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

Martin Burda,[†] Matthew Harding,[‡] Jerry Hausman[§]

March 15, 2010

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 purchase volume – 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.

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 Princeton, Stanford, and 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`

[§]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 contend 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 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 irregularities 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].

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

[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).

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-Hernndez (2005), and Jochmann (2006), 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 purchase volume – 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 reviews some key definitions that are used in subsequent analysis. Section 3 introduces the mixed Poisson model with its analyzed properties and the efficient recursive estimation procedure. Section 4 elaborates on the tools of Bayesian analysis used in model implementation, Section 5 elaborates on the issues on identification and posterior consistency, Section 6 discusses the application results and Section 7 concludes.

2. **Poisson Mixtures**

In this Section we establish notation and briefly review several relevant concepts and definitions that will serve as reference for subsequent analysis. In the base-case 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 $\{Y(t), t \geq 0\}$ where $Y(t)$ represents the total number of events that have occurred before t. The Poisson assumption stipulates stationary and independent increments for $Y(t)$ 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 $Y(t)$ within the interval Δ is given by

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

Hence the probability of an event occurring in an infinitesimal time interval dt is λdt 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 $\widetilde{\lambda}(t)$. Then the probability of a unit addition to the count process $Y(t)$ within the interval Δ is given by

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

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

$$\lambda_t = \int_0^t \widetilde{\lambda}(s)ds \tag{2.2}$$

it holds that the p.m.f. of the resulting Y on the interval $[0, t)$ is given again by the base-case $P(Y = y)$ in (2.1).

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 $y \in \mathbb{N}_0$. The mixture component $g(\lambda)$ accounts for overdispersion typically present in count data. This definition naturally extends to probability functions other than Poisson but we do not consider such cases here. Parametrizing $g(\lambda)$ in (2.3) as the gamma density yields the Negative Binomial model. For a number of existing mixed Poisson specifications applied in other model contexts, see Karlis and Xekalaki (2005), Table 1.

An additional 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

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

3. Model

3.1. Count Intensity

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. The bottom level of the individual decision process is formed by the utility-maximizing choice among the various store alternatives. Let the latent continuous-time *potential* utility of an individual i at time instant $\tau \in (t-1, t]$ be given by

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

where X_{itj} are key variables of interest, D_{itj} are other relevant (individual-)alternative-specific variables, $j = 1, \ldots, J$ is the index over alternatives, and $\widetilde{\varepsilon}_{itj}$ is a stochastic disturbance term with a strictly stationary marginal extreme value type 1 density $f_{\widetilde{\varepsilon}}$. Let the index $c \in \mathcal{C} \subseteq \mathcal{J}$ label the chosen utility-maximizing alternative, with $c = 1, \ldots, C_{it}$ for each i, t. The $\widetilde{U}_{itj}(\tau)$ is rationalized as specifying an individual's internal utility ranking for the choice alternatives at the instant τ as a function of individual-product characteristics, product attributes and a continuously evolving strictly stationary idiosyncratic component process. As in the logit model, the parameters $\widetilde{\beta}_i$ and $\widetilde{\theta}_i$ are only identified up to a common scale.

The trip count intensity choice forms the top level of our model. An individual faces various time constraints on the number of trips they can possibly make for the purpose of shopping. We do not attempt to model such constraints explicitly as households' shopping patters can be highly irregular – people can make unplanned spontaneous visits of grocery stores or cancel pre-planned trips on a moment's notice. Instead, we treat the actual occurrences of shopping trips as realizations of an underlying continuous-time Poisson process whereby the *probability* of taking the trip to store j in the next instant $d\tau$ is given by the continuous-time shopping intensity $\widetilde{\lambda}_{itj}(\tau)$. Such assumption appears warranted for our application where one time period t spans the lengths of a month over the duration of a number of potential shopping cycles. The individual is then viewed as making a joint decision on the store choice and the shopping intensity.

Under our model framework the shopping intensity $\widetilde{\lambda}_{itc}(\tau)$ is interpreted as a latent continuous-time utility of the trip count that is intrinsically linked to the utility $\widetilde{U}_{itc}(\tau)$. At the instant τ of the count incidence (new trip), the agent chooses the alternative $c = j$ that maximizes $\widetilde{U}_{itj}(\tau)$. We specify the link between $\widetilde{\lambda}_{itc}(\tau)$ and $\widetilde{U}_{itc}(\tau)$ as a monotonic invertible mapping h that takes the form

$$\begin{aligned} \widetilde{\lambda}_{itc}(\tau) &= h(\widetilde{U}_{itc}(\tau)) \\ &= \gamma' Z_{it}(\tau) + \omega_{1i}\widetilde{\beta}_i' X_{itc}(\tau) + \omega_{2i}\widetilde{\theta}_i' D_{itc}(\tau) + \omega_{3i}\widetilde{\varepsilon}_{itc}(\tau) \\ &= \gamma' Z_{it}(\tau) + \beta_i' X_{itc}(\tau) + \theta_i' D_{itc}(\tau) + \varepsilon_{itc}(\tau) \end{aligned} \tag{3.1}$$

for $\widetilde{U}_{itc}(\tau) \geq 0$, where ω_{1i}, ω_{2i}, and ω_{3i} are unknown factors of proportionality. Higher utility derived from the *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. Such 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 (3.1) 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 (3.2), 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.

