}) &= \sum_{k=1}^{y_{itc}} \kappa_w(\varepsilon_{itck}) \\ &= \left(\frac{1}{y_{itc}}\right)^{w-1} (w-1)!\zeta(w)\end{aligned}$$

For the uncentered cumulants, conditionally on $\overline{V}_{itc}$, we obtain

$$\kappa_1(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) = \overline{V}_{itc} + \kappa_1(\overline{\varepsilon}_{itc}) \tag{9.6}$$

while for $w > 1$

$$\kappa_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) = \kappa_w(\overline{\varepsilon}_{itc}) \tag{9.7}$$

[see the Technical Appendix for details on the additivity properties of cumulants.]

Substituting for $\kappa_1(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ and $\kappa_{w-q}(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ with $w > 1$ from from (9.6) and (9.7) in (9.1), canceling the term $(w-i-1)!$, yields

$$\begin{aligned}\eta'_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) &= \sum_{q=0}^{w-2} \frac{(w-1)!}{q!} \left(\frac{1}{y_{itc}}\right)^{w-q-1} \zeta(w-q)\eta'_q(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) \\ &\quad + [\overline{V}_{itc} + \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} \log(\nu_{itck}) + \gamma_e]\eta'_{w-1}(\overline{\varepsilon}_{itc}; \overline{V}_{itc})\end{aligned} \tag{9.8}$$

Note that the appearance (and hence the possibility of cancellation) of the explosive term $(w-q-1)!$ in both in the recursion coefficient and in the expression for all the cumulants κ_{w-q} is a special feature of Gumbel distribution which further adds to its analytical appeal.

Let

$$\widetilde{\eta}'_{r+y_{itc}}(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) = \frac{(-1)^r}{r!y_{itc}!} \delta_{itc}^{r+y_{itc}} \eta'_{r+y_{itc}}(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) \tag{9.9}$$

denote the scaled raw moment obtained by scaling $\eta'_{r+y_{itc}}(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ in (9.8) with $(-1)^r \delta_{itc}^{r+y_{itc}}/(r!y_{itc}!)$. Summing the expression (9.9) over $r = 1, \ldots, \infty$ would now give us the desired series representation for (2.10). The expression (9.9) relates unscaled moments expressed in terms of cumulants to scaled ones. We will now elaborate on a recursive relation based on (9.9) expressing higher-order scaled cumulants in terms of their lower-order scaled counterparts. The recursive scheme will facilitate fast and easy evaluation of the series expansion for (2.10).

The intuition for devising the scheme weights is as follows. If the simple scaling term $(-1)^r/(r!y_{itc}!)$ were to be used for calculating $\eta'_{r+y_{itc}}(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ in (9.8), the former would be transferred to $\eta'_{r+y_{itc}+1}$ along with a new scaling term for higher r in any recursive evaluation of higher-order scaled moments. To prevent this compounding of scaling terms, it is necessary to adjust scaling for each w appropriately.

Let

$$\widetilde{\eta}_0' = \frac{1}{y_{itc}!}\eta_0'$$

with $\eta_0' = 1$ and let

$$B_{y_{itc},r,q} = (-1)^r \frac{(y_{itc}+r-1)!}{q!}\left(\frac{1}{y_{itc}}\right)^{y_{itc}+r-q-1}\zeta(y_{itc}+r-q)$$

Let $p = 1,\ldots,r+y_{itc}$, distinguishing three different cases:

(1) For $p \leq y_{itc}$ the summands in $\widetilde{\eta}_p'$ from (9.8) do not contain r in their scaling terms. Hence to scale η_p' to a constituent term of $\widetilde{\eta}_{r+y_{itc}}'$ these need to be multiplied by the full factorial $1/r!$ which then appears in $\widetilde{\eta}_{r+y_{itc}}'$. In this case,

$$Q_{y_{itc},r,q} = \frac{1}{r!}B_{y_{itc},r,q}$$

(2) For $p > y_{itc}$ (i.e. $r > 0$) but $p \leq r+y_{itc}-2$ the summands in $\widetilde{\eta}_p'$ already contain scaling by $1/(q-y_{itc})!$ transferred from lower-order terms. Hence these summands are additionally scaled only by $1/r!^{(q-y_{itc})}$ where $r!^{(q-y_{itc})} \equiv \prod_{c=q-y_{itc}}^{r} c$ in order to result in the sum $\widetilde{\eta}_p'$ that is fully scaled by $1/r!$. In this case,

$$Q_{y_{itc},r,q} = \frac{1}{r!^{(q-y_{itc})}}B_{y_{itc},r,q}$$

(3) The scaling term on the first cumulant $\kappa_1(\overline{\varepsilon}_{itc};\overline{V}_{itc})$ is r^{-1} for each $p = 1,\ldots,y_{itc}+r$. Through the recursion up to $\widetilde{\eta}_{y_{itc}+r}'$ the full scaling becomes $r!^{-1}$. In this case,

$$Q_{y_{itc},r,q} = \frac{1}{r}(-1)^r$$

Denoting $\widetilde{\eta}_{y_{itc},r-2}' = (\widetilde{\eta}_0',\ldots,\widetilde{\eta}_{y_{itc}+r-2}')^T$ and $\mathbf{Q}_{y_{itc},r-2} = (Q_{y_{itc},r,q},\ldots,Q_{y_{itc}r,,r-2})^T$ the recursive updating scheme

$$\widetilde{\eta}_{y_{itc}+r}' = \delta_{itc}^{r+y_{itc}}\left[\mathbf{Q}_{y_{itc},r-2}^T\widetilde{\eta}_{y_{itc},r-2}' + (-1)^r r^{-1}\kappa_1(\nu_{itc})\,\widetilde{\eta}_{y_{itc}+r-1}'\right]$$

yields the expression

$$\begin{aligned}\widetilde{\eta}_{y_{itc}+r}'(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \quad &= \quad (-1)^r\delta_{itc}^{r+y_{itc}}\sum_{q=0}^{y_{itc}+r-2}\frac{(y_{itc}+r-1)!}{r!q!}+\frac{\zeta(y_{itc}+r-q)}{y_{itc}^{y_{itc}+r-q-1}}\widetilde{\eta}_q'(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \\ &\quad +(-1)^r\delta_{itc}^{r+y_{itc}}\frac{1}{r!}[\overline{V}_{itc}+\frac{1}{y_{itc}}\sum_{k=1}^{y_{itc}}\log(\nu_{itck})+\gamma_e]\widetilde{\eta}_{y_{itc}+r-1}'(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \end{aligned} \tag{9.10}$$

for a generic $y_{itc}+r$ which is equivalent to our target term in (9.9) that uses the substitution for $\eta_w'(\overline{\varepsilon}_{itc};\overline{V}_{itc})$ from (9.8). However, unlike the unscaled moments $\eta_q'(\overline{\varepsilon}_{itc};\overline{V}_{itc})$, the terms on the right-hand side of (9.10) are bounded and yield a convergent sum over $r = 1,\ldots,\infty$ required for evaluation of (2.10), as verified in Lemma 2. An illustrative example of our recursive updating scheme for $y_{itc} = 4$ follows.

