ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian nonparametric mixed random utility models

George Karabatsos [a,*], Stephen G. Walker [b]

[a] Educational Psychology, University of Illinois-Chicago, 60304 Chicago, IL, United States

[b] Statistics, University of Kent, United Kingdom

ARTICLE INFO

Article history:
Received 5 May 2010
Received in revised form 14 October 2011
Accepted 15 October 2011
Available online xxxx

Keywords:
Mixed multinomial logit model
Stick-breaking priors
Bayesian nonparametrics

ABSTRACT

We propose a mixed multinomial logit model, with the mixing distribution assigned a general (nonparametric) stick-breaking prior. We present a Markov chain Monte Carlo (MCMC) algorithm to sample and estimate the posterior distribution of the model's parameters. The algorithm relies on a Gibbs (slice) sampler that is useful for Bayesian nonparametric (infinite-dimensional) models. The model and algorithm are illustrated through the analysis of real data involving 10 choice alternatives, and we prove the posterior consistency of the model.

Published by Elsevier B.V.

1. Introduction

The random utility model is fundamental to most theories of choice behavior. Given a finite set of choice alternatives $\mathcal{C} = \{c = 1, \ldots, C\}$, this general model assumes that the (choice) probability $\Pr(c|\mathcal{C})$ that an individual selects alternative c is defined by:

$$\Pr(c|\mathcal{C}) = \Pr\left(U_c = \max_{l=1,\ldots,C} U_l | \mathcal{C}\right), \tag{1}$$

for $c = 1, \ldots, C$, where $(U_1, \ldots, U_C)$ are continuous random utility variables having joint distribution F (Train, 2003). For example, McFadden and Train (2000) introduced the mixed multinomial logit (MMNL) model,

$$\begin{aligned}\Pr(c|\mathcal{C}, \mathbf{x}) &= \int_{\mathbb{R}^K} \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta}) \mathrm{d}G(\boldsymbol{\beta}) \\ &= \int_{\mathbb{R}^K} \left[e^{\mathbf{x}_c'\boldsymbol{\beta}} \Big/ \sum_{l=1}^{C} e^{\mathbf{x}_l'\boldsymbol{\beta}} \right] \mathrm{d}G(\boldsymbol{\beta}), \quad c = 1, \ldots, C,\end{aligned}$$

which is a general random utility model that has seen many applications. Here, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_C)$, with each $\mathbf{x}_c$ a $K \times 1$ covariate vector describing the attributes of the choice alternative c and the decision-maker, $\boldsymbol{\beta}$ is a K-variate random parameter, and G is the mixing distribution.

McFadden and Train proved that any discrete choice model derived from a random utility model has choice probabilities that can be approximated arbitrarily-well by the MMNL model, provided that this model assigns sufficiently large support

* Corresponding author.
E-mail address: georgek@uic.edu (G. Karabatsos).

0167-9473/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.csda.2011.10.014

to the space of mixing distributions,[1] $\{G\}$. However, as they mention, the theorem does not explicitly specify such an MMNL model. In many applications of the MMNL model (Train, 2003), G is assumed to be a K-variate normal distribution with mean vector $\boldsymbol{\mu}$ and $K \times K$ covariance matrix $\boldsymbol{T}$. This assumption can be violated in real data sets, and precludes the possibility of accounting for a wider class of mixing distributions, including distributions that are heavier-tailed, skewed, or multimodal.

From a Bayesian perspective, an obvious way to address this problem is to specify a prior distribution that supports a large class of mixing distribution, G. This approach was taken by Rossi et al. (2005, Ch.5), and by Burda et al. (2008), who specifically considered the Dirichlet process (DP) prior (Ferguson, 1973). (This DP approach to the MMNL model can be implemented using the rhierMnlDP() routine in the bayesm package Rossi, 2011 of the R software R Development Core Team, 2011). In this article, we consider inference with a general random utility model under a more general, stick-breaking prior for G, which includes the DP prior as a special case. For the posterior sampling of G, we implement Walker's (2007) Gibbs sampling algorithm, along with improvements in the algorithm (Kalli et al., 2010). This algorithm can be applied to a general random utility model, that is assigned any specific stick-breaking prior.

Next, we review the stick-breaking prior and present the MMNL model in terms of this prior. In Section 3, we present a Markov chain Monte Carlo (MCMC) sampling algorithm, which can be used to infer the model's posterior distribution. In Section 4 we illustrate this model through the analysis of real data, where we also compare the fit of MMNL models under different priors for the mixing distribution, namely, the Pitman–Yor stick-breaking process, the DP, and a multivariate normal prior. In Section 5, we study the consistency of the model under general choices of stick-breaking prior. This is necessary because Bayesian models with infinite-dimensional prior distributions can give rise to consistency problems (e.g., see Barron et al., 1999). We conclude in Section 6.

2. The model

A stick-breaking prior, denoted by Stick-Breaking($\mathbf{a}, \mathbf{b}, H$), is a prior distribution defined on the space of distributions, $\{G\}$, and has parameters $(\mathbf{a}, \mathbf{b}, H)$, where $\mathbf{a} = (a_1, a_2, \ldots, a_j, \ldots)'$ and $\mathbf{b} = (b_1, b_2, \ldots, b_j, \ldots)'$ are vectors of positive-valued parameters, and H is a distribution with density h defined with respect to Lebesgue measure (Ishwaran and James, 2001). This stick-breaking prior supports the entire space of (measurable) distributions, including symmetric distributions, skewed distributions, multimodal distributions, skewed and multimodal distributions, and so forth.