In our application of supermarket trip choice and count, Z_{it} include various demographic characteristics, X_{itck} is composed of price, driving distance, and purchase volume, while D_{itck} are formed by store indicator variables. Since count data records a discrete number of choices y_{itc} observed during the period t for an individual household i, we treat

$$\widetilde{U}_{itck} = \widetilde{\beta}_i' X_{itck} + \widetilde{\theta}_i' D_{itck} + \widetilde{\varepsilon}_{itck} \tag{3.2}$$

as realizations of the continuous time process yielding $\widetilde{U}_{itc}(\tau)$ for $\tau \in (t-1, t]$ where $k = 1, \ldots, Y_{itc}$ is the index over the undertaken choice occasions. Referring to the result on integrated intensity (2.2), letting

$$\lambda_{itc} = \int_{t-1}^{t} \widetilde{\lambda}_{itc}(\tau) d\tau \tag{3.3}$$

enables us to use the λ_{itc} as the integrated count intensity for individual i during period t due to the Poisson assumption on the count process evolution. Since information on $\widetilde{\lambda}_{itc}(\tau)$ is available only on the filtration $\mathcal{T}_{it} \equiv \{\tau_1, \ldots, \tau_{y_{it}}\} \subset (t-1, t]$, using the assumption of Poisson count increments we approximate the integral in (3.3) by

$$\lambda_{itj} \simeq y_{itj}^{-1} \sum_{k=1}^{y_{itj}} \lambda_{itjk}$$

for $\lambda_{itck} \geq 0$. Hence,

$$\begin{aligned} \lambda_{itck} &= \max\{0, \lambda_{itck}^*\} \\ \lambda_{itck}^* &= \gamma' Z_{it} + \beta_i' X_{itck} + \theta_{ic} D_{itck} + \varepsilon_{itck} \end{aligned}$$

and

$$\begin{aligned} \lambda_{itc}^* &\simeq \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} [\gamma' Z_{it} + \beta_i' X_{itck} + \theta_{ic} D_{itck} + \varepsilon_{itck}] \\ &= \gamma' Z_{it} + \beta_i' \overline{X}_{itc} + \theta_i \overline{D}_{itc} + \overline{\varepsilon}_{itc} \\ &= \overline{V}_{itc} + \overline{\varepsilon}_{itc} \end{aligned} \tag{3.4}$$

where $\overline{X}_{itc} = \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} X_{itck}$, $\overline{D}_{itc} = \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} D_{itck}$, and $\overline{\varepsilon}_{itc} = \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} \varepsilon_{itck}$. Here λ_{it}^* corresponds to the actual ex-post realized latent utility. 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} reflect the various constraints the consumers face regarding the realized shopping frequency which are assumed to be of the Poisson nature.

3.2. **Count Probability Function**

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 for a small fraction of the time period, but the count intensity associated with such alternative was not sufficiently high to result in a unit increase of its count process. Capturing this effect introduces a continuous-scale latent measurement of the probability of selection associated with each alternative. This allows us to conduct counterfactual experiments that alter the micro-foundations of any given alternative as given by alternative-specific 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. Due to the assumption of extreme value type 1 distribution of the residual ε_{itj}, 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

$$\widetilde{V}_{itc} = \widetilde{\beta}_i' X_{itck} + \widetilde{\theta}_i' D_{itck}$$

is the deterministic part of the utility function (3.2). Even though $\widetilde{\beta}_i$ and $\widetilde{\theta}_i$ could in principle be estimated parametrically as in the standard logit model, we use the semiparametric estimation logit-probit procedure of Burda, Harding, and Hausman (2008) in order to preserve the overall model flexibility.

Based on the proportional Poisson pmf (2.4) and the specification of λ_{itc} in (3.4), the count probability mass function is given by

$$P(Y_{itc} = y_{itc}|\delta_{itc}) = \int \frac{\exp\left(-\delta_{itc}\lambda_{itc}\right)\left(\delta_{itc}\lambda_{itc}\right)^{y_{itc}}}{y_{itc}!} g(\lambda_{itc})d(\lambda_{itc}) \tag{3.5}$$

which, conditional on δ_{itc}, is a particular case of the Poisson mixture model (2.3) with a mixing distribution $g(\lambda_{itc})$ that arises from the underlying individual utility maximization problem. However, the marginal density $g(\lambda_{itc})$ takes on a very complicated form. From (3.4), 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 then 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 (3.4) 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 $\gamma, \theta_i, \beta_i$ is 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,

$$\begin{aligned} P(Y_{itc} &= y_{itc}|\delta_{itc}) = \int_{\Lambda} f(y_{itc}|\overline{\varepsilon}_{itc}, \overline{V}_{itc}) g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) g(\overline{V}_{itc}) d(\overline{\varepsilon}_{itc}, \overline{V}_{itc}) \\ &= \int_{\mathcal{V}} \int_{\varepsilon} f(y_{itc}|\overline{\varepsilon}_{itc}, \overline{V}_{itc}) g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} g(\overline{V}_{itc}) d\overline{V}_{itc} \\ &= \int_{\mathcal{V}} E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) g(\overline{V}_{itc}) d\overline{V}_{itc} \end{aligned} \tag{3.6}$$

[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.

where

$$E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) = \int_{\varepsilon} f(y_{itc}|\overline{\varepsilon}_{itc}, \overline{V}_{itc}) g(\overline{\varepsilon}_{itc}|\overline{V}_{itc}) d\overline{\varepsilon}_{itc} \tag{3.7}$$

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

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

In the remainder of this Section we derive a novel approach for analytical evaluation of $E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc})$ in (3.7). 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 (3.5) admits a representation in terms of an infinite sequence of *moments* of the mixing distribution

$$E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc}) = \sum_{r=0}^{\infty} \frac{(-1)^r}{y_{itc}!r!} \delta_{itc}^{r+y_{itc}} \eta'_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) \tag{3.8}$$

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 (3.8) 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 (3.8) is smaller than some numerical tolerance level δ local to zero in the implementation.

3.3. 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, (3.8) 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].

[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.

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 (3.8) 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 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 of extreme value type 1 distribution on the stochastic disturbances $\varepsilon_{itj}(\tau)$ in the underlying individual choice model (3.1). 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 our model assumptions,* $f_{\max}(\varepsilon_{itck})$ *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

$$K_{\varepsilon_{itck}}(s) = s \log(\nu_{itck}) - \log \Gamma(1-s) \tag{3.9}$$

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 (3.8) is given by the following Theorem.

THEOREM 1. *Under our model assumptions,*

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

where

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

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. *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, thus entering $\mathbf{Q}_{y_{itc},r}$. 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 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.