9.4. Illustrative Example of Recursive Updating:

Let $\xi = (\beta, \theta, \gamma)$. Each column in the following table represents a vector of terms that sum up in each column to obtain the scaled moment $\widetilde{\eta}'_p$, up to $\delta_{itc}^{r+y_{itc}}$. This example is for $y_{itc} = 4$, with $r_k = k$.

r	q	$p:1$	2	3	4	5	6	7	8
0	0	$\kappa_1(\xi)\widetilde{\eta}'_0$	$B_{4,0,0}\widetilde{\eta}'_0$	$B_{4,0,0}\widetilde{\eta}'_0$	$B_{4,0,0}\widetilde{\eta}'_0$	$\frac{1}{r_1}B_{4,1,0}\widetilde{\eta}'_0$	$\frac{1}{r_1}\frac{1}{r_2}B_{4,2,0}\widetilde{\eta}'_0$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}B_{4,3,0}\widetilde{\eta}'_0$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}\frac{1}{r_4}B_{4,4,0}\widetilde{\eta}'_0$
0	1	$=\widetilde{\eta}'_1$	$\kappa_1(\xi)\widetilde{\eta}'_1$	$B_{4,0,1}\widetilde{\eta}'_1$	$B_{4,0,1}\widetilde{\eta}'_1$	$\frac{1}{r_1}B_{4,1,1}\widetilde{\eta}'_1$	$\frac{1}{r_1}\frac{1}{r_2}B_{4,2,1}\widetilde{\eta}'_1$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}B_{4,3,1}\widetilde{\eta}'_1$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}\frac{1}{r_4}B_{4,4,1}\widetilde{\eta}'_1$
0	2		$=\widetilde{\eta}'_2$	$\kappa_1(\xi)\widetilde{\eta}'_2$	$B_{4,0,2}\widetilde{\eta}'_2$	$\frac{1}{r_1}B_{4,1,2}\widetilde{\eta}'_2$	$\frac{1}{r_1}\frac{1}{r_2}B_{4,2,2}\widetilde{\eta}'_2$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}B_{4,3,2}\widetilde{\eta}'_2$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}\frac{1}{r_4}B_{4,4,2}\widetilde{\eta}'_2$
0	3			$=\widetilde{\eta}'_3$	$\kappa_1(\xi)\widetilde{\eta}'_3$	$\frac{1}{r_1}B_{4,1,3}\widetilde{\eta}'_3$	$\frac{1}{r_1}\frac{1}{r_2}B_{4,2,3}\widetilde{\eta}'_3$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}B_{4,3,3}\widetilde{\eta}'_3$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}\frac{1}{r_4}B_{4,4,3}\widetilde{\eta}'_3$
0	4				$=\widetilde{\eta}'_4$	$\frac{1}{r_1}\kappa_1(\xi)\widetilde{\eta}'_4$	$\frac{1}{r_1}\frac{1}{r_2}B_{4,2,4}\widetilde{\eta}'_4$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}B_{4,3,4}\widetilde{\eta}'_4$	$\frac{1}{r_1}\frac{1}{r_2}\frac{1}{r_3}\frac{1}{r_4}B_{4,4,4}\widetilde{\eta}'_4$
1	5					$=\widetilde{\eta}'_5$	$\frac{1}{r_2}\kappa_1(\xi)\widetilde{\eta}'_5$	$\frac{1}{r_2}\frac{1}{r_3}B_{4,3,5}\widetilde{\eta}'_5$	$\frac{1}{r_2}\frac{1}{r_3}\frac{1}{r_4}B_{4,4,5}\widetilde{\eta}'_5$
2	6						$=\widetilde{\eta}'_6$	$\frac{1}{r_3}\kappa_1(\xi)\widetilde{\eta}'_6$	$\frac{1}{r_3}\frac{1}{r_4}B_{4,4,5}\widetilde{\eta}'_6$
3	7							$=\widetilde{\eta}'_7$	$\frac{1}{r_4}\kappa_1(\xi)\widetilde{\eta}'_7$
4	8								$=\widetilde{\eta}'_8$

Note on color coding: The terms in green are pre-computed and stored in a memory array before the MCMC run. The one term in violet is updated with each MCMC draw. The terms in red are computed recursively by summing up the columns above and updating the red term in the following column, respectively, within each MCMC step.

9.5. Proof of Lemma 2

From (9.10) we have

$$
\begin{aligned}
\widetilde{\eta}'_{y_{itc}+r}(\bar{\varepsilon}_{itc}; \overline{V}_{itc}) &= \sum_{q=0}^{y_{itc}+r-2} O(q!^{-1}) O(y_{i1}^{-r}) O(1) \widetilde{\eta}'_q(\bar{\varepsilon}_{itc}; \overline{V}_{itc}) \\
&\quad + O(r!^{-1}) O(1) \widetilde{\eta}'_{y_{itc}+r-1}(\bar{\varepsilon}_{itc}; \overline{V}_{itc})
\end{aligned}
$$

as r grows large, with dominating term $O(y_{i1}^{-r})$. For $y_{itc} > 1$, $O(y_{i1}^{-r}) = o(1)$. For $y_{itc} = 1$, using (9.10) in (2.10), for R large enough to evaluate $E_{\bar{\varepsilon}} f(y_{itc}|\overline{V}_{itc})$ with a numerical error smaller than some tolerance level, switch the order of summation between r and q to obtain a triangular array

$$
\begin{aligned}
E_{\bar{\varepsilon}} f(y_{itc} &= 1|\overline{V}_{itc}) \approx \sum_{r=0}^{R} \widetilde{\eta}'_{1+r}(\bar{\varepsilon}_{itc}; \overline{V}_{itc}) \\
&= \sum_{q=0}^{R-1} \widetilde{\eta}'_q(\bar{\varepsilon}_{itc}; \overline{V}_{itc}) \sum_{r=0}^{R} (-1)^r \frac{(r+1-q)!}{r!q!} \zeta(r+1-q) \\
&\quad + \widetilde{\eta}'_R(\bar{\varepsilon}_{itc}; \overline{V}_{itc}) \sum_{r=0}^{R} (-1)^r \frac{1}{r!} [\overline{V}_{itc} + \log(\nu_{itc1}) + \gamma_e] \\
&= \sum_{q=0}^{r-1} \widetilde{\eta}'_q(\bar{\varepsilon}_{itc}; \overline{V}_{itc}) \sum_{r=0}^{q-1} (-1)^r \frac{(r+1-q)!}{r!q!}
\end{aligned}
$$

with zero elements $\widetilde{\eta}'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc}) = 0$ for $q = r, r+1, \ldots, R$. Substitute for $\zeta(r + y_{itc} - q)$ from (9.5) and split the series expression for $p = 0$ and $p \geq 1$ to yield

$$\begin{aligned}
E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) \approx\ & \sum_{q=0}^{R-1} \widetilde{\eta}'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \sum_{r=0}^{R} (-1)^r \frac{(r+1-q)!}{r!q!} \sum_{p=0}^{\infty} \frac{1}{(1+p)^{r+1-q}} \\
& + \widetilde{\eta}'_R(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \sum_{r=0}^{R} (-1)^r \frac{1}{r!} [\overline{V}_{itc} + \log(\nu_{itc1}) + \gamma_e] \\
=\ & \sum_{q=0}^{R-1} \widetilde{\eta}'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \sum_{r=0}^{R} (-1)^r \frac{(r+1-q)!}{r!q!} \\
& + \sum_{q=0}^{R-1} \widetilde{\eta}'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \sum_{r=0}^{R} (-1)^r \frac{(r+1-q)!}{r!q!} \sum_{p=1}^{\infty} \frac{1}{(1+p)^{r+1-q}} \\
& + \widetilde{\eta}'_R(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \sum_{r=0}^{R} (-1)^r \frac{1}{r!} [\overline{V}_{itc} + \log(\nu_{itc1}) + \gamma_e]
\end{aligned}$$