The stick-breaking prior is most conveniently represented by viewing a random distribution, G, as being constructed by a particular stochastic process. Specifically, let $v_j \sim \text{Beta}(a_j, b_j)$ independently, and let $\boldsymbol{\beta}_j \sim H$ independently for $j = 1, 2, \ldots$. Then such a random distribution is constructed by taking $G(\cdot) = \sum_{j=1}^{\infty} w_j \delta_{\boldsymbol{\beta}_j}(\cdot)$, where

$$w_j = v_j \prod_{l=1}^{j-1}(1 - v_l), \quad j = 1, 2, \ldots,$$

and where $\delta_{\boldsymbol{\beta}}$ denotes the degenerate distribution supporting $\boldsymbol{\beta}$. Using the standard terminology of mixture modeling, this random distribution G is a mixture distribution with an infinite number of components, having mixing weights $w_1, w_2, \ldots, w_j, \ldots$, with $\sum_{j=1}^{\infty} w_j = 1$, and having $\delta_{\boldsymbol{\beta}_1}, \delta_{\boldsymbol{\beta}_2}, \ldots, \delta_{\boldsymbol{\beta}_j}, \ldots$ as the component densities of the mixture. The construction described above is called a "stick-breaking" procedure because at each stage j, we randomly break what is left of a stick of unit length, and then assign the length of this break to the value of w_j. Also, it is obvious from this construction that a stick-breaking prior supports discrete distributions, with probability 1.

The stick-breaking prior is a general nonparametric prior which include other nonparametric priors as special cases. Here, we give two examples. First, the two-parameter Pitman–Yor (Poisson–Dirichlet) process (Pitman, 1996) is defined by the specifications $a_j = 1 - a$ and $b_j = b + ja$ for all $j = 1, 2, \ldots$, where a and b satisfy $0 \leq a < 1$ and $b > -a$. Second, the DP (Sethuraman, 1994) is defined by the specifications $a = 0$ and $b = \alpha$ (Sethuraman, 1994), and has mean $\text{E}[G(\cdot)] = H(\cdot)$ and variance $\text{Var}[G(\cdot)] = \{H(\cdot) + [1 - H(\cdot)]\}/(\alpha + 1)$, where α represents the prior degree of belief in H (Ferguson, 1973).

The MMNL model, with a general stick-breaking prior assigned to the mixing distribution G, can be represented by:

$$\Pr(c|\mathcal{C}, \mathbf{x}) = \int_{\mathbb{R}^K} \left[e^{\mathbf{x}_c'\boldsymbol{\beta}} \Big/ \sum_{l=1}^{C} e^{\mathbf{x}_l'\boldsymbol{\beta}} \right] \text{d}G(\boldsymbol{\beta}),$$
$$G|\boldsymbol{\Psi} \sim \text{Stick-Breaking}(\mathbf{a}, \mathbf{b}, H(\boldsymbol{\Psi}))$$
$$\boldsymbol{\Psi} \sim \pi(\boldsymbol{\Psi}|\boldsymbol{\Gamma}),$$

with baseline distribution H depending on parameter $\boldsymbol{\Psi}$, assigned a hyper-prior density $\pi(\boldsymbol{\Psi}|\boldsymbol{\Gamma})$ depending on hyper-parameter $\boldsymbol{\Gamma}$. Via Bayes' theorem, a set of data $\boldsymbol{D} = \{c_i, \mathbf{x}_i\}_{i=1}^n$ with likelihood $\prod_{i=1}^n \Pr(c_i|\mathcal{C}, \mathbf{x}_i)$ combines with the prior distribution $\Pi(G, \boldsymbol{\Psi})$ to yield a posterior distribution $\Pi(G, \boldsymbol{\Psi}|\boldsymbol{D})$. Since $\boldsymbol{\beta}$ is a K-dimensional vector with space $\mathbb{R}^K$, a convenient choice of baseline distribution $H(\boldsymbol{\Psi})$ is given by the K-variate normal distribution, with $H(\boldsymbol{\Psi}) = \text{Normal}_K(\boldsymbol{\mu}, \boldsymbol{T})$,

[1] Also, a reviewer points out that in order for McFadden and Train's result to hold, one must include arbitrary functions of the covariates in the probabilities of the MNL model.

Please cite this article in press as: Karabatsos, G., Walker, S.G., Bayesian nonparametric mixed random utility models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.014

where $\boldsymbol{\Psi} = (\boldsymbol{\mu}, \boldsymbol{T})$ and $\boldsymbol{T} = (\tau_{jk})_{K\times K}$. Also, the mean vector $\boldsymbol{\mu}$ can be assigned a K-variate Normal$_K(\boldsymbol{\xi}, \boldsymbol{\Xi})$ hyper-prior distribution, and the inverse covariance matrix $\boldsymbol{T}^{-1}$ can be assigned a Wishart hyper-prior distribution Wishart$_K(\nu, \boldsymbol{\Sigma}^{-1})$. Thus, here, the hyper-prior $\pi(\boldsymbol{\Psi}|\boldsymbol{\Gamma})$ for $\boldsymbol{\Psi} = (\boldsymbol{\mu}, \boldsymbol{T})$ is parameterized by $\boldsymbol{\Gamma} = (\boldsymbol{\xi}, \boldsymbol{\Xi}, \nu, \boldsymbol{\Sigma}^{-1})$.

3. MCMC sampling and model selection

3.1. Gibbs sampling $(G, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{T})$

The Gibbs sampling algorithm relies on the introduction of strategic latent variables (Walker, 2007; Kalli et al., 2010), as described below. The starting form is

$$f_G(c|\mathcal{C}, \mathbf{x}) = \int \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta})\, \mathrm{d}G(\boldsymbol{\beta}), \tag{2}$$

for a set of choice-alternatives for $c = 1, \ldots, C$, where G is a stick-breaking process given by

$$G(\boldsymbol{\beta}) = \sum_{j=1}^{\infty} w_j \delta_{\boldsymbol{\beta}_j}$$

so that

$$f_G(c|\mathcal{C}, \mathbf{x}) = \sum_{j=1}^{\infty} w_j \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta}_j),$$

where the random $\{w_j\}$ are obtained as described in Section 2.