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

4.2. **Prior Structure**

Denote the model hyperparameters by W and their joint prior by $k(W)$. From (3.6),

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

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 as follows:

$$\begin{aligned} \beta_i &\sim F^0 \\ \theta_i &\sim N(\underline{\mu}_\theta, \underline{\Sigma}_\theta) \\ \gamma &\sim N(\underline{\mu}_\gamma, \underline{\Sigma}_\gamma) \end{aligned}$$

Denote the respective priors by $k(\beta_i)$, $k(\theta_i)$, $k(\gamma)$. The assumption on the distribution of β_i implies the following hyperparameter model (Neal, 2000):

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

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 $\underline{\Sigma}_\gamma$.

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 hyperparameters of the DP prior for β_i. For α use the updates described in Escobar and West (1995).
- Update the parameter δ_{itc} as in Burda, Harding, and Hausman (2008)
- Update the remaining hyperparameters based on the identifed 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), 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 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 simply 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 of β_i, which forms the nonparametric component of our model under the Dirichlet process prior, by verifying the conditions necessary for a consitency result by Schwartz (1964) as detailed in Ghosal (2009). We first construct the sieve estimate of an assumed true posterior as implied by our model. Then we show 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. We also verify that the model likelihood is bounded in the appropriate sense. We then invoke the result of Ghosal (2009) who proves the existence of Kullback-Leibler type positivity for the DP prior which along with the above conditions yield posterior consistency for our model. The result is summarized in the following Theorem with the proof provided in the Appendix.

THEOREM 2. *Under our model assumptions, for the posterior $K(\beta_i|\cdot)$ and an arbitrary neighborhood V_0 or the true posterior $K_0(\beta_i|\cdot)$ it holds that $P(K(\beta_i|\cdot) \notin V_0) \to 0$ as the sample size tends to infinity.*

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 imbalanced 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.

We dropped from the sample households with fewer than 6 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 536,976 for a total 1,210 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.

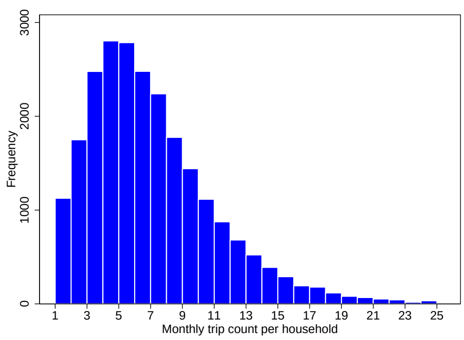

FIGURE 1. Histogram of the monthly total number of trips to a store per month for the households in the sample.

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

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

the corresponding supermarket; and *purchase volume* which captures the total amount of money spent on a given shopping trip to a store, accounting for the different price levels of stores that were not visited on the trip .

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.

The purpose of the *purchase volume* variable is to capture information on the overall magnitude of a given shopping trip in terms of the amount of expenditure of the household on that trip, while accounting for the differences in price sensitivity that individuals exhibit at varying amounts of the overall quantity of produce purchased. To achieve both goals we constructed the variable as an interactive term *price* $\times$ *quantity* (both in logs) where the *quantity* variable was calculated dividing the actual total expenditure of a shopping trip by the unit price of the shopping basket. In

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

addition to the actual expenditure incurred, we compute what the total cost and hence hypothetical expenditure would have been at alternatives that were not chosen by multiplying the expenditure at the alternative that was chosen by the ratio of the price index at the alternative that was not chosen divided by the price index for chosen store.

6.2. **Results**

First we consider the estimated densities of our key parameters of interest on price, distance and purchase volume. Plots of marginal densities of our key parameters of interest on price, distance, and purchase volume (logs) are presented in Figure 2. Plots of joint densities of pairs of these parameters (price vs distance, distance vs volume, price vs volume) are given in Figure 3. All plots attest to the existence of several sizeable preference clusters of consumers. 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.

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 three distinct preference clusters in the price-distance preference space. Consumer who dislike higher prices fall into two categories. One group of consumers is prepared to invest the effort to search for a better deal and are thus willing and eager to travel longer distances to make their preferred purchases. The other group of consumers is insensitive to travel distance while remaining sensitive to the price. The last cluster of consumers is very convenience focused and thus willing to pay higher prices in exchange for not having to travel very far. Moreover, with each additional unit of quantity purchased, price-sensitive consumers become even more price sensitive, while price-insensitive consumers take even less regard to higher unit prices.

This trade-off is further exemplified by the graph pair illustrating the joint distribution of the coefficients on distance and purchase volume (lower graph pair in Figure 3). While most consumers are strongly motivated by the desire for convenience and are willing to make large volume purchase decisions even if they end up paying a higher price as a result, a non-trivial cluster of consumers is willing to travel longer distances in order to economize on large volume transactions.

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 purchase volume coefficient β_{3i}. A 3D animation is available at http://dl.getdropbox.com/u/716158/pde7722b.wmv while a 2D contour animation is at http://dl.getdropbox.com/u/716158/pde7722ctb.wmv. We find the preference clusters corresponding to the trade-offs between price and distance of different consumers to

