Online Appendices for "Modeling dynamic preferences. A Bayesian robust dynamic latent ordered probit model"

Daniel Stegmueller
Department of Government
University of Essex
Wivenhoe Park, Colchester CO4 3SQ, UK
mail@daniel-stegmueller.com

A. INITIAL OBSERVATIONS

To explicate the role of initial observations, rewrite the dynamic model

$$z_{it} = \phi z_{it-1} + \boldsymbol{x}_{it}\beta + \xi_i + \epsilon_{it}, t = 1, \ldots, T$$

in its explicit distributed lag representation by successive backward substitution (e.g., following Harris et al. 2008: 251):

$$z_{it} = \phi^t z_{i0} + \sum_{j=0}^{t-1} \phi^j \boldsymbol{x}_{it-j}\boldsymbol{\beta} + \frac{1-\phi^t}{1-\phi}\xi_i + \eta_{it} \tag{1}$$

with $\eta_{it} = \phi\eta_{it-1} + \epsilon_{it}$ with $\eta_{i0} = 0$.

This makes obvious that each observation of z_i can be expressed as the sum of several factors. The first part of equation (1), $\phi^t z_{i0}$ depends on the initial observation of the panel, while the second part depends on current and past covariate values. The third part $\frac{1-\phi^t}{1-\phi}\xi_i$ indicates proportional dependence on unobserved individual specific effects.

Direct estimation of (1) would require sufficiently large T and that ϕ^t decays sufficiently rapidly with t. Alternatively, one can specify an empirical approximation of z_{i0} (Pudney 2008: 27). Heckman's (1981) approximation for $z_{i0}|\boldsymbol{x}_{it}, \xi_i$,

$$z_{i0} = \boldsymbol{\delta}'\boldsymbol{w}_i + \lambda\xi_i + \epsilon_{i0}, \tag{2}$$

as given in the main text, is obtained by first writing

$$z_{i0} = \boldsymbol{\delta}'\boldsymbol{w}_i + \eta_i \tag{3}$$

where $\boldsymbol{w}_i = (\boldsymbol{x}_{i0}, \boldsymbol{v}_i)$ is a vector of initial condition covariates comprised of covariate values at sample entry $\boldsymbol{x}_{i0}$ and additional background information $\boldsymbol{v}_i$. η_i is an individual error component at the initial condition. Next, decompose η_i into an individual specific (time-constant) random effect and a stochastic disturbance at $t = 0$. Instead of introducing a second individual random effect, Heckman employs the orthogonal projection

$$\eta_i = \lambda\xi_i + \epsilon_{i0} \tag{4}$$

which specifies η_i as resulting from random disturbance ϵ_{i0} and individual specific effect ξ_i. The random disturbance term at the initial condition ϵ_{i0} is now uncorrelated with ξ_i by design, and assumed uncorrelated with other errors, i.e. $\text{Cov}(\epsilon_{i0}, \epsilon_{it}) = 0, \ \forall t > 0$. The individual specific random effects ξ_i are allowed to have a different scaling in the initial conditions equations by including a scale factor λ. Substituting (4) into (3) yields the reduced form equation (2) for initial observations used in the main text.

B. DIRICHLET PROCESS

In this appendix I describe the Dirichlet process in more detail.[1] A Dirichlet process random effects model can be understood as a (countably) *infinite* mixture of points. Thus I start from specifying a *finite* mixture of points model for random effects and set up the Dirichlet process model from there by letting the number of points $K \to \infty$.

A finite nonparametric random effects prior Start by specifying some flexible distribution G for the random effects:

$$\xi_i \sim G(\boldsymbol{\phi}) \tag{5}$$

with hyperparameters $\boldsymbol{\phi}$. G can be approximated arbitrarily close by specifying a finite sum of K point masses and weights π_k,

$$G(\boldsymbol{\pi}, \boldsymbol{\zeta}) = \sum_{k=1}^{K} \pi_k \delta_{\zeta_k} \tag{6}$$

with $\sum_{k=1}^{K} \pi_k = 1$ and where δ_{ζ_k} is the Dirac delta function yielding a point mass at ζ_k. Here, $\boldsymbol{\phi} = (\boldsymbol{\zeta}, \boldsymbol{\pi})$ and random effects ξ_i are sampled from this distribution and are equal to one of the ζ_k.