We first introduce the latent variable u such that

$$f_G(c, u|\mathcal{C}, \mathbf{x}) = \sum_{j=1}^{\infty} \mathbb{I}(u < w_j) \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta}_j),$$

with $\mathbb{I}(\cdot)$ the indicator function, and the importance here is that the number of w_j greater than u is finite. Note that integrating over u with respect to the Lebesgue measure returns us $f_G(c|\mathcal{C}, \mathbf{x})$. Alternatively, it is possible to write $f_G(c, u|\mathcal{C}, \mathbf{x}) = \sum_{j=1}^{\infty} w_j \mathrm{U}(u|0, w_j) \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta}_j)$, and so with probability w_j, u and c are independent, and the uniform random variate u has marginal density $\sum_{j=1}^{\infty} w_j \mathrm{U}(u|0, w_j) = \sum_{j=1}^{\infty} \mathbb{I}(u < w_j)$ (Walker, 2007). We now introduce another latent variable d which picks out the component from which c comes;

$$f_G(c, u, d|\mathcal{C}, \mathbf{x}) = \mathbb{I}(u < w_d) \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta}_d),$$

and marginalizing over (u, d) yields a density $f_G(c|\mathcal{C}, \mathbf{x})$ that agrees with our model in Eq. (2). Hence the likelihood function based on n observations is given by

$$\prod_{i=1}^{n} \mathbb{I}(u_i < w_{d_i}) \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_{d_i}).$$

The Gibbs sampler can now be described. Starting with the (d_i), we sample $[v, u]$ by sampling $[v]$ then $[u|v]$. Hence, the conditional for v_j; for $j = 1, \ldots, M$, where $M = \max_i d_i$, is beta$(a_j + n_j, b_j + m_j)$, where $n_j = \sum_i \mathbb{I}(d_i = j)$ and $m_j = \sum_i \mathbb{I}(d_i > j)$. We then sample $u_i \sim \mathrm{Uniform}(0, w_{d_i})$.

Before discussing the sampling of the $(\boldsymbol{\beta}_j)$ and how many more of the v_j we need to sample, let us consider the sampling of the (d_i). Now

$$\Pr(d_i = j) \propto \mathbb{I}(w_j > u_i) \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_j).$$

Hence we need to sample as many v_js until we are sure we have all the w_j which are greater than u_i. We will know we have all of these when we have N_i such that

$$\sum_{j=1}^{N_i} w_j > 1 - u_i.$$

Hence, we keep sampling $v_{M+1}, \ldots, V_N$ and also sample $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$, where $N = \max_i N_i$. Note that for $j < M + 1$ we have

$$[\boldsymbol{\beta}_j] \propto \prod_{d_i=j} \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_j)\, h(\boldsymbol{\beta}_j)$$

and this can be sampled via a random-walk Metropolis–Hastings algorithm. For $j > M$, $[\boldsymbol{\beta}_j] \propto h(\boldsymbol{\beta}_j)$ which can be sampled directly. Note that the new M will be less than the old N and hence the Metropolis–Hastings algorithm can always be implemented.

Please cite this article in press as: Karabatsos, G., Walker, S.G., Bayesian nonparametric mixed random utility models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.014

In order to obtain samples from the predictive distribution of $\boldsymbol{\beta}$, we can, at each iteration of the Gibbs sampler, sample from the random G. In particular, we sample a ρ from the uniform distribution on (0, 1) and take $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_j$ if $\sum_{l=1}^{j-1} w_l < \rho < \sum_{l=1}^{j} w_l$, otherwise, if $\rho > \sum_{l=1}^{N} w_l$, then we take $\boldsymbol{\beta}$ to be distributed as $h(\boldsymbol{\beta})$.

As mentioned in Section 2, in the stick-breaking prior, we can specify the baseline distribution H as the K-variate, $\text{Normal}_K(\boldsymbol{\mu}, \boldsymbol{T})$ distribution, with a $\text{Normal}_K(\boldsymbol{\xi}, \boldsymbol{\Xi})$ hyper-prior distribution assigned to $\boldsymbol{\mu}$, and a $\text{Wishart}_K(\nu, \boldsymbol{\Sigma}^{-1})$ hyper-prior distribution assigned to $\boldsymbol{T}^{-1}$. According to standard Bayesian theory for the multivariate normal distribution (Evans, 1965), the full conditional posterior distribution of $\boldsymbol{\mu}$ is multivariate normal, and the full conditional posterior of $\boldsymbol{T}^{-1}$ is Wishart.

3.2. MCMC algorithm

We describe the MCMC sampling methods in terms of a 8-step algorithm, given below.

(Initialization) Initialize with starting values, by setting $s = 0$, $\boldsymbol{\mu}^{(0)} = \mathbf{0}$, $\boldsymbol{T}^{(0)} = \text{diag}_K(10)$, by drawing $d_i^{(0)} \sim \text{Uniform}\{1, \ldots, n\}$ for $i = 1, \ldots, n$, and by drawing $\boldsymbol{\beta}_j^{(0)} \sim \text{Normal}_K(\boldsymbol{\mu}^{(0)}, \boldsymbol{T}^{(0)})$ for $j = 1, \ldots, M^{(s)} = \max_i d_i^{(0)}$.

(Step 1) Set $s = s + 1$.

(Step 2) Draw mixture weights $w_j^{(s)}$ from their full-conditional posterior distribution using the following slice sampling method (Walker, 2007; Kalli et al., 2010). For $j = 1, \ldots, M^{(s)} = \max_i d_i^{(s-1)}$, take $n_j^{(s)} = \sum_i \mathbb{I}(d_i^{(s)} = j)$, $m_j^{(s)} = \sum_i \mathbb{I}(d_i^{(s)} > j)$, draw $u_i^{(s)} \sim \text{Uniform}(0, w_{d_i^{(s-1)}}^{(s)})$ for $i = 1, \ldots, n$, and then draw $v_j^{(s)} \sim \text{Beta}(a_j + n_j^{(s)}, b_j + m_j^{(s)})$, for $j = 1, 2, \ldots, M^{(s)}, M^{(s)} + 1, \ldots$, until an integer value $N^{(s)}$ is found to satisfy $\sum_{j=1}^{N^{(s)}} w_j^{(s)} > \max_i(1 - u_i^{(s)})$, where $w_j^{(s)} = v_j^{(s)} \prod_{l<j}(1 - v_j^{(s)})$ for $j = 1, 2, \ldots, N^{(s)}$.