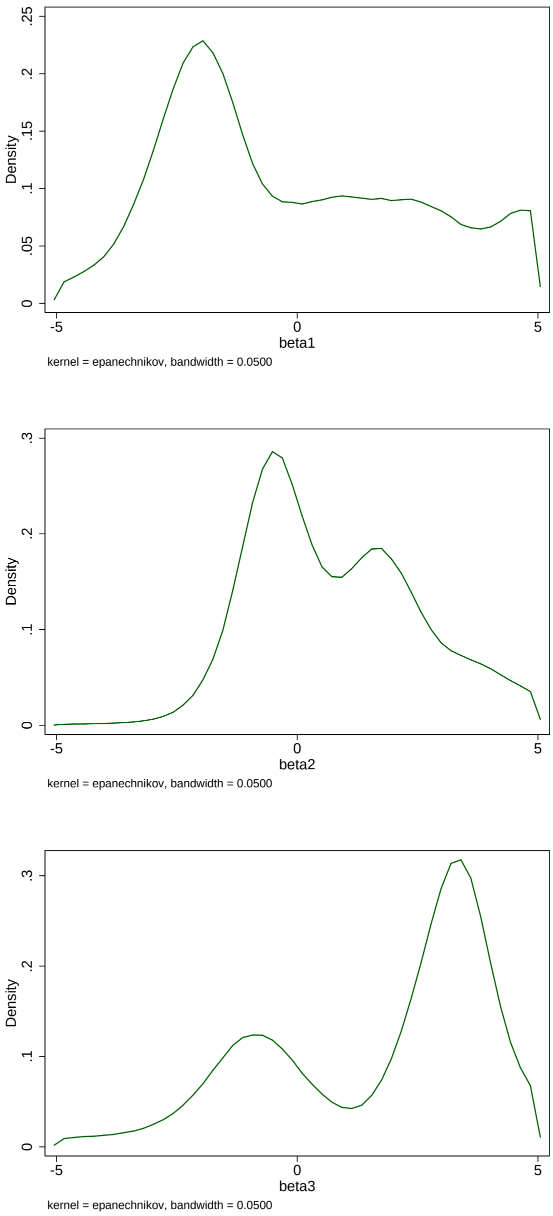

FIGURE 2. Density of MC draws of the individual-specific coefficients on price β_{i1} (top), distance β_{i2} (middle), and purchase volume β_{i3} (bottom) variables. The modes are -2.035, -0.497, and 3.274, respectively.

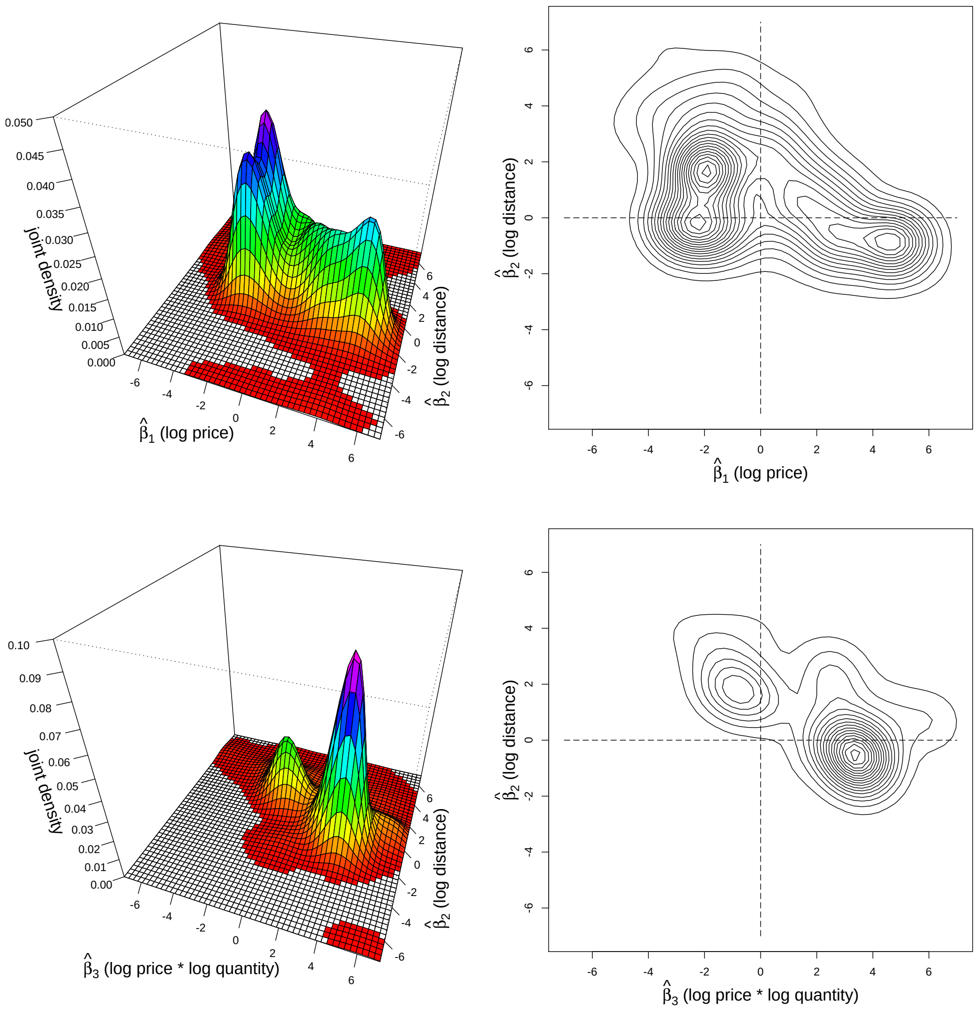

FIGURE 3. Joint density of MC draws of β_1 vs β_2 and β_2 vs β_3

be remarkably stable at different levels of purchase volume. At the same time we find substantial heterogeneity in the trade-offs between price and distance at different levels of the sensitivity of households with respect to purchase volume.

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 95% Bayesian Credible Sets (BCS, corresponding to 0.025 and 0.975 quantiles) for coefficients

γ on demographic variables are presented in Table 3 with their marginal counterparts incorporating the price interaction effects in Table 4.

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

At the same time, following a price increase, households increase their shopping count intensity for their chosen stores (Table 4). 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 3 and 4). Households that belong to the high age or unemployed category intensify their search most when faced with higher prices. The search effect further increases at higher price levels for the unemployed while somewhat abating for the aged. The opportunity cost of time relative to other household types is a likely factor at play. High income households substantially increase their search intensity only for high price levels while remaining virtually insensitive to price changes at low price levels. The opposite pattern is exhibited by the educated household category which could reflect the differences in the effective target price basket levels demanded by these two household types. The non-white and Hispanic household categories maintain overall medium search intensity, albeit differing in their search adjustment at high price levels. Households that are large, with children, or have one household head absent increase their search intensity the least following a price increase. This could reflect the various trade-offs among household members regarding their preferences for the former category, while for the latter two the opportunity cost of time is a plausible factor affecting their behavioral patterns.