For any given $q < r$, the sum over r in the first term is zero for any odd R.The sum over p in the second term is $O(1)$ as r grows large, while the sum over r is $o(1)$ as q grows large with r. For $q \geq r$ the elements of the array are zero by construction. The third term is $O(r!^{-1})$, completing the claim of the Lemma.

9.6. **Proof of Theorem 2**

The proof follows the application of Theorem 4.4.2 (weak consistency) and Theorem 4.4.4 (strong consistency) in Ghosh and Ramamoorthi (2003), henceforth GR. These results extend the classic theory of Schwartz (1965) based on the existence of uniformly exponentially consistent tests of the true parameter versus alternatives. An illuminating exposition of extensions of the Schwartz theory with applications can also be found in Ghosal (2010).

GR Theorems 4.4.2 and 4.4.4 are sufficiently generic to cover the consistency our parameter of interest – the posterior density of β. The weak consistency case follows immediately from GR Theorem 4.4.2 by imposing a Kullback-Leibler positivity condition on the prior. In verifying the sufficient condisions for the strong consistency case we follow the GR treatment of density estimation (GR chapter 5) which differs from our case in two aspects. In GR, the target posterior density is constructed directly as functional of the observables in the form of a Normal kernel mixture centered on the observables. In our case the target posterior density of β is also expressed as a functional of the observables, but its form is obtained from the economic model given by Assumptions 1–6. This difference necessitates a remake of the section of the GR proof where the posterior functional form is explicitly used, which is the first part of the GR Lemma 5.6.2 showing that the pointwise distance between the posterior density at two different parameter values can be bounded by some small δ.

The second difference is that using a different functional form of the target posterior density requires the construction of a slightly different sieve and hence bound on the corresponding metric entropy growth. The sieve only differs in its support but not its form, though, and hence we keep relatively close to GR on this aspect. Replacing the relevant parts of Theorem 4.4.4 in GR then constitutes our strong consistency proof. The notation throughout closely follows GR and Ghosal (2010).

The setup is similar to the one considered in GR (p. 121), and Ghosal (2010, p. 55). Denote by y_i the set of observables $\{Y_{itk} : k = 1, \ldots, C_{it}\}$, by $w_i = \{Y_i, X_i, Z_i, D_i\}$ the set of observables for the individual i, and by w_n the set of observables for all n individuals. The observables $\{X, Z, D\}$ are assumed distributed i.i.d., exogenously to our model for y. Denote by $\Omega \subset \mathbb{R}^d$ the parameter space of β. Denote by D_α the probability measure defined on the set of probability measures $M(\mathbb{R})$ generated by the Dirichlet Process $DP(\alpha, G_0)$. The existence and uniqueness of D_α is guaranteed e.g. by Theorem 3.2.1 in GR. Let $f(\psi_i|w_i)$ denote the marginal posterior density of β given y_i for which, from Assumptions 1–6 and hence 2.9,

$$f(\psi|w_i) \propto \int_{\mathcal{V}\setminus\Omega} E_{\overline{\varepsilon}}[f_y(y_{itc}|\overline{V}_{itc})]g(\overline{V}_{itc}|\psi)d(\gamma,\theta)k(\beta;\psi)$$

An analytical expression for $E_{\overline{\varepsilon}}[f_y(y_{itc}|\overline{V}_{itc})]$ is given in (3.1). We will use the short-hand notation f_0 for the true density of β, and f for a generic density of β. Given f, w_i are i.i.d. with a common distribution P_f. From Assumption 5, the hyperparameters $\psi_1, \ldots, \psi_n$ are i.i.d. G given G, and $w_i \sim P_f(\cdot, \psi_i)$ given $\psi_1, \ldots, \psi_n, G$. Let P_f^∞ denote the joint distribution of the sequence $\{w_i\}_{i=1}^\infty$. Assume that $P_f(\cdot, \psi_i)$ is absolutely continuous with respect to the Lebesgue measure yielding the density $p_f(w)$ over the domain of the observables.

Let $L_\mu = \{f : f \text{ is measurable}, \ f \geq 0, \ \int f d\mu = 1\}$ denote the space of densities with respect to a σ-finite measure μ on Ω. Every $f \in L_\mu$ then corresponds to the probability measure P_f. Equip L_μ with the L_1-metric

$$\|f - g\| = \int |f - g| \, d\mu$$

Under the L_1-metric, L_μ is complete and separable. On L_μ the L_1-metric is equivalent to the total variation metric (GR, p. 58). Denote by $\Pi(f)$ the prior on L_μ, which in our case is induced by D_α, and let $\Pi(f|w_n)$ denote the posterior. Let U be a set containing f_0. The following two definitions make the posterior consistency concept precise in the weak and the strong sense.

Definition 1 (GR Definition 4.2.1)**.** *$\{\Pi(\cdot|w_n)\}$ is said to be strongly or L_1-consistent at P_{f_0} if there is $\Omega_0 \subset \Omega$ such that $P_{f_0}^\infty(\Omega_0) = 1$ and for $\omega \in \Omega_0$*

$$\Pi(U|w_n) \to 1$$

for all total variation neighborhoods of P_{f_0}.

Definition 2 (GR Definition 4.2.2)**.** *$\{\Pi(\cdot|w_n)\}$ is said to be weakly consistent at P_{f_0} if there is $\Omega_0 \subset \Omega$ such that $P_{f_0}^\infty(\Omega_0) = 1$ and for $\omega \in \Omega_0$*

$$\Pi(U|w_n) \to 1$$

for all weak neighborhoods of P_{f_0}.