(Step 3) Update each $\boldsymbol{\beta}_j^{(s)}$ $(j = 1, \ldots, M^{(s)})$ with a Metropolis–Hastings algorithm, by first drawing $\boldsymbol{\beta}_j^*$ from a $\text{Normal}_K(\mathbf{0}, \text{diag}(\xi_{\text{prop}}))$ proposal distribution (e.g., $\xi_{\text{prop}} = 0.1$), and then taking:

$$\boldsymbol{\beta}_j^{(s)} = \begin{cases} \boldsymbol{\beta}_j^* & \text{with probability } \Pr(\boldsymbol{\beta}_j^{(s-1)}, \boldsymbol{\beta}_j^*) \\ \boldsymbol{\beta}_j^{(s-1)} & \text{with probability } 1 - \Pr(\boldsymbol{\beta}_j^{(s-1)}, \boldsymbol{\beta}_j^*), \end{cases}$$

where:

$$\Pr(\boldsymbol{\beta}_j^{(s-1)}, \boldsymbol{\beta}_j^*) = \min\left[1, \frac{\prod_{d_i=j} \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_j^*)\, \text{normal}_K(\boldsymbol{\beta}_j^*|\boldsymbol{\mu}^{(s-1)}, \boldsymbol{T}^{(s-1)})}{\prod_{d_i=j} \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_j^{(s-1)})\, \text{normal}_K(\boldsymbol{\beta}_j^{(s-1)}|\boldsymbol{\mu}^{(s-1)}, \boldsymbol{T}^{(s-1)})}\right].$$

Also, if $N^{(s)} > M^{(s)}$, then draw $\boldsymbol{\beta}_j^{(s)} \sim \text{Normal}_K(\boldsymbol{\mu}^{(s-1)}, \boldsymbol{T}^{(s-1)})$, for $j = M^{(s)} + 1, \ldots, N^{(s)}$. The proposal variance ξ_{prop} should be chosen to yield an acceptance rate of about 0.44 when K is 1 or 2, and an acceptance rate of about 0.234 when K is greater than 2 (Roberts and Rosenthal, 2001). An adaptive Metropolis method (Atchadé and Rosenthal, 2005) is used to estimate the proposal variance that yields the desired acceptance rate, over MCMC iterations.

(Step 4) Sample $\boldsymbol{\beta}^{(s)}$ from G, i.e., from its posterior predictive distribution, by drawing $\rho^{(s)} \sim \text{Uniform}(0, 1)$ and then taking $\boldsymbol{\beta}^{(s)} = \boldsymbol{\beta}_j^{(s)}$ when $\sum_{l=1}^{j-1} w_l^{(s)} < \rho^{(s)} < \sum_{l=1}^{j} w_l^{(s)}$, otherwise, taking $\boldsymbol{\beta}^{(s)}$ as a draw from a $\text{Normal}_K(\boldsymbol{\mu}^{(s-1)}, \boldsymbol{T}^{(s-1)})$ distribution.

(Step 5) For $i = 1, \ldots, n$, sample $d_i^{(s)} = j$ with probability proportional to $\mathbb{I}(w_j > u_i^{(s)}) \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_j^{(s)})$, $j = 1, \ldots, N^{(s)}$.

(Step 6) Draw $\boldsymbol{\mu}^{(s)} \sim \text{Normal}_K(\boldsymbol{\Xi}^*(n_{\text{clus}}^{(s)}(\boldsymbol{T}^{-1})^{(s-1)}\widehat{\boldsymbol{\mu}}_{\text{clus}}^{(s)} + \boldsymbol{\Xi}^{-1}\boldsymbol{\xi}), \boldsymbol{\Xi}^*)$, where $\boldsymbol{\Xi}^* = (n_{\text{clus}}^{(s)}(\boldsymbol{T}^{-1})^{(s-1)} + \boldsymbol{\Xi}^{-1})^{-1}$, $n_{\text{clus}}^{(s)}$ is the number of distinct vectors in the set $\{\boldsymbol{\beta}_{d_i}^{(s)}\}_{i=1}^n$, and $\widehat{\boldsymbol{\mu}}_{\text{clus}}^{(s)}$ denotes the mean of these vectors (Evans, 1965; Escobar and West, 1998).

(Step 7) Draw $(\boldsymbol{T}^{-1})^{(s)} \sim \text{Wishart}_K(\nu + n_{\text{clus}}^{(s)} + K - 1, \{\boldsymbol{\Sigma} + (\boldsymbol{B}^{(s)} - \boldsymbol{M}^{(s)})'(\boldsymbol{B}^{(s)} - \boldsymbol{M}^{(s)})\}^{-1})$, where $\boldsymbol{B}^{(s)}$ is the $n_{\text{clus}}^{(s)} \times K$ matrix of the $n_{\text{clus}}^{(s)}$ distinct vectors in $\{\boldsymbol{\beta}_{d_i}^{(s)}\}_{i=1}^n$, and $\boldsymbol{M}^{(s)}$ is the $n_{\text{clus}}^{(s)} \times K$ matrix of row vectors $\boldsymbol{\mu}'^{(s)}$ (Evans, 1965; Escobar and West, 1998).

(Step 8) Repeat Steps 1 though 7 until the resulting Markov chain $\{(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{T})^{(s)}\}_{s=1}^{S-1}$ has mixed well, and provides acceptably-small standard errors for the univariate marginal posterior estimates of $(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{T})$.

Recall that the DP is a special case of the general stick-breaking process, i.e., is defined by $a_j = 1 - a$ and $b_j = b + ja$ for all $j = 1, 2, \ldots$, with $a = 0$ and $b = \alpha$. In this case, α may also be assigned a prior distribution, such as a gamma distribution with shape and scale parameters (a_α, b_α). Then the full conditional posterior distribution of α is a gamma distribution with shape $a_\alpha + n_{\text{clus}} - \mathbb{I}(u > \{O/(1 + O)\})$ and scale $b_\alpha - \log(\eta)$, given draws $\eta \sim \text{Beta}(\alpha + 1, n)$, $u \sim \text{Uniform}(0, 1)$, and $O = (a_\alpha + n_{\text{clus}} - 1)/(\{b_\alpha - \log(\eta)\}n)$ (Escobar and West, 1995, p. 584). Therefore, for an MMNL model assigned a DP prior, with a gamma prior for the precision parameter α, the full conditional of α can be easily sampled through an extra step in the MCMC sampling algorithm.

The MCMC algorithm described in this section assumes that each individual makes a single choice from the choice set $\mathcal{C}$. However, the algorithm can be easily extended to a setting where each individual makes a sequence of $T_i \geq 1$ choices

Please cite this article in press as: Karabatsos, G., Walker, S.G., Bayesian nonparametric mixed random utility models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.014

Table 1
Summary statistics of the 10 choice alternatives and family background variables, for the sample of 788 individuals from 100 families.