Table 5 shows the posterior means, medians, standard deviations and 95% Bayesian Credible Sets for the means of b_θ and Table 6 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 and Walmart have the lowest store effect means but also the largest variance of the means, reflecting the diversity of preferences regarding the shopping intensity at these two store types on the part of the pool of households. Randalls exhibits the highest store effect which likely stems from its one-stop-shopping business concept featuring an array of specialty departments, once its price level – by far the highest among all the store types – has been controlled for. H.E.B., Pantry Foods, and Other belong to the mid-range category in terms of store shopping intensity preference. The store effects also exhibit various interesting correlation patterns (Table 6). The highest correlation falls on H.E.B. and Pantry Foods: indeed, Pantry Foods is a subsidiary owned by H.E.B., even though the former are conceptually separate limited-assortment stores with reduced surface area and facilities. H.E.B. and Randalls also exhibit relatively high positive correlation, competing for the same segment of customers in the mainstream market. In contrast, Kroger and Pantry Foods are highly negatively correlated, reflecting customer loyalty to either one or the other store type.

Figure 4 shows the kernel density estimate of the MC draws of the Dirichlet process latent class model hyperparameter α (left) and the density of the number of latent classes obtained at each MC step in the Dirichlet process latent class sampling algorithm (right). Thus, we conclude the empirical analysis by highlighting the fact that the clustering in the preference distributions is not particularly sensitive to the choice of α. Indeed the posterior distribution of this Dirichlet parameter is tight and takes values similar to those suggested by other statistical applications.

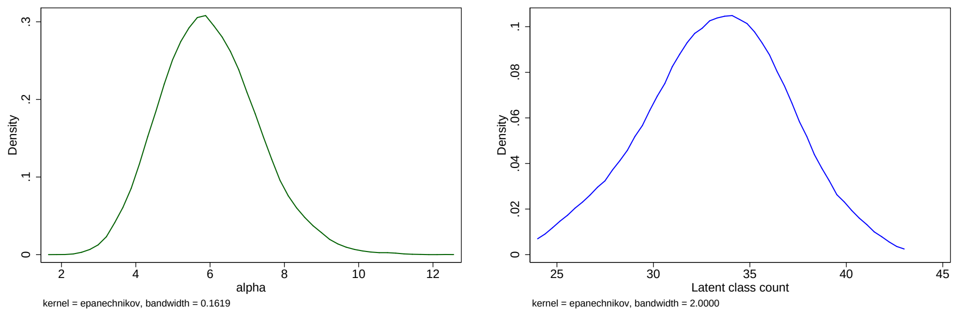

FIGURE 4. Density of MC draws of α (left), and density of the latent class count (right).

7. 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 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 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 various clusters of consumers regarding their preferences over the price-distance trade-off, and their joint density for diverse levels of overall shopping expenditure. 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.

Variable	Mean	Median	Std.Dev.	95% BCS
Large household	1.481	1.487	0.335	(0.850, 2.217)
Children	0.036	0.130	0.369	(-0.579, 0.599)
Absent head	1.459	1.428	0.201	(1.174, 1.875)
Non-white	-2.034	-2.057	0.179	(-2.310,-1.662)
Hispanic	1.109	0.980	0.528	(0.339, 1.996)
Unemployed	1.251	1.350	0.601	(-0.198, 2.046)
Education	2.035	2.146	0.383	(1.127, 2.460)
Age	1.634	1.606	0.230	(1.266, 2.071)
Income	-1.016	-1.268	0.525	(-1.751,-0.090)
$logP\times$ Large household	-4.152	-4.326	0.683	(-5.064,-2.781)
$logP\times$ Children	0.697	0.500	0.712	(-0.371, 1.780)
$logP\times$ Absent head	-3.223	-3.186	0.537	(-4.336,-2.183)
$logP\times$ Non-white	8.846	8.805	0.762	(7.534,10.402)
$logP\times$ Hispanic	-0.874	-0.807	1.082	(-2.342, 1.024)
$logP\times$ Unemployed	0.292	-0.370	1.815	(-1.565, 5.000)
$logP\times$ Education	-4.331	-4.640	1.176	(-5.624,-1.608)
$logP\times$ Age	-0.517	-0.579	0.521	(-1.423, 0.232)
$logP\times$ Income	5.814	6.233	1.344	(3.558, 7.637)

TABLE 3. Coefficients γ on demographic variables. $logP$ denotes interaction term with price.

Variable	Mean	Median	Std.Dev.	95% BCS
Large household	0.109	0.015	0.336	(-0.407,0.821)
Children	0.266	0.254	0.209	(-0.065,0.658)
Absent head	0.401	0.379	0.131	(0.167,0.684)
Non-white	0.880	0.904	0.290	(0.461,1.373)
Hispanic	0.821	0.798	0.281	(0.323,1.286)
Unemployed	1.348	1.393	0.267	(0.753,1.906)
Education	0.606	0.593	0.107	(0.377,0.826)
Age	1.463	1.497	0.173	(1.098,1.771)
Income	0.899	0.905	0.177	(0.588,1.249)

TABLE 4. Marginal coefficients γ on demographic variables.

Parameter	Mean	Median	Std.Dev.	95% BCS
$b_{\theta 1}$ (HEB)	-5.406	-5.728	2.126	(-8.941,-1.667)
$b_{\theta 2}$ (Kroger)	-6.773	-6.304	2.593	(-11.156,-2.500)
$b_{\theta 3}$ (Randalls)	-0.605	-0.959	0.986	(-2.062, 1.592)
$b_{\theta 4}$ (Walmart)	-8.348	-8.406	2.545	(-12.273,-3.690)
$b_{\theta 5}$ (Pantry Foods)	-2.643	-2.515	1.093	(-4.444,-0.532)
$b_{\theta 6}$ (other)	-4.285	-4.828	1.225	(-6.558,-2.302)

TABLE 5. Means b_θ of distributions of store indicator variable coefficients θ_i.