A useful approach to proving posterior consistency due to Schwartz (1965) imposes conditions on the support of the prior in the sense of the Kullback-Leibler divergence

$$K(f_1, f_2) = \int f_1 \log(f_1/f_2) d\mu$$

with $f_1, f_2 \in L_\mu$. The Schwartz approach has been extended and applied by a number of authors to various types of models. In general, the posterior probability of the complement U^c of U (and in general any set in place of U^c) can be expressed as

$$\Pi(f \in U^c | w_n) = \frac{\int_{U^c} \prod_{i=1}^n \frac{p_f(w_i)}{p_{f_0}(w_i)} \Pi(df)}{\int_{L_\mu} \prod_{i=1}^n \frac{p_f(w_i)}{p_{f_0}(w_i)} \Pi(df)} \tag{9.11}$$

To prove consistency it suffices to show that $\Pi(f \in U^c|w_n) \to 0$ a.s. $P_{f_0}^\infty$ for which the sufficient conditions are

$$\liminf_{n\to\infty} e^{nb} \int_{L_\mu} \prod_{i=1}^n \frac{p_f(w_i)}{p_{f_0}(w_i)} \Pi(df) = \infty \ \text{ a.e. } P_{f_0}^\infty \ , \forall b > 0 \tag{9.12}$$

and

$$\lim_{n\to\infty} e^{nb_0} \int_{U^c} \prod_{i=1}^n \frac{p_f(w_i)}{p_{f_0}(w_i)} \Pi(df) = 0 \ \text{ a.e. } \ P_{f_0}^\infty \ \text{ for some } b_0 > 0 \tag{9.13}$$

These correspond to (4.1) and (4.2) in GR, respectively. (9.13) ensures that the numerator in (9.11) converges to zero e^{nb_0} exponentially fast for some $b > 0$ while (9.12) ensures that the denominator multiplied by e^{nb} converges to infinity for all $b > 0$. We need to provide sufficient conditions for (9.12) and (9.13) to hold. The weak and strong consistency case differ in the metric utilized.

9.7. Controlling the Denominator

A sufficient condition for (9.12) is referred to as the Kullback-Leibler (or K-L) positivity property of the prior. We first define the concept of K-L support as in Ghosal (2010), p. 55 and then state sufficient conditions for the property to hold in an assumption on the true density of β. Denote by $K_\varepsilon(f)$ the neighborhood $\{g : K(p_f, p_g) < \varepsilon\}$.

Definition 3. *Let f_0 be in L_μ. f_0 is said to be in the K-L support of the prior* Π, *if for all* $\varepsilon > 0$, $\Pi(K_\varepsilon(f_0)) > 0$.

We impose the K-L support condition by the following assumption.

ASSUMPTION 7. $\Pi(K_\varepsilon(f_0)) > 0$ *for all* $\varepsilon > 0$.

Since ε can be chosen arbitrarily small, it immediately follows that the denominator in (9.11) multiplied by is exponentially large. Assumption 7 is thus sufficient to control the denominator.

9.8. Controlling the Numerator

Schwartz's original approach was to link (9.13) with the power of uniformly exponentially consistent tests for the hypothesis $H_0 : f = f_0$ versus $H_1 : f \in U^c$. Under the existence of these tests, Schwartz (1965) showed that the ratio of the marginal density of the observations with f conditioned to lie outside U to the true joint density is exponentially small except on a set with exponentially small sampling probability, which is sufficient to control (9.13) as required.

Under the weak topology, GR Theorem 4.4.2 shows that the K-L positivity condition given in Assumption 7 is sufficient for the existence of such uniformly consistent tests (for a formal definition see GR Definition 4.4.2) which in turn yield posterior consistency since Assumption 7 is also sufficient for controlling (9.12).

However, if U is a neighborhood of f_0 under the strong topology of total variation or the L_1-metric then these uniformly consistent tests will not exist, as shown by LeCam (1973) and Barron (1999) (GR, p. 132). In this case, the Schwartz approach will still go through but instead of U^c one needs to use its truncation to the sieve space $\mathcal{F}_n$ with the property that $\mathcal{F}_n^c \subset U^c$. Before proceeding further we will define a useful concept:

Definition 4 (GR Definition 4.4.5)**.** *Let $\mathcal{G} \subset L_\mu$. For $\delta > 0$, the L_1-metric entropy $J(\delta, \mathcal{G})$ is defined as the logarithm of the minimum of all n such that there exist $f_1, \ldots, f_n$ in L_μ with the property $\mathcal{G} \subset \cup_1^n \{f : \|f - f_i\| < \delta\}$.*

The formal statement for satisfying (9.13) in this case and hence strong consistency is given by GR Theorem 4.4.3 with sufficient conditions provided by GR Theorem 4.4.4. Since the proof of our Theorem 2 involves showing that the conditions of GR Theorem 4.4.4 are satisfied, we will restate the latter here:

THEOREM 3 (Theorem 4.4.4 of GR)**.** *Let Π be a prior on L_μ. Suppose $f_0 \in L_\mu$ and $\Pi(K_\varepsilon(f_0)) > 0$ for all $\varepsilon > 0$. If for each $\varepsilon > 0$, there is a $\delta < \varepsilon$, $c_1, b_1 > 0$, $b < \varepsilon^2/2$, and $\mathcal{F}_n^c \subset L_\mu$ such that, for all n large,*

$$\Pi(\mathcal{F}_n^c) < c_1 e^{-nb_1} \tag{9.14}$$

$$J(\delta, \mathcal{F}_n) < nb \tag{9.15}$$

then the posterior is strongly consistent at f_0.

The condition 9.14 imposes a restriction on the probability mass assigned to the complement of the sieve space by the prior, while 9.15 requires a bound of the metric entropy limiting the rate of growth of the sieve. Theorem 4.4.4 of GR is generic and covers a wide range of applications. In verifying its sufficient conditions we will proceed similarly to the density estimation case of GR (chapter 5) but with the key differences described above.

9.9. **Entropy Bound**

In order to apply Theorem 4.4.4 of GR, given

$$U = \{f : \|f - f_0\| < \varepsilon\}$$

for some $\delta < \varepsilon/4$, we need to construct sieves $\{\mathcal{F}_n : n \geq 1\}$ such that $J(\delta, \mathcal{F}_n) \leq nb$ and $\mathcal{F}_n^c$ has an exponentially small prior probability. The DP prior has the property that $D_\alpha\{G : G[-a_n, a_n] > 1 - \delta\} \to 1$ as $a_n \to \infty$. Hence, a natural candidate for $\mathcal{F}_n$ is

$$\mathcal{F}_n = \cup_n \mathcal{F}_n^{a_n} \tag{9.16}$$

with

$$\mathcal{F}_n^{a_n} = \left\{ \int f(\psi|w) dG(\psi) : G[-a_n, a_n] > 1 - \delta \right\}$$

where $a_n \in \mathbb{R}^d$ such that $a_n \to \infty$. This sieve is very similar to the one considered in GR (section 5.6.2). The following regularity condition will be useful in verifying the entropy bound.

ASSUMPTION 8. *Given the prior structure of Assumption 5,*

$$\|f(\psi_1|\cdot) - f(\psi_2|\cdot)\| \leq c_1 \|\psi_1 - \psi_2\|_E$$

where $c_1 < \infty$ is a constant that does not depend on n, and $\|\cdot\|_E$ is the Euclidean norm.