Choice	Count	Price Mean (s.d.)	Choice	Count	Price Mean (s.d.)
Pk_Stk	336	0.52 (0.15)	Imp_Stk	12	0.79 (0.13)
BB_Stk	128	0.55 (0.12)	SS_Tub	55	0.82 (0.07)
Fl_Stk	50	1.01 (0.04)	Pk_Tub	62	1.08 (0.03)
Hse_Stk	71	0.44 (0.12)	Fl_Tub	35	1.19 (0.01)
Gen_Stk	35	0.34 (0.03)	Hse_Tub	4	0.57 (0.07)
	Income	FamSize	College	Whitecollar	Retired
Mean (s.d.)	26.03 (16.04)	3.01 (1.30)	0.29 (0.46)	0.47 (0.50)	0.30 (0.46)

from $\mathcal{C}$. In this case, each instance of the term $\Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_j^*)$ in the MCMC algorithm is simply replaced with the term $\prod_{t=1}^{T_i} \Pr(c_{i_t}|\mathcal{C}, \mathbf{x}_{i_t}, \boldsymbol{\beta}_j^*)$, for $i = 1, \ldots, n$. Otherwise, the sampling algorithm proceeds the same manner.

We have written code in MATLAB (2010, The MathWorks, Natick, MA) to implement this entire Gibbs sampling algorithm. It can be requested from the corresponding author.

3.3. Model selection

Given a set of data $\boldsymbol{D} = \{(c_i, \mathbf{x}_i)\}_{i=1}^n$, it may be of interest to compare the predictive utility between versions of the MMNL model, which differ in their choice of prior for the mixing distribution, G. The predictive utility of a model can be assessed using the log-predictive marginal likelihood (LPML), which is based on leave-one-out cross-validation (Geisser and Eddy, 1979). The LPML, defined here by $\text{LPML} = \sum_{i=1}^n \log \int \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\gamma}) \Pi(\mathrm{d}\boldsymbol{\gamma}|\boldsymbol{D}_{-i})$ where $\boldsymbol{\gamma}$ denotes a model's parameter vector and $\boldsymbol{D}_{-i}$ denotes the data set excluding data point $(c_i, \mathbf{x}_i)$, evaluates a model's predictive power of future data point y_i given $\mathbf{x}_i$ and parameters fit to the $n-1$ remaining data points, for all data points indexed by $i = 1, \ldots, n$. Also, $\frac{1}{n}$LPML is an estimate of a model's expected posterior predictive utility under the logarithmic utility function (Bernardo and Smith, 1994, Section 6.1.6). Given a set of S samples from the MCMC algorithm, the LPML estimate is $\sum_{i=1}^n \log(\{\frac{1}{S}\sum_{s=1}^S \Pr(c_i|\mathcal{C}, \mathbf{x}_i, \boldsymbol{\beta}_{d_i}^{(s)})^{-1}\}^{-1})$ (Gelfand, 1995). In contrast to the LPML, the Bayes factor compares models by evaluating each model on the ability to predict the full data set $\boldsymbol{D}$, given no data and the model's prior (Bernardo and Smith, 1994, Section 6.1.6). Therefore, the Bayes factor is quite sensitive to the choice of prior distribution. In this study, we will use the LPML to compare models, because it is less sensitive to the choice of prior, and is far easier to estimate for nonparametric models.

4. Illustration

Here we analyze panel data on choices among 10 margarine alternatives, including Parkay (Pk), Blue Bonnet (BB), Fleischmanns (Fl), house (Hse), generic (Gen), Imperial (Imp), Shed Spread (SS), each of which are either in stick (Stk) or tub (Tub) form. The data are described in Allenby and Rossi (1991), and were obtained from the bayesm R package. Table 1 presents the descriptive statistics on 788 individuals from 100 randomly-selected families, with each individual making a single choice among the 10 alternatives. The aim of the data analysis is to infer how choices depend on purchase price, family income, family size, and three variables indicating (0–1) whether the family head of household is college educated, white collar, and retired, respectively. Each of the family variables are entered into the model such that they differ across the 10 choices; for a given choice alternative, the family variable takes on its original value if the individual made that choice, and equals zero otherwise.

We analyzed the data using three versions of the MMNL model, which differ in their choice of prior for the mixing distribution G. The first model is defined by a Pitman–Yor stick-breaking process with parameters $(a = 1/4, b = 10)$, for $a_j = 1 - a$ and $b_j = b + ja, j = 1, 2, \ldots$. The second model assigns a weakly-informative DP prior, defined by $(a = 0, b = \alpha)$, with prior $\alpha \sim \text{Gamma}(0.01, 0.01)$ prior. In both models, the baseline distribution H was specified as the six-variate $\text{Normal}_6(\boldsymbol{\mu}, \boldsymbol{T})$ distribution, with a $\text{Normal}_6(\mathbf{0}, \text{diag}(100))$ hyper-prior for $\boldsymbol{\mu}$, and a $\text{Wishart}_6(1, \text{diag}(10))$ hyper-prior for $\boldsymbol{T}^{-1}$. Both of these nonparametric models allow for individuals to cluster into groups on common values of $\boldsymbol{\beta}$, across families. The third model assumes a $\text{Normal}_6(\boldsymbol{\mu}, \boldsymbol{T})$ prior for G, with the same hyper-priors for $(\boldsymbol{\mu}, \boldsymbol{T})$. Each of the three models were fit using 50,000 MCMC iterations. The posterior of the normal mixture model was sampled using the MCMC algorithm described in Train (2003, Chapter 12).