Parameter	Mean	Median	Std.Dev.	95% BCS
$\Sigma_{\theta 1 \theta 1}$ (HEB)	2.654	2.643	0.188	(2.306, 3.056)
$\Sigma_{\theta 1 \theta 2}$ (HEB & Kroger)	-0.029	-0.030	0.102	(-0.233, 0.170)
$\Sigma_{\theta 1 \theta 3}$ (HEB & Randalls)	0.731	0.730	0.136	(0.474, 1.006)
$\Sigma_{\theta 1 \theta 4}$ (HEB & Walmart)	-0.225	-0.222	0.100	(-0.430,-0.030)
$\Sigma_{\theta 1 \theta 5}$ (HEB & Pantry Foods)	1.647	1.638	0.172	(1.327, 2.006)
$\Sigma_{\theta 1 \theta 6}$ (HEB & other)	0.634	0.632	0.128	(0.395, 0.895)
$\Sigma_{\theta 2 \theta 2}$ (Kroger)	2.075	2.073	0.154	(1.788, 2.387)
$\Sigma_{\theta 2 \theta 3}$ (Kroger & Randalls)	-0.150	-0.149	0.107	(-0.363, 0.056)
$\Sigma_{\theta 2 \theta 4}$ (Kroger & Walmart)	0.244	0.239	0.091	(0.072, 0.434)
$\Sigma_{\theta 2 \theta 5}$ (Kroger & Pantry Foods)	-0.803	-0.802	0.153	(-1.104,-0.510)
$\Sigma_{\theta 2 \theta 6}$ (Kroger & other)	-0.321	-0.319	0.095	(-0.517,-0.139)
$\Sigma_{\theta 3 \theta 3}$ (Randalls)	3.039	3.042	0.248	(2.545, 3.530)
$\Sigma_{\theta 3 \theta 4}$ (Randalls & Walmart)	0.060	0.060	0.127	(-0.182, 0.317)
$\Sigma_{\theta 3 \theta 5}$ (Randalls & Pantry Foods)	-0.196	-0.174	0.210	(-0.629, 0.161)
$\Sigma_{\theta 3 \theta 6}$ (Randalls & other)	-0.249	-0.247	0.127	(-0.510, 0.001)
$\Sigma_{\theta 4 \theta 4}$ (Walmart)	1.958	1.949	0.157	(1.680, 2.286)
$\Sigma_{\theta 4 \theta 5}$ (Walmart & Pantry Foods)	0.591	0.590	0.120	(0.361, 0.837)
$\Sigma_{\theta 4 \theta 6}$ (Walmart & other)	0.291	0.293	0.096	(0.103, 0.479)
$\Sigma_{\theta 5 \theta 5}$ (Pantry Foods)	3.145	3.125	0.261	(2.701, 3.712)
$\Sigma_{\theta 5 \theta 6}$ (Pantry Foods & other)	-0.491	-0.485	0.134	(-0.771,-0.243)
$\Sigma_{\theta 6 \theta 6}$ (other)	2.387	2.376	0.169	(2.081, 2.752)

TABLE 6. Covariances Σ_θ of distributions of store indicator variable coefficients θ_i.

8. Appendix

8.1. 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 12 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. [12]

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 $\mu_\beta^{\phi_i} \sim N(\underline{\mu}_\beta, \underline{\Sigma}_\beta)$, $\Sigma_\beta^{\phi_i} \sim IW(\underline{\Sigma}_\beta^{\phi_i}, v_{0\Sigma_\beta})$, $\underline{\mu}_\beta = 0$, $\underline{\Sigma}_\beta = diag(100)$, $\underline{\Sigma}_\beta^{\phi_i} = diag(1/3)$, 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 an 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 expresses 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 $\mu_\theta \sim N(\underline{\mu}_\theta, \underline{\Sigma}_\theta)$, $\underline{\mu}_\theta = 0$, $\underline{\Sigma}_\theta = diag(1/3)$, $\Sigma_\theta \sim IW(\underline{\Sigma}_\theta, v_{0\Sigma_\theta})$, 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 number of latent classes decreased in the first half of the burn-in section to just under 40 and remained at that level for the rest of the run, with mean of 33.2 and standard deviation of 3.1. 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

[12] 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.

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 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.

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

We have

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

Therefore

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

Defining $z_{itck} = \exp(-\varepsilon_{itck})$ 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

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

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:

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

Hence

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

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}.

8.3. **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 (3.8) 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 (3.8).

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

$$
\begin{aligned}
\eta'_w(\overline{\varepsilon}_{itc};\overline{V}_{itc}) = \sum_{q=0}^{w-2}\frac{(w-1)!}{q!(w-1-q)!}\kappa_{w-q}(\overline{\varepsilon}_{itc};\overline{V}_{itc})\eta'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \\
+\kappa_1(\overline{\varepsilon}_{itc};\overline{V}_{itc})\eta'_{w-1}(\overline{\varepsilon}_{itc};\overline{V}_{itc})
\end{aligned}
\tag{8.1}
$$

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

$$\begin{aligned}\kappa_w(\overline{\varepsilon}_{itc}) &= \kappa_w\left(\frac{1}{y_{itc}}\sum_{k=1}^{y_{itc}}\varepsilon_{itck}\right)\\ &= \left(\frac{1}{y_{itc}}\right)^w \kappa_w\left(\sum_{k=1}^{y_{itc}}\varepsilon_{itck}\right)\\ &= \left(\frac{1}{y_{itc}}\right)^w \sum_{k=1}^{y_{itc}}\kappa_w\left(\varepsilon_{itck}\right) \qquad (8.2)\end{aligned}$$

[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

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

yielding for $w = 1$

$$\kappa_1(\varepsilon_{itck}) = \log(\nu_{itck}) + \gamma_e \qquad (8.3)$$

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

$$\begin{aligned}\kappa_w(\varepsilon_{itck}) &= \left.-\frac{d^w}{ds^w}\log\Gamma(1-s)\right|_{s=0}\\ &= (-1)^w\psi^{(w-1)}(1)\\ &= (w-1)!\zeta(w) \qquad (8.4)\end{aligned}$$

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

$$\zeta(w) = \sum_{p=0}^{\infty}\frac{1}{(1+p)^w} \qquad (8.5)$$

(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 (8.3) and (8.4) in (8.2) 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

$$\kappa_1(\overline{\varepsilon}_{itc};\overline{V}_{itc}) = \overline{V}_{itc} + \kappa_1(\overline{\varepsilon}_{itc}) \tag{8.6}$$

while for $w > 1$

$$\kappa_w(\overline{\varepsilon}_{itc};\overline{V}_{itc}) = \kappa_w(\overline{\varepsilon}_{itc}) \tag{8.7}$$