This Lipschitz continuity type assumption controls the influence of the prior hyperparameters ψ on the posterior density. The assumption is easily satisfied if $f(\psi|\cdot)$ is a smooth density, such as a smooth mixture of Normals as in our application.

The following Lemma provides a bound on $J(\delta, \mathcal{F}_n^{a_n})$.

LEMMA 3. *Under the Assumptions 1–8,*

$$J(\delta, \mathcal{F}_n^{a_n}) \leq c_2 a_n$$

where $c_2 > 0$ is a constant.

Proof. The proof proceeds in two steps, along the lines of GR (section 5.6.2). First we consider the case of $G[-a_n, a_n] = 1$ for some fixed finite $a_n \in \mathbb{R}^d$. This is then generalized to $G[-a_n, a_n] > 1 - \delta$ while permitting a_n to increase with n at a rate that preserves the previously established inequality.

Let

$$\mathcal{F}_{n,1}^{a_n} = \left\{ \int f(\psi|w) dG(\psi) : G[-a_n, a_n] = 1 \right\}$$

Given δ, let N be the smallest integer greater than $c_1 a_n/\delta$. Cover $(-a_n, a_n]$ with a set of balls E_i of radius a_n/N so that if $\psi, \psi' \in E_i$ then $\|\psi - \psi'\|_E < a_n/N$ and consequently by Assumption 8, $\|f(\psi|\cdot) - f(\psi'|\cdot)\| < \delta$.

The rest of the proof proceeds as in the proof of GR Lemma 5.6.2 and GR Lemma 5.6.3. In particular, let

$$\mathcal{G}_N = \left\{ (P_1, \ldots, P_N)\} : P_i \geq 0, \sum_{i=1}^{N} P_i = 1 \right\}$$

be the N-dimensional probability simplex and let $\mathcal{G}_N^*$ be a δ-net in $\mathcal{G}_N$, i.e. given $P \in \mathcal{G}_N$, there is $P^* = (P_1^*, \ldots, P_N^*) \in \mathcal{G}_N^*$ such that $\sum_{i=1}^N |P_i - P_i^*| < \delta$. Let $\mathcal{F}^* = \left\{ \sum_{i=1}^N P_i^* f(\psi_i|w_i) : P^* \in \mathcal{G}_N^* \right\}$.

Then $\mathcal{F}^*$ is 2δ-net in $\mathcal{F}_{n,1}^{a_n}$. To show this, note that if $\int f(\psi|w)dG(\psi) \in \mathcal{F}_{n,1}^{a_n}$ then set $P_i = P(E_i)$ and let $P^* \in \mathcal{G}_N^*$ be such that $\sum_{i=1}^N |P_i - P_i^*| < \delta$. Then, similarly to GP (p. 171),

$$\begin{aligned}
\left\| \int f(\psi|w)dG(\psi) - \sum_{i=1}^N P_i^* f(\psi|w_i) \right\| &\leq \left\| \int f(\psi|w)dG(\psi) - \sum_{i=1}^N \int I_{E_i} f(\psi_i|w_i)dG(\psi) \right\| \\
&\quad + \left\| \sum_{i=1}^N P_i f(\psi|w_i) - \sum_{i=1}^N P_i^* f(\psi|w_i) \right\| \\
&\leq \int \sum_{i=1}^N I_{E_i} \left\| f(\psi_i|w_i) - f(\psi_i|w_i) \right\| dG(\psi) + \sum_{i=1}^N |P_i - P_i^*| \\
&\leq 2\delta
\end{aligned}$$

where I_{E_i} is an indicator function taking the value of 1 over the set E_i and 0 elsewhere. This shows that

$$J(\delta, \mathcal{F}_{n,1}^{a_n}) \leq J(2\delta, \mathcal{G}_N) \tag{9.17}$$

The bound for $J(\delta, \mathcal{G}_N)$ is then given by

$$J(2\delta, \mathcal{G}_N) \leq N^d \left(c_1 a_n/\delta + 1\right) \left(1 + \log \frac{1+\delta}{\delta}\right) \tag{9.18}$$

using the approach of Barron, Schervish, and Wasserman (1999) cited in GR. Using GR Lemma 5.6.3,

$$J(3\delta, \mathcal{F}_n^{a_n}) \leq J(\delta, \mathcal{F}_{n,1}^{a_n})$$

which together with (9.17) and (9.18) completes the proof of the Lemma. □

Given these results, we can now proceed to the statement of the proof of Theorem 2.

Proof of Theorem 2.

(a) Under the weak topology, GR Theorem 4.4.2 shows that the K-L positivity condition given here in Assumption 7 is sufficient for weak posterior consistency.
(b) Under the strong topology with the L_1-metric, we apply GR Theorem 4.4.4. Its sufficient conditions are satisfied as follows: the K-L positivity property by Assumption 7, Condition (9.14) by Assumptions 5, 7, and the definition of $\mathcal{F}_n$, and condition (9.15) by Lemma 3 and (9.16).

□

10. **Technical Appendix**

10.1. **Poisson mixture in terms of a moment expansion**

Applying the series expansion

$$\exp(x) = \left(\sum_{r=0}^{\infty} \frac{(x)^r}{r!} \right)$$

to our Poisson mixture in (2.8) yields

$$\begin{aligned}
P(Y_{itc} &= y_{itc}|\delta_{itc}) = \int_{\Lambda} \frac{1}{y_{itc}!} \exp(-\delta_{itc}\lambda_{itc}) \, (\delta_{itc}\lambda_{itc})^{y_{itc}} \, g(\lambda_{itc}) d\lambda_{itc} \\
&= \int_{(\mathcal{V}\times\varepsilon)} \frac{1}{y_{itc}!} \exp(-\delta_{itc}(\overline{\varepsilon}_{itc} + \overline{V}_{itc}))\delta_{itc}^{y_{itc}} (\overline{\varepsilon}_{itc} + \overline{V}_{itc})^{y_{itc}} g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) g(\overline{V}_{itc}) d(\overline{\varepsilon}_{itc}, \overline{V}_{itc}) \\
&= \int_{\mathcal{V}} \int_{\varepsilon} \frac{1}{y_{itc}!} \left(\sum_{r=0}^{\infty} \frac{(-\delta_{itc}(\overline{\varepsilon}_{itc} + \overline{V}_{itc}))^r}{r!} \right) \delta_{itc}^{y_{itc}} (\overline{\varepsilon}_{itc} + \overline{V}_{itc})^{y_{itc}} g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} g(\overline{V}_{itc}) d\overline{V}_{itc} \\
&= \int_{\mathcal{V}} \sum_{r=0}^{\infty} \int_{\varepsilon} \frac{(-1)^r \delta_{itc}^{r+y_{itc}} (\overline{\varepsilon}_{itc} + \overline{V}_{itc})^{r+y_{itc}}}{r! y_{itc}!} g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} g(\overline{V}_{itc}) d\overline{V}_{itc} \\
&= \int_{\mathcal{V}} \sum_{r=0}^{\infty} \frac{(-1)^r}{r! y_{itc}!} \int_{\varepsilon} \delta_{itc}^{r+y_{itc}} (\overline{\varepsilon}_{itc} + \overline{V}_{itc})^{r+y_{itc}} g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} g(\overline{V}_{itc}) d\overline{V}_{itc} \\
&= \int_{\mathcal{V}} \sum_{r=0}^{\infty} \frac{(-1)^r}{r! y_{itc}!} \delta_{itc}^{r+y_{itc}} \eta'_{r+y_{itc}} (\overline{\varepsilon}_{itc}; \overline{V}_{itc}) g(\overline{V}_{itc}) d\overline{V}_{itc}
\end{aligned}$$

whereby $\sum_{r=0}^{\infty} \frac{(-1)^r}{r! y_{itc}!} \delta_{itc}^{r+y_{itc}} \eta'_{r+y_{itc}} (\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ is equivalent to $E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc})$ in (2.10).