For the MMNL model assigned Pitman–Yor stick-breaking prior, Fig. 1 presents the trace plots of several key elements of the MCMC algorithm, and shows that the samples of all the parameters seemed to have stabilized after 30,000 MCMC iterations. The samples of the two other models showed similar sampling patterns. All results described below are based on 20,000 additional MCMC samples. For the MMNL model assigned Pitman–Yor prior, Fig. 2 presents univariate marginal density estimates of the mixing distribution G. The univariate marginals of G seem to exhibit skewness, heavy tails, and multiple modes. Also, the figure shows that all covariates tended to be correlated with choice, with positive association shown by price, income, college education, white collar status, and retirement status, and negative association shown by family size. Table 2 presents the marginal posterior mean estimates of the mean vector $\boldsymbol{\mu}$ and variance–covariance $\boldsymbol{T}$, along

Please cite this article in press as: Karabatsos, G., Walker, S.G., Bayesian nonparametric mixed random utility models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.014

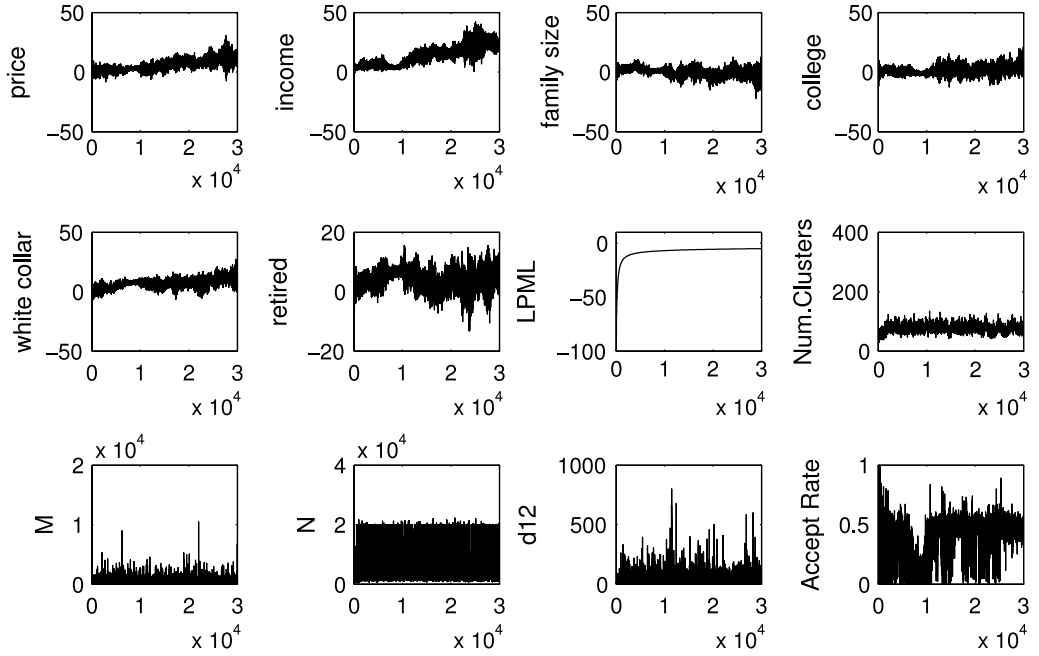

Fig. 1. For key random elements, trace plots for the first 30,000 MCMC samples. A covariate name (e.g., price) refers to the random coefficient β for that covariate. "Accept Rate" refers to the proportion of the proposed $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)^{(s)}$ that is accepted under the Metropolis algorithm, for each MCMC iteration s.

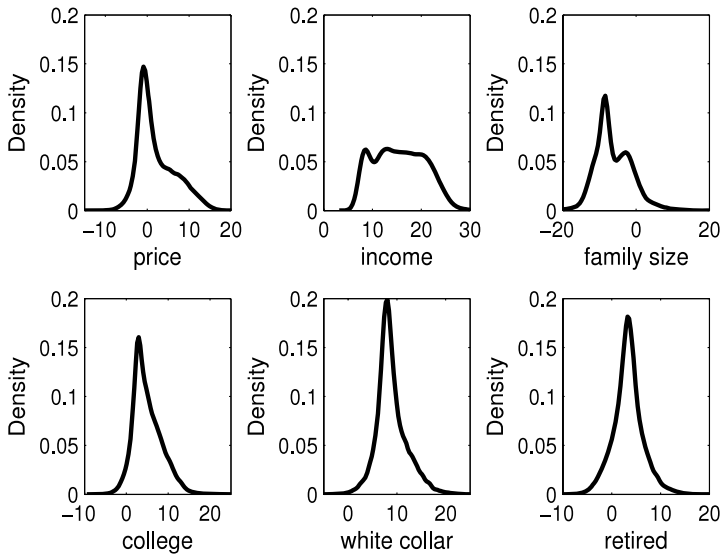

Fig. 2. Marginal density estimates of G, based on the last 20,000 samples of a run of 50,000 MCMC iterations.

with 95% Monte Carlo confidence intervals computed via the batch means method (Flegal and Jones, 2011). Also, under this model, among the 788 individuals, the posterior mean estimate of the number of distinct clusters of $\boldsymbol{\beta}$ was 80 ($\pm$1.35), with posterior standard deviation estimate 12.26 ($\pm$0.66).

The MMNL model with Pitman–Yor stick-breaking prior had an LPML of -0.11 ($\pm$0.01), and performed best in terms of predictive utility, compared to the MMNL model assigned a DP prior (LPML $= -0.85 \pm 0.02$), and compared to the MMNL model assigned a multivariate normal prior (LPML $= -1.11 \pm 0.03$). For the MMNL model assigned a DP prior, the precision parameter α had posterior mean estimate 462.31 ($\pm$0.46) and posterior standard deviation estimate of 36.18 ($\pm$0.53). Table 3 compares the posterior mean estimate of the 10 choice probabilities across all three models analyzed, conditional on different values of the covariate vector, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_c, \ldots, \mathbf{x}_{10})$. For illustration, we considered four covariate vectors denoted by $\mathbf{x}_{\min}$, $\mathbf{x}_{0.25}$, $\mathbf{x}_{0.50}$, $\mathbf{x}_{\max}$, which are respectively defined by the component-wise empirical minimum, first quartile, median, and maximum of the 6 individual covariates. From this table we see that the estimated choice probabilities of the

Please cite this article in press as: Karabatsos, G., Walker, S.G., Bayesian nonparametric mixed random utility models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.014

Table 2
Marginal posterior mean estimates of the mean and variance–covariance parameters of the baseline distribution (95% Monte Carlo confidence interval).