[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 (8.6) and (8.7) in (8.1), canceling the term $(w-i-1)!$, yields

$$\begin{aligned}\eta'_w(\overline{\varepsilon}_{itc};\overline{V}_{itc}) &= \sum_{q=0}^{w-2}\frac{(w-1)!}{q!}\left(\frac{1}{y_{itc}}\right)^{w-q-1}\zeta(w-q)\eta'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc})\\ &\quad +[\overline{V}_{itc}+\frac{1}{y_{itc}}\sum_{k=1}^{y_{itc}}\log(\nu_{itck})+\gamma_e]\eta'_{w-1}(\overline{\varepsilon}_{itc};\overline{V}_{itc})\end{aligned} \tag{8.8}$$

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

$$\widetilde{\eta}'_{r+y_{itc}}(\overline{\varepsilon}_{itc};\overline{V}_{itc}) = \frac{(-1)^r}{r!y_{itc}!}\delta_{itc}^{r+y_{itc}}\eta'_{r+y_{itc}}(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \tag{8.9}$$

denote the scaled raw moment obtained by scaling $\eta'_{r+y_{itc}}(\overline{\varepsilon}_{itc};\overline{V}_{itc})$ in (8.8) with $(-1)^r\delta_{itc}^{r+y_{itc}}/(r!y_{itc}!)$. Summing the expression (8.9) over $r = 1,\ldots,\infty$ would now give us the desired series representation for (3.7). The expression (8.9) relates unscaled moments expressed in terms of cumulants to scaled ones. We will now elaborate on a recursive relation based on (8.9) 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 (3.7).

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 $\eta'_{r+y_{itc}}(\bar{\varepsilon}_{itc};\overline{V}_{itc})$ in (8.8), 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 $p = 1,\dots,r+y_{itc}$, distinguishing three different cases:

(1) For $p \leq y_{itc}$ the summands in $\widetilde{\eta}'_p$ from (8.8) 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 $\kappa_1(\bar{\varepsilon}_{itc};\overline{V}_{itc})$ is r^{-1} for each $p = 1,\dots,y_{itc}+r$. 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 $\widetilde{\eta}'_{y_{itc},r-2} = (\widetilde{\eta}'_0,\dots,\widetilde{\eta}'_{y_{itc}+r-2})^T$ and $\mathbf{Q}_{y_{itc},r-2} = (Q_{y_{itc},r,q},\dots,Q_{y_{itc}r,,r-2})^T$ the recursive updating scheme

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

yields the expression

$$\begin{aligned}
\widetilde{\eta}'_{y_{itc}+r}(\overline{\varepsilon}_{itc};\overline{V}_{itc}) &= (-1)^r \delta_{itc}^{r+y_{itc}} \sum_{q=0}^{y_{itc}+r-2} \frac{(y_{itc}+r-1)!}{r!q!} + \left(\frac{1}{y_{itc}^{y_{itc}+r-q-1}}\right) \zeta(y_{itc}+r-q)\widetilde{\eta}'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc}) \\
&\quad +(-1)^r \delta_{itc}^{r+y_{itc}} \frac{1}{r!}[\overline{V}_{itc} + \frac{1}{y_{itc}} \sum_{k=1}^{y_{itc}} \log(\nu_{itck}) + \gamma_e]\widetilde{\eta}'_{y_{itc}+r-1}(\overline{\varepsilon}_{itc};\overline{V}_{itc})
\end{aligned} \tag{8.10}$$

for a generic $y_{itc}+r$ which is equivalent to our target term in (8.9) that uses the substitution for $\eta'_w(\overline{\varepsilon}_{itc};\overline{V}_{itc})$ from (8.8). However, unlike the unscaled moments $\eta'_q(\overline{\varepsilon}_{itc};\overline{V}_{itc})$, the terms on the right-hand side of (8.10) are bounded and yield a convergent sum over $r = 1,\ldots,\infty$ required for evaluation of (3.7), as verified in Lemma 2. An illustrative example of our recursive updating scheme for $y_{itc} = 4$ follows.

8.4. 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.

8.5. Proof of Lemma 2

From (8.10) we have

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

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 (8.10) in (3.7), for R large enough to evaluate $E_{\overline{\varepsilon}} f(y_{itc}|\overline{V}_{itc})$ with a numerical error smaller than some tolerance level, switch the order of summation between r and q to obtain a triangular array

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

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 (8.5) and split the series expression for $p = 0$ and $p \geq 1$ to yield

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

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.

8.6. **Proof of Theorem 2**

Our proof proceeds along the lines of Ghosal (2009) whose approach based on Schwartz (1964) requires showing four conditions:

A: The prior probability mass assigned to a complement of the sieve space implied by the model is exponentially small and the model sieve approaches the true population value of the parameter as the sample size grows without bound;

B: The model sieve satisfies an entropy condition binding the rate of growth of the sieve space in terms of its $logN(\epsilon/2)$-covering number;

C: The model likelihood for β_i is bounded in an appropriate sense;

D: The Kullback-Leibler positivity property of the prior is satisfied.

Ghosal (2009) by an application of Fubini's theorem shows that these conditions are sufficient to yield weak posterior consistency for the density of β_i . However, in our case the latent variable β_i drawn from the Dirichlet process mixture (DPM) model is subject to MC updates implying also updates for its DPM hyperparameters ψ_i. Hence the posterior of β_i is here a convolution of $F(\psi_i)$ for all MC draws of ψ_i. In contrast, the counterpart of β_i in Ghosal (2009) and other literature on Bayesian density estimation was fixed data within the realm of kernel density estimation with appropriate bandwidths. Therefore, our proof requires us to show conditions A and B using different means. Our model likelihood is new in the literature and hence condition C also needs to be verified independently. Condition D was shown by Ghosal (2009) to hold for the Dirichlet Process prior with a Gaussian base measure which we adopt in our case as well.