10.2. **Evaluation of Conditional Choice Probabilities Based on Moments**

The moments $\eta'_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc})$ can be evaluated by deriving the Moment Generating Function (MGF) $M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s)$ of the composite random variable $\overline{\varepsilon}_{itc}$ and then taking the $w-$th derivative of $M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s)$ evaluated at $s = 0$:

$$\eta'_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) = \left. \frac{d^w}{ds^w} M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s) \right|_{s=0} \tag{10.1}$$

The expression for $M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s)$ can be obtained as the composite mapping

$$\begin{aligned}
M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s) &= F_1(M_{\overline{\varepsilon}_{it}}(s)) \\
&= F_1(F_2(M_{\varepsilon_{itck}}(s)))
\end{aligned} \tag{10.2}$$

where $M_{\varepsilon_{itck}}(s)$ is the MGF for the centered moments of ε_{itck}, $M_{\overline{\varepsilon}_{itc}}(s)$ is the MGF of the centered moments of $\overline{\varepsilon}_{itc}$, and F_1 and F_2 are functionals on the space C^{∞} of smooth functions.

Let $e_{itc} = \sum_{k=1}^{y_{itc}} \varepsilon_{itck}$ so that $\overline{\varepsilon}_{itc} = y_{itc}^{-1} e_{itc}$. Using the properties of an MGF for a composite random variable (Severini, 2005) and the independence of ε_{itck} over k conditional on V_{it}

$$\begin{aligned} M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s) &= \exp(\overline{V}_{it}s) M_{e_{itc}}(y_{itc}^{-1}s) \\ &= \exp(\overline{V}_{it}s) \prod_{k=1}^{y_{itc}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \end{aligned} \tag{10.3}$$

for $|s| < \kappa / y_{itc}^{-1}$ for some small $\kappa \in \mathbb{R}_+$. Let $r_n = r + y_{itc}$. Substituting and using the product rule for differentiation we obtain

$$\begin{aligned} f(y_{itc}|\overline{V}_{itc}) &= \sum_{r=0}^{\infty} \frac{(-1)^r}{r! y_{itc}!} \eta'_{r_n}(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) \\ &= \sum_{r=0}^{\infty} \frac{(-1)^r}{r! y_{itc}!} \frac{d^{r_n}}{ds^{r_n}} M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s) \Big|_{s=0} \\ &= \sum_{r=0}^{\infty} \frac{(-1)^r}{r! y_{itc}!} \frac{d^{r_n}}{ds^{r_n}} \exp(\overline{V}_{it}s) M_{e_{it}}(y_{itc}^{-1}s) \Big|_{s=0} \\ &= \sum_{r=0}^{\infty} \frac{(-1)^r}{r! y_{itc}!} \left\{ \sum_{w=0}^{r_n} \frac{r_n!}{w!(r_n - w)!} \overline{V}_{it}^{(r_n - w)} \frac{d^w}{dt^w} M_{e_{it}}(y_{itc}^{-1}s) \Big|_{s=0} \right\} \end{aligned}$$

Using the expression for $M_{e_{it}}(s)$ in (10.3) and the Leibniz generalized product rule for differentiation yields

$$\begin{aligned} \frac{d^w}{dt^w} M_{e_{it}}(y_{itc}^{-1}s) \Big|_{s=0} &= \frac{d^w}{dt^w} \prod_{k=1}^{y_{itc}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \Big|_{s=0} \\ &= \sum_{w_1 + \dots + w_{y_{itc}} = w} \frac{w!}{w_1! w_2! \dots w_{y_{itc}}!} \prod_{k=1}^{y_{itc}} \frac{d^{w_k}}{dt^{w_k}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \Big|_{s=0} \end{aligned} \tag{10.4}$$

Using $M_{\varepsilon_{itck}}(s)$, Lemma 1, and the form of the MGF for Gumbel random variables,

$$\frac{d^{w_k}}{dt^{w_k}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \Big|_{s=0} = \sum_{p=0}^{w_k} \frac{w_k!}{p!(w_k - p)!} \left(y_{itc}^{-1} \log(\nu_{itck}) \right)^{(w_k - p)} (-y_{itc}^{-1})^p \Gamma^{(p)}(1) \tag{10.5}$$

Moreover,

$$\Gamma^{(p)}(1) = \sum_{j=0}^{p-1} (-1)^{j+1} j! \widetilde{\zeta}(j+1)$$

with

$$\widetilde{\zeta}(j+1) = \begin{cases} -\gamma_e \text{ for } j = 0 \\ \zeta(j+1) \text{ for } j \geq 1 \end{cases}$$

where $\zeta(j+1)$ is the Riemann zeta function, for which $|\widetilde{\zeta}(j+1)| < \frac{\pi^2}{6}$ and $\widetilde{\zeta}(j+1) \to 1$ as $j \to \infty$. Using $\Gamma^{(p)}(1)$ in (10.5) and canceling $p!$ with $j!$ we obtain

$$\frac{d^{w_k}}{dt^{w_k}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \Big|_{s=0} = \sum_{p=0}^{w_k} \frac{w_k!}{(w_k - p)!} \alpha_1(w_k, p)$$

where

$$\begin{aligned} \alpha_1(w_k,p) &\equiv \left(y_{itc}^{-1}\log(\nu_{itck})\right)^{(w_k-p)}(-y_{itc}^{-1})^p\sum_{j=0}^{p-1}(-1)^{j+1}\frac{1}{p!^{(j)}}\widetilde{\zeta}(j+1) \\ p!^{(j)} &\equiv \prod_{c=j+1}^{p} c \end{aligned}$$

for $c \in \mathbb{N}$.