	Price	Income	Famsize	College	Whitecollar	Retired
Mean $\boldsymbol{\mu}$	2.14 (±0.64)	15.58 (±0.77)	−6.04 (±0.57)	5.06 (±0.39)	8.86 (±0.33)	3.27 (±0.29)
Variance–covariance $\boldsymbol{T}$:						
Price	6.28 (±0.87)					
Income	−0.14 (±0.44)	4.93 (±0.66)				
Famsize	1.03 (±1.09)	−0.42 (±0.65)	15.61 (±3.66)			
College	1.22 (±0.58)	−0.35 (±0.47)	3.47 (±1.78)	7.13 (±1.30)		
Whitecollar	−0.46 (±0.44)	0.19 (±0.41)	−2.66 (±1.26)	−2.41 (±0.87)	6.06 (±0.83)	
Retired	−0.57 (±0.56)	−0.46 (±0.45)	−4.57 (±1.46)	−2.04 (±0.90)	1.33 (±0.72)	6.63 (±0.91)

Table 3
Posterior mean estimates of choice probabilities, for different covariate vectors x. They are compared between the stick-breaking (Pitman–Yor) mixture model (labeled S), the Dirichlet process mixture model (D), and the normal mixture model (N). In general, the posterior mean estimates each had a 95% Monte Carlo Confidence interval with half-width not exceeding 0.01.

	Probability of choice, given $\mathbf{x}$											
	$\mathbf{x}_{\min}$			$\mathbf{x}_{0.25}$			$\mathbf{x}_{0.50}$			$\mathbf{x}_{\max}$		
	S	D	N	S	D	N	S	D	N	S	D	N
Pk_Stk	0.08	0.12	0.20	0.07	0.11	0.13	0.07	0.11	0.12	0.00	0.14	0.12
BB_Stk	0.08	0.12	0.20	0.07	0.11	0.13	0.07	0.11	0.12	0.00	0.06	0.02
Fl_Stk	0.10	0.08	0.01	0.10	0.08	0.02	0.10	0.09	0.03	0.15	0.31	0.38
Hse_Stk	0.07	0.12	0.20	0.09	0.13	0.26	0.08	0.12	0.19	0.00	0.08	0.06
Gen_Stk	0.07	0.11	0.16	0.09	0.12	0.23	0.10	0.13	0.29	0.00	0.16	0.22
Imp_Stk	0.06	0.09	0.03	0.07	0.10	0.06	0.07	0.10	0.07	0.11	0.04	0.06
SS_Tub	0.06	0.10	0.07	0.07	0.09	0.04	0.07	0.09	0.05	0.00	0.06	0.02
Pk_Tub	0.12	0.08	0.01	0.13	0.08	0.02	0.14	0.08	0.02	0.00	0.02	0.03
Fl_Tub	0.28	0.07	0.01	0.25	0.08	0.01	0.24	0.08	0.01	0.83	0.13	0.15
Hse_Tub	0.01	0.11	0.12	0.01	0.11	0.10	0.07	0.11	0.00	0.00	0.00	0.00

best-fitting stick-breaking MMNL model can differ substantially from the corresponding choice probabilities of the DP and normal MMNL models.

5. Posterior consistency

Here we establish consistency for the model

$$f_G(c|\mathbf{x}) = \int \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta})\, \mathrm{d}G(\boldsymbol{\beta})$$

with G assigned any given nonparametric prior. By this we mean that the posterior distribution for G accumulates in a neighborhood of G_0, the assumed true mixing distribution, as the sample size increases. Without this, there seems little point in collecting more data, and statistical inference, if it could use large data sets, the prohibitive factor typically being cost, would do so. But the crucial aspect then is that the posterior can be shown to move toward G_0, rather than be left in the strange situation whereby more data are being purchased yet the posterior is heading off in the wrong direction. Such a phenomena is possible when dealing with infinite dimensional parameters such as a random distribution function.

We need to establish what exactly accumulate means here and we will use a distance between the density functions; namely between f_G and f_{G_0} and will make this the L_1 distance.

We will assume that $\mathbf{x} \sim m$, where m is a density function on the covariates, and show that consistency holds in the sense that the posterior distribution has the convergence property

$$\Pi\left(G : d_1(f_G, f_{G_0}) > \epsilon | \{(\mathbf{x}_i, c_i)\}_{i=1}^n\right) \to 0 \quad \text{a.s.}$$

for all $\epsilon > 0$. Here, we slightly modify f_G to be $f_G(\mathbf{z}) = f_G(c|\mathbf{x})\, m(\mathbf{x})$, where $\mathbf{z} = (\mathbf{x}, c)$, and d_1 denotes the L_1 distance between probability density functions.

On the other hand, if a single normal distribution is chosen for G, then a prior is put on the mean and covariance matrix. If the true mixing distribution is normal then no problems will arise. However, this is rarely known and to cover deviations from normal it is wise and safe to use a larger model. A stick-breaking prior is one such model and the inference via MCMC is extremely possible and not difficult.

Since $\Pr(\cdot|\cdots)$ is bounded by 1 we can work out the result of consistency using material to be found in Walker et al. (2005).

Please cite this article in press as: Karabatsos, G., Walker, S.G., Bayesian nonparametric mixed random utility models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.014

On the assumption that $\kappa(\mathbf{z}|\boldsymbol{\beta}) = \Pr(c|\mathcal{C}, \mathbf{x}, \boldsymbol{\beta})m(\mathbf{x})$ is continuous in $\mathbf{z}$, we know it is already known to be bounded, and hence this establishes that if G_N converges weakly to G, for some sequence of distribution functions (G_N), then $f_{G_N}(\mathbf{z})$ converges point-wise to $f_G(\mathbf{z})$ and hence, by Scheffé's Theorem, f_{G_N} converges to f_G with respect to the L_1 distance.

The converse is also easy to see; namely that if G' is outside a weak neighborhood of G_0 then $f_{G'}$ is outside an L_1 neighborhood of f_G. Section 3.2 of Walker et al. (2005) then establishes the required consistency result for all G_0 in the Kullback–Leibler support of the prior; that is for all G_0 satisfying

$$\Pi(G : d_K(f_G, f_{G_0}) > \epsilon) > 0$$

for all $\epsilon > 0$, and where d_K denotes the Kullback–Leibler distance between density functions.