Let $\mathcal{P}$ be a class of density functions where the possible values of p lie and $p_0 \in \mathcal{P}$ stand for the true density function. For our problem we define the sieve space $\mathcal{P}_n$ as a collection of all mixture densities of β_i constructed as convolutions of $F(\psi_{is})$ over $i = 1, \ldots, N$, for $s = 1, \ldots, S$ MC draws. This sieve follows naturally from the model (3.5). Since $\psi_i|G \sim G$ with $G \sim DP(\alpha, G_0)$ and in our case G_0 is the Gaussian base measure while the kernel F is also Gaussian, the Dirichlet process assigns an exponentially small probability to the complement $\mathcal{P}_n^c \subset \mathcal{P}$ of $\mathcal{P}_n$, satisfying condition **A**.

The Gaussianity of $F(\psi_i)$ also allows us to benefit from the result of Ghosal and van der Vaart (2001), Theorem 3.3, who derived the bounds on the bracketing entropies of the class of normal mixtures. Since our sieve falls into the category considered by these authors, the log $\epsilon/2-$covering number of our sieve in Hellinger distance is bounded by log $\left(\frac{1}{\varepsilon}\right)^{4\gamma+1}$ for $0 < \varepsilon < 1/2$ and $\gamma \geq 1/2$, satisfying condition **B**.

Lemma 1, 3 and the joint Gaussianity of $g(\overline{V}_{itc})$ for each i and t imply that the model likelihood (3.5) is uniformly bounded, complying with condition **C.** Ghosal (2009) establishes the Kullback-Leibler property of Dirichlet mixture of normal prior at a true density p_0 (condition **D**) by approximating p_0 by p_m defined as the convolution of p_0 truncated to $[-m, m]$ for some large m and the Gaussian kernel with a small bandwidth. This convolution, which is itself a normal mixture, approximates p_0 pointwise as well as in the Kullback-Leibler sense under mild conditions on p_0. A Kullback-Leibler neighborhood around p_m includes a set which can be described in terms of a weak neighborhood around p_0 truncated to $[-m, m]$. Since the Dirichlet process has large weak support, the resulting weak neighborhood will have positive probability, proving the Kullback-Leibler property.

9. Technical Appendix

9.1. 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 (3.5) 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 (3.7).

9.2. 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 $s = 0$:

$$\eta'_w(\overline{\varepsilon}_{itc}; \overline{V}_{itc}) = \frac{d^w}{ds^w} M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s)\Big|_{s=0} \tag{9.1}$$

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

$$\begin{aligned}
M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s) &= F_1(M_{\overline{\varepsilon}_{it}}(s)) \\
&= F_1(F_2(M_{\varepsilon_{itck}}(s)))
\end{aligned} \tag{9.2}$$

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}

$$
\begin{aligned}
M_{\overline{\varepsilon}_{itc}|\overline{V}_{itc}}(s) &= \exp(\overline{V}_{it}s) M_{e_{itc}}(y_{itc}^{-1}s) \\
&= \exp(\overline{V}_{it}s) \prod_{k=1}^{y_{itc}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s)
\end{aligned}
\tag{9.3}
$$

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 (9.3) and the Leibniz generalized product rule for differentiation yields

$$
\begin{aligned}
\frac{d^w}{dt^w} M_{e_{it}}(y_{itc}^{-1}s) \Big|_{s=0} &= \frac{d^w}{dt^w} \prod_{k=1}^{y_{itc}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \Big|_{s=0} \\
&= \sum_{w_1 + \cdots + w_{y_{itc}} = w} \frac{w!}{w_1! w_2! \ldots w_{y_{itc}}!} \prod_{k=1}^{y_{itc}} \frac{d^{w_k}}{dt^{w_k}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \Big|_{s=0}
\end{aligned}
\tag{9.4}
$$

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

$$
\frac{d^{w_k}}{dt^{w_k}} M_{\varepsilon_{itck}}(y_{itc}^{-1}s) \Big|_{s=0} = \sum_{p=0}^{w_k} \frac{w_k!}{p!(w_k - p)!} \left(y_{itc}^{-1} \log(\nu_{itck}) \right)^{(w_k - p)} (-y_{itc}^{-1})^p \Gamma^{(p)}(1)
\tag{9.5}
$$

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 (9.5) 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 (9.4) yields

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

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 (9.1) and (3.8), canceling $w!$ and terms in $r_n!$ we obtain

$$\begin{aligned}(9.6)\qquad E_{\overline{\varepsilon}}f(y_{itc}|\overline{V}_{itc}) &= \sum_{r=0}^{\infty}\frac{(-1)^r}{r!y_{itc}!}\left\{\sum_{w=0}^{r_n}\frac{r_n!}{w!(r_n-w)!}\overline{V}_{itc}^{(r_n-w)}\frac{d^w}{dt^w}M_{e_{it}}(y_{itc}^{-1}s)\Big|_{s=0}\right\}\\ &= \sum_{r=0}^{\infty}\frac{(-1)^r}{r!}\sum_{w=0}^{r_n}\frac{r_n!^{(y_{itc})}}{(r+y_{itc}-w)!}\overline{V}_{itc}^{(r_n-w)}\alpha_2(y_{itc})\end{aligned}$$

where

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

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

9.3. **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,

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

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,

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

9.4. **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.

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," .

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.

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. (2009): "Dirichlet process, related priors and posterior asymptotics," in *Bayesian Nonparametrics in Practice*, ed. by N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker. Cambridge University Press (to appear).

Ghosal, S., and A. W. van der Vaart (2001): "Entropies And Rates Of Convergence For Maximum Likelihood And Bayes Estimation For Mixtures Of Normal Densities," *Annals of Statistics*, 29(5), 1233–1263.

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.

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. (2006): "Three Essays on Bayesian Nonparametric Modeling in Microeconometrics," Phd thesis, Universitt Konstanz.

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.

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-Hernndez (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. (1964): "On Bayes procedures," *Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 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): *Econometrics Analysis of Count Data.* Springer, 5 edn.