Substituting into (10.4) yields

$$\begin{aligned} \frac{d^w}{dt^w}M_{e_{it}}(y_{itc}^{-1}s)\bigg|_{s=0} &= \sum_{w_1+\cdots+w_{y_{itc}}=w}\frac{w!}{w_1!w_2!\ldots w_{y_{itc}}!}\prod_{k=1}^{y_{itc}}\sum_{p=0}^{w_k}\frac{w_k!}{(w_k-p)!}\alpha_1(w_k,p) \\ &= w!\sum_{w_1+\cdots+w_{y_{itc}}=w}\frac{1}{w_1!w_2!\ldots w_{y_{itc}}!}\prod_{k=1}^{y_{itc}}w_k!\sum_{p=0}^{w_k}\frac{1}{(w_k-p)!}\alpha_1(w_k,p) \\ &= w!\sum_{w_1+\cdots+w_{y_{itc}}=w}\prod_{k=1}^{y_{itc}}\sum_{p=0}^{w_k}\frac{1}{(w_k-p)!}\alpha_1(w_k,p) \\ &= w!\alpha_2(y_{itc}) \end{aligned}$$

where

$$\alpha_2(y_{itc}) \equiv \sum_{w_1+\cdots+w_{y_{itc}}=w}\prod_{k=1}^{y_{itc}}\sum_{p=0}^{w_k}\frac{1}{(w_k-p)!}\alpha_1(w_k,p)$$

Substituting into (10.1) and (3.1), canceling $w!$ and terms in $r_n!$ we obtain

$$\begin{aligned} (10.6) \qquad E_{\overline{\varepsilon}}f(y_{itc}|\overline{V}_{itc}) &= \sum_{r=0}^{\infty}\frac{(-1)^r}{r!y_{itc}!}\left\{\sum_{w=0}^{r_n}\frac{r_n!}{w!(r_n-w)!}\overline{V}_{itc}^{(r_n-w)}\frac{d^w}{dt^w}M_{e_{it}}(y_{itc}^{-1}s)\bigg|_{s=0}\right\} \\ &= \sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\sum_{w=0}^{r_n}\frac{r_n!^{(y_{itc})}}{(r+y_{itc}-w)!}\overline{V}_{itc}^{(r_n-w)}\alpha_2(y_{itc}) \end{aligned}$$

where

$$r_n!^{(y_{itc})} \equiv \prod_{c=y_{itc}+1}^{r_n} c$$

for $c \in \mathbb{N}$.

10.3. **Result C: Moments of Gumbel Random Variables**

Let $f^G(X;\mu,\sigma)$ denote the Gumbel density with mean μ and scale parameter σ. The moment-generating function of $X \sim f^G(X;\mu,\sigma)$ is

$$M_X(t) = E[\exp(tX)] = \exp(t\mu)\Gamma(1-\sigma t) \quad , \quad \text{for } \sigma|t| < 1.$$

(Kotz and Nadarajah, 2000).

Then,

$$
\begin{aligned}
\eta'_r(X) &= \frac{d^r}{dt^r} M_X(t)\bigg|_{t=0} \\
&= \frac{d^r}{dt^r} \exp(\mu t)\Gamma(1-\sigma t)\bigg|_{t=0} \\
&= \sum_{w=0}^{r} \binom{r}{w} \left[\frac{d^{r-w}}{dt^{r-w}} \exp(\mu t) \frac{d^w}{dt^w}\Gamma(1-\sigma t)\right]\bigg|_{t=0} \\
&= \sum_{w=0}^{r} \frac{r!}{w!(r-w)!} \left[\mu^{(r-w)} \exp(\mu t)(-\sigma)^w \Gamma^{(w)}(1-\sigma t)\right]\bigg|_{t=0} \\
&= \sum_{w=0}^{r} \frac{r!}{w!(r-w)!} \mu^{(r-w)}(-\sigma)^w \Gamma^{(w)}(1)
\end{aligned}
$$

where $\Gamma^{(w)}(1)$ is the w^{th} derivative of the gamma function around 1.

$$\Gamma^{(m)}(1) = \sum_{j=0}^{m-1} \psi_j(1)$$

$\psi_j(1)$ for $j = 1, 2$, can be expressed as

$$\psi_j(1) = (-1)^{j+1} j! \zeta(j+1)$$

where $\zeta(j+1)$ is the Riemann zeta function

$$\zeta(j+1) = \sum_{c=1}^{\infty} \frac{1}{c^{(j+1)}}$$

(Abramowitz and Stegun, 1964). Hence,

$$\Gamma^{(m)}(1) = \sum_{j=0}^{m-1} (-1)^{j+1} j! \widetilde{\zeta}(j+1)$$

where

$$\widetilde{\zeta}(j+1) = \begin{cases} -\gamma_e \text{ for } j = 0 \\ \zeta(j+1) \text{ for } j \geq 1 \end{cases}$$

for which $|\widetilde{\zeta}(j+1)| < \frac{\pi^2}{6}$ and $\widetilde{\zeta}(j+1) \to 1$ as $j \to \infty$ (Abramowitz and Stegun, 1964). Note that the NAG fortran library can only evaluate $\psi_m(1)$ for $m \leq 6$.

Moreover,

$$\begin{aligned}
\left.\frac{d^r}{dt^r}M_X(ct)\right|_{t=0} &= \left.\frac{d^r}{dt^r}\exp(\mu ct)\Gamma(1-\sigma ct)\right|_{t=0} \\
&= \sum_{w=0}^{r}\binom{r}{w}\left.\left[\frac{d^{r-w}}{dt^{r-w}}\exp(\mu ct)\frac{d^w}{dt^w}\Gamma(1-\sigma ct)\right]\right|_{t=0} \\
&= \sum_{w=0}^{r}\frac{r!}{w!(r-w)!}\left.\left[(\mu c)^{(r-w)}\exp(\mu ct)(-\sigma c)^w\Gamma^{(w)}(1-\sigma ct)\right]\right|_{t=0} \\
&= \sum_{w=0}^{r}\frac{r!}{w!(r-w)!}(\mu c)^{(r-w)}(-\sigma c)^w\Gamma^{(w)}(1)
\end{aligned}$$

10.4. **Properties of Cumulants**

The cumulants κ_n of a random variable X are defined by the cumulant-generating function (CGF) which is the logarithm of the moment-generating function (MGF), if it exists:

$$\begin{aligned}
CGF(t) &= log\left(E\left[e^{tX}\right]\right) \\
&= \sum_{n=0}^{\infty}\kappa_n\frac{t^n}{n!}
\end{aligned}$$

The cumulants κ_n are then given by the derivatives of the $CGF(t)$ at $t=0$. Cumulants are related to moments by the following recursion formula:

$$\kappa_n = \mu'_n - \sum_{k=1}^{n-1}\binom{n-1}{k-1}\kappa_k\mu'_{n-k}$$

Cumulants have the following properties not shared by moments (Severini, 2005):

(1) *Additivity:* Let X and Y be statistically independent random vectors having the same dimension, then

$$\kappa_n(X+Y) = \kappa_n(X) + \kappa_n(Y)$$

i.e. the cumulant of their sum $X+Y$ is equal to the sum of the cumulants of X and Y. This property also holds for the sum of more than two independent random vectors. The term "cumulant" reflects their behavior under addition of random variables.