In a Bayesian setup (e.g. Lo 1984), one has to specify priors for the weights, such as:

$$\zeta_k \sim G_0 \tag{7}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{8}$$

where each of K discrete locations ζ_k are sampled from some base distribution G_0. The prior over weights is a Dirichlet distribution of dimension K with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$:

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \mathscr{L}(\boldsymbol{\alpha})^{-1} \left(\prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \right) \mathbf{1}(\boldsymbol{\pi}) \tag{9}$$

where $\mathbf{1}(\boldsymbol{\pi})$ is an indicator function equal to one if weights sum to one and zero otherwise. $\mathscr{L}$ is a normalizing function given by:[2]

$$\mathscr{L}(\boldsymbol{\alpha}) = \int \left(\prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \right) \mathbf{1}(\boldsymbol{\pi})\, d\boldsymbol{\pi} = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)} \tag{10}$$

[1] This section builds on the excellent presentation in Navarro et al. (2006).

[2] See, e.g. Gill (2008: 180). Γ is the gamma function, which is a generalization of the factorial function: for a non-negative integer n, $\Gamma(n) = (n-1)!$.

The Dirichlet prior for the weights $\boldsymbol{\pi}$ is taken to be symmetric, i.e. we use a parameter vector of length K with $(\alpha/K, \ldots, \alpha/K)$, thus ensuring that the sum of the parameter vector will always be α (e.g. Ishwaran and Zarepour 2002).

Moving to the infinite case Having specified a prior for the finite case, we elicit a prior specification for the infinite point mixture case by letting $K \to \infty$.

First, to make the clustering structure of the model explicit, define membership indicators s_i, which indicate to which subcluster the ith random effect is assigned. For a random effect of individual i the probability of belonging to subcluster k is given by the weight π_k, and thus

$$p(s_i = k|\boldsymbol{\pi}) = \pi_k. \tag{11}$$

Using membership indicators, the prior in (7)–(8) becomes:

$$\zeta_k \sim G_0 \tag{12}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha/K) \tag{13}$$

$$s_i \sim \text{Multinomial}(\boldsymbol{\pi}) \tag{14}$$

where membership indicators are sampled from a multinomial with size one.

Second, we integrate out the subcluster weights $\boldsymbol{\pi}$ to get the conditional subcluster assignment probability when having already observed $N-1$ random effects assignments $S_{[i]} = \{s_1, \ldots, s_{N-1}\}$:

$$p(s_i = k|S_{[i]}, \alpha, K) = \int p(s_i = k|\boldsymbol{\pi}) p(\boldsymbol{\pi}|S_{[i]}, \alpha, K) d\boldsymbol{\pi} \tag{15}$$

To solve the integral, note that the first term of the integrand is π_k (cf. equation (11)). The second term is the posterior probability

$$p(\boldsymbol{\pi}|S_{[i]}, \alpha, K) \propto p(S_{[i]}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha, K), \tag{16}$$

i.e. the product of a multinomial and Dirichlet distribution, which implies that the posterior distribution is also a Dirichlet (i.e. conjugacy of the resulting posterior).

Denote by $m_k = \#\{\xi_1 = \zeta_k\}$ the number of random effects assigned to subcluster k, and let $\boldsymbol{m} = (m_1, \ldots, m_K)$ be a 'member size' vector giving the number of individuals in each subcluster. The posterior probability $p(\boldsymbol{\pi}|\boldsymbol{s}_{/i}, \alpha, K)$ is distributed Dirichlet with

parameter vector $\boldsymbol{s} + \alpha/K$. Thus

$$p(s_i = k|\boldsymbol{s}_{/i}, \alpha, K) \tag{17}$$

$$= \mathscr{L}(\boldsymbol{m} + \alpha/K)^{-1} \int \pi_k \left(\prod_l \pi_l^{m_l + \alpha/K - 1} \right) \mathbf{1}(\pi) d\pi \tag{18}$$

$$= \frac{\mathscr{L}(\boldsymbol{m} + \alpha/K + \mathbf{1}(k))}{\mathscr{L}(\boldsymbol{m} + \alpha/K)} \tag{19}$$

$$= \frac{m_k + \alpha/K}{N - 1 + \alpha} \tag{20}$$

where $\mathbf{1}(k)$ is an indicator vector (with length K) with a 1 at position k and zero otherwise.