This result could also be obtained using results from Kiefer and Wolfowitz (1956). It is clear that $\widehat{G}$, the MLE, exists and given the nature of $\kappa(\mathbf{z}|\boldsymbol{\beta})$ it is not difficult to see that the assumptions required for

$$\sup_{d^*(G,G_0)>\rho} \frac{f_G(\mathbf{z}_i)}{f_{G_0}(\mathbf{z}_i)} < h_\rho^n \quad \text{a.s.}$$

for all large n for some $0 < h_\rho < 1$ are satisfied when

$$-\int \log f_{G_0}(\mathbf{z})\, f_{G_0}(\mathbf{z})\, \mathrm{d}\mathbf{z} < +\infty.$$

Here $d^*(G, G_0) = \int |G(\boldsymbol{\beta}) - G_0(\boldsymbol{\beta})| e^{-|\boldsymbol{\beta}|}\, \mathrm{d}\boldsymbol{\beta}$. Since the posterior for the mass assigned to $A_\epsilon = \{G : d^*(G, G_0) > \epsilon\}$ is given by

$$\frac{\int_{A_\epsilon} \prod_{i=1}^n f_G(\mathbf{z}_i)/f_{G_0}(\mathbf{z}_i)\, \Pi(\mathrm{d}G)}{\int \prod_{i=1}^n f_G(\mathbf{z}_i)/f_{G_0}(\mathbf{z}_i)\, \Pi(\mathrm{d}G)}$$

we note that the numerator can be bounded above by h_ϵ^n. The Kullback–Leibler support including G_0 yields the denominator being bounded below by e^{-nc} a.s. for all large n for any $c > 0$. Hence, choosing $c < -\log h_\epsilon$ we have that

$$\Pi(G : d_1(f_G, f_{G_0}) > \epsilon | \{(\mathbf{x}_i, c_i)\}_{i=1}^n) \to 0 \quad \text{a.s.}$$

for all $\epsilon > 0$. Hence, provided the prior puts positive mass around G_0 then consistency is assured. It seems obvious that such a condition is needed since of the prior does not put positive mass around G_0, then neither can the posterior and so consistency would be impossible.

6. Conclusions

We introduced a general Bayesian nonparametric approach to perform flexible statistical inference with a general random utility model, and illustrated the approach through applications of the MMNL model, which is popular in econometrics. Compared to previous MMNL models assigned either a DP or a normal prior, we demonstrated that a more general stick-breaking prior can provide an MMNL model with better predictive fit, while maintaining consistency in the posterior distribution.

Acknowledgments

We thank the Editor, the anonymous reviewers and the Associate Editor for comments that have helped improve the presentation of the manuscript. The writing of this article began on September 2008, during a sabbatical visit of the first author to the second author. This work was partially supported by NSF grant SES-0242030 from the Program in Methodology, Measurement, and Statistics, awarded to the first author as principal investigator.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.10.014.

References

Allenby, G., Rossi, P., 1991. Quality perceptions and asymmetric switching between brands. Marketing Science 10, 185–204.
Atchadé, Y., Rosenthal, J., 2005. On adaptive Markov chain Monte Carlo algorithms. Bernoulli 11, 815–828.
Barron, A., Schervish, M., Wasserman, L., 1999. The consistency of posterior distributions in nonparametric problems. Annals of Statistics 27, 536–561.
Bernardo, J., Smith, A., 1994. Bayesian Theory. Wiley, Chichester, England.
Burda, M., Harding, M., Hausman, J., 2008. A Bayesian mixed logit-probit model for multinomial choice. Tech. Rep.. University of Toronto: Department of Economics.
Escobar, M., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.

Escobar, M., West, M., 1998. Computing nonparametric hierarchical models. In: Dey, D., Müller, P., Sinha, D. (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics. Springer, New York, pp. 1–22.
Evans, I., 1965. Bayesian estimation of parameters of a multivariate normal distribution. Journal of the Royal Statistical Society: Series B 27, 279–283.
Ferguson, T., 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1, 209–230.
Flegal, J., Jones, G., 2011. Implementing Markov chain Monte Carlo: estimating with confidence. In: Brooks, S., Gelman, A., Jones, G., Meng, X. (Eds.), Handbook of Markov Chain Monte Carlo. CRC, Boca Raton, FL, pp. 175–197.
Geisser, S., Eddy, W., 1979. A predictive approach to model selection. Journal of the American Statistical Association 74, 153–160.
Gelfand, A., 1995. Model determination using sampling-based methods. In: Gilks, W., Richardson, S., Spiegelhalter, D. (Eds.), Markov Chain Monte Carlo in Practice. Chapman and Hall, London, pp. 145–161.
Ishwaran, H., James, L., 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96, 161–173.
Kalli, M., Griffin, J., Walker, S., 2010. Slice sampling mixture models. Statistics and Computing 21, 93–105.
Kiefer, J., Wolfowitz, J., 1956. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. Annals of Mathematical Statistics 27, 887–906.
McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15, 447–470.
Pitman, J., 1996. Some developments of the Blackwell–MacQueen urn scheme. In: Ferguson, T., Shapeley, L., MacQueen, J. (Eds.), Statistics, Probability and Game Theory. Papers in Honor of David Blackwell. Institute of Mathematical Sciences, Hayward, CA, pp. 245–268.
R Development Core Team, 2011. R: a language and environment for statistical computing. Vienna, Austria.
Roberts, G., Rosenthal, J., 2001. Optimal scaling of various Metropolis–Hastings algorithms. Statistical Science 16, 351–367.
Rossi, P., 2011. Bayesm: Bayesian inference for marketing/micro-econometrics. R package version 2.2-4.
Rossi, P., Allenby, G., McCulloch, R., 2005. Bayesian Statistics and Marketing. John Wiley and Sons, New York.
Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639–650.
Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press, Cambridge, United Kingdom.
Walker, S., 2007. Sampling the Dirichlet mixture model with slices. Communications in Statistics: Simulation and Computation 36, 45–54.
Walker, S., Lijoi, A., Prünster, I., 2005. Data tracking and the understanding of Bayesian consistency. Biometrika 92, 765–778.

Please cite this article in press as: Karabatsos, G., Walker, S.G., Bayesian nonparametric mixed random utility models. Computational Statistics and Data Analysis (2011), doi:10.1016/j.csda.2011.10.014