(2) *Homogeneity:* The n^{th} cumulant is homogenous of degree n, i.e. if c is any constant, then

$$\kappa_n(cX) = c^n\kappa_n(X)$$

(3) *Affine transformation:* Cumulants of order $n \geq 2$ are semi-invariant with respect to affine transformations. If κ_n is the n^{th} cumulant of X, then for the n^{th} cumulant of the affine transformation $a+bX$ it holds that, independent of a,

$$\kappa_n(a+bX) = b^n\kappa_n(X)$$

References

Abramowitz, M., and I. A. Stegun (1964): *Handbook of mathematical functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edn.

Aguiar, M., and E. Hurst (2007): "Life-Cycle Prices and Production," *American Economic Review*, 97(5), 1533–1559.

Aldrich, J. (2002): "How Likelihood and Identification Went Bayesian," *International Statistical Review*, 70(2), 79–98.

Allenby, G. M., P. E. Rossi, and R. E. McCulloch (2005): "Hierarchical Bayes Models: A Practitioners Guide," Ssrn working paper, Ohio State University, University of Chicago.

Antoniak, C. E. (1974): "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 1, 1152–1174.

Bajari, P., J. T. Fox, K. Kim, and S. Ryan (2009): "The Random Coefficients Logit Model is Identified," .

Barron, A. R. (1999): "Information-theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems," in *Bayesian Statistics 6*, ed. by J. Bernardo, J. Berger, A. Dawid, and A. Smith, pp. 27–52. Oxford University Press, New York.

Barron, A. R., M. J. Schervish, and L. Wasserman (1999): "The consistency of posterior distributions in nonparametric problems," *Annals of Statistics*, 27(2), 536–561.

Bernardo, J. M., and A. F. M. Smith (1994): *Bayesian Theory*. Wiley, New York.

Berry, S. T., and P. A. Haile (2010): "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," .

Briesch, R. A., P. K. Chintagunta, and R. L. Matzkin (2010): "Nonparametric Discrete Choice Models With Unobserved Heterogeneity," *Journal of Business and Economic Statistics*, 28(2), 291–307.

Broda, C., E. Leibtag, and D. E. Weinstein (2009): "The Role of Prices in Measuring the Poor's Living Standards," *Journal of Economic Perspectives*, 23(2), 77–97.

Burda, M., M. C. Harding, and J. A. Hausman (2008): "A Bayesian Mixed Logit-Probit Model for Multinomial Choice," *Journal of Econometrics*, 147(2), 232–246.

Chiappori, P.-A., and I. Komunjer (2009): "On the Nonparametric Identification of Multiple Choice Models," .

David, H. A., and H. N. Nagaraja (2003): *Order Statistics*. Wiley, 3 edn.

Diaconis, P., and D. Freedman (1986a): "On inconsistent Bayes estimates of location," *Annals of Statistics*, 14, 68–87.

——— (1986b): "On the consistency of Bayes estimates," *Annals of Statistics*, 14, 1–67.

——— (1990): "On the uniform consistency of Bayes estimates for multinomial probabilities," *Annals of Statistics*, 18, 1317–1327.

Doob, J. L. (1949): "Application of the theory of martigales," in *Le calcul des probabilites et ses applications*, pp. 22–28. Centre national de la recherche scientifique (France).

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

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

Fox, J. T., and A. Gandhi (2010): "Nonparametric Identification and Estimation of Random Coefficients in Nonlinear Economic Models," .

Ghosal, S. (2010): "Dirichlet process, related priors and posterior asymptotics," in *Bayesian Nonparametrics*, ed. by N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. Cambridge University Press, accessed at www4.stat.ncsu.edu/ sghosal/papers/BayesAsymp.pdf.

Ghosh, J., and R. Ramamoorthi (2003): *Bayesian Nonparametrics*. Springer-Verlag, New York.

Greene, W. (2007): "Functional Form and Heterogeneity in Models for Count Data," .

Gurmu, S., P. Rilstone, and S. Stern (1999): "Semiparametric estimation of count regression models," *Journal of Econometrics*, 88(1), 123–150.

Harding, M. C., and M. Lovenheim (2010): "Living on a Budget: National Evidence on Price Search and Consumer Heterogeneity from Homescan Data," Working paper, Stanford University.

Hausman, J., B. H. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52(4), 909–938.

Hausman, J. A. (1997): "Valuation of New Goods under Perfect and Imperfect Competition," in *The Economics of New Goods*, ed. by T. Bresnahan, and R. Gordon. University of Chicago Press.

Jochmann, M., and R. León-González (2004): "Estimating the demand for health care with panel data: a semiparametric Bayesian approach," *Health Economics*, 13, 1003–1014.

Kadane, J. B. (1974): "The role of identification in Bayesian theory," in *Studies in Bayesian Econometrics and Statistics*, ed. by S. E. Fienberg, and A. Zellner, pp. 175–191. Amsterdam: North-Holland.

Karlis, D., and E. Xekalaki (2005): "Mixed Poisson Distributions," *International Statistical Review*, 73(1), 35–58.

Kotz, S., and S. Nadarajah (2000): *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London, UK, 1 edn.

LeCam, L. (1973): "Convergence of Estimates under dimensionality restrictions," *Annals of Statistics*, 1, 38–53.

Lewbel, A. (2000): "Semiparametric Qualitative Response Model Estimation With Instrumental Variables and Unknown Heteroscedasticity," *Journal of Econometrics*, 97(0), 145–177.

Lukacs, E. (1970): *Characteristic Functions*. Griffin, London, UK, 2 edn.

Mannering, F. L., and M. M. Hamed (1990): "Occurence, frequency, and duration of commuters' work-to-home departure delay," *Transportation Research Part B: Methodological*, 24(2), 99–109.

Munkin, M. K., and P. K. Trivedi (2003): "Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: an application to the demand for healthcare," *Journal of Econometrics*, 114(2), 197–220.

Nadarajah, S. (2008): "Exact distribution of the linear combination of p Gumbel random variables," *International Journal of Computer Mathematics*, 85(9), 1355–1362.

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

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69(2), 307–42.

Rendon, S. (2002): "Fixed and Random Effects in Classical and Bayesian Regression," Economics Working Papers 613, Department of Economics and Business, Universitat Pompeu Fabra.

Romeu, A., and M. Vera-Hernández (2005): "Counts with an endogenous binary regressor: A series expansion approach," *Econometrics Journal*, 8(1), 1–22.

Sapatinas, T. (1995): "Identifiability of mixtures of power-series distributions and related characterizations," *Annals of the Institute of Statistical Mathematics*, 47(3), 447–459.

Schwartz, L. (1965): "On Bayes procedures," *Probability Theory and Related Fields*, 4, 10–26.

Severini, T. A. (2005): *Elements of Distribution Theory*. Cambridge University Press.

Smith, P. J. (1995): "A Recursive Formulation of the Old Problem of Obtaining Moments from Cumulants and Vice Versa," *The American Statistician*, 49(2), 217–218.

Teicher, H. (1960): "On the Mixture of Distributions," *The Annals of Mathematical Statistics*, 31(1), 55–73.

Terza, J. V. (1998): "Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects," *Journal of Econometrics*, 84(1), 129–154.

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

Winkelmann, R. (2008): *Econometric Analysis of Count Data*. Springer, 5 edn.