Having integrated out the weights, consider now the limiting probability that random effect ξ_i gets assigned value(s) ζ_k of an existing subcluster k with $m_k \geq 1$:

$$p(s_i = k|S_{[i]}, \alpha) = \lim_{K \to \infty} \left(\frac{m_k + \alpha/K}{N - 1 + \alpha} \right) \tag{21}$$

$$= \frac{m_k}{N - 1 + \alpha} \tag{22}$$

Conversely, consider the limit probability that ξ_i gets assigned values from a new subcluster. Let $K_{/i}$ be the realized number of subclusters when $N - 1$ random effects have already been assigned. Denote by $\mathscr{S}$ the set of subclusters with $m_k = 0$ (i.e. the $K - K_{[i]}$ empty subclusters). The assignment probability for the ith random effect is then

$$p(s_i \in \mathscr{S}|S_{[i]}, \alpha) = \lim_{K \to \infty} \left(\sum_{l \in \mathscr{S}} \frac{m_l + \alpha/K}{N - 1 + \alpha} \right) \tag{23}$$

$$= \frac{\alpha}{N - 1 + \alpha} \lim_{K \to \infty} \left(\frac{K - K_{[i]}}{K} \right) \tag{24}$$

$$= \frac{\alpha}{N - 1 + \alpha} \tag{25}$$

Integrating out subcluster assignment indicator variables s_i yields the prior distribution for assigning a value to random effect ξ_i given that all other random effects $\boldsymbol{\xi}_{[i]}$ have already been assigned. This distribution is a mixture of the base distribution G_0 and the empirical distribution of $N - 1$ previously assigned random effect values:

$$\xi_i|\boldsymbol{\xi}_{[i]}, \alpha, G_0 \sim \frac{\alpha}{N - 1 + \alpha} G_0 + \sum_{k=1}^{K_{[i]}} \frac{m_k}{N - 1 + \alpha} \delta_{\zeta_k}. \tag{26}$$

Drawing a sequence of random effects assignments from (26) yields a Polya urn scheme

with parameters α and G_0 (Blackwell and MacQueen 1973). Using this scheme allows us to choose a prior for the random effects distribution G. We require that the marginal prior over parameters $(\zeta_1, \ldots, \zeta_\infty)$ follows a Polya urn scheme. Blackwell and MacQueen (1973) show that the Dirichlet process does, and we can thus specify the Dirichlet process as nonparametric random effects prior:

$$\xi_i \sim G \tag{27}$$

$$G \sim DP(\alpha, G_0) \tag{28}$$

Dirichlet process The Dirichlet process is a stochastic process (a distribution over function spaces) whose sample paths (i.e. random functions draws) are probability measures with probability 1 (Ferguson 1973, 1974). Intuitively, it is a distribution over distributions, where each draw yields a Dirichlet distribution. More formally, let $(\Sigma, \mathscr{B})$ be a (measurable) space, and let G_0 be a random probability measure over it, and let α be a positive real number. A Dirichlet Process is a distribution G over $(\Sigma, \mathscr{B})$ such that for every (finite measurable) partition $(B_1, \ldots, B_N)$:

$$(G(B_1), \ldots, G(B_N)) \sim \text{Dirichlet}(\alpha G_0(B_1), \ldots, \alpha G_0(B_N)). \tag{29}$$

G_0 can be interpreted as mean of the process, since for any measurable B, $E(G(B)) = G_0(B)$. The 'dispersion', 'strength' or 'prior mass' parameter α can be understood as inverse variance, since $V(G(B)) = G_0(B)/(\alpha + 1)$, so that larger values of α imply a tighter concentration of the DP around G_0.

The posterior process for a drawing G from the DP and a subsequent random effect draw ξ_1 from G is a standard Dirichlet update (see Schervish 1995):

$$G|\xi_1 \sim DP\left(\alpha G_0 + \delta_{\xi_1}\right). \tag{30}$$

Iterating the updating yields

$$G|\xi_1, \ldots, \xi_N \sim DP\left(\alpha G_0 + \sum_{i=1}^{N} \delta_{\xi_i}\right). \tag{31}$$

To see the connection to the infinite mixture model consider the predictive distribution for a new ξ_{N+1} given previous random effect realizations $\boldsymbol{\xi}$, with G marginalized out. For any $B \subset \Sigma$ we again get the Polya/Blackwell MacQueen (1973) urn scheme (cf. equation

26):

$$E(G(B)|\xi_1,\ldots,\xi_N) = \frac{\alpha G_0(B) + \sum_{i=1}^{N} \delta_{\xi_i}(B)}{\alpha + N} \tag{32}$$

$$\rightarrow \sum_{k=1}^{\infty} \pi_k \, \delta_{\zeta_k}(B) \tag{33}$$

with $\pi_k = \lim_{N\to\infty} m_k/N$, and where ζ_k represents one unique random effect value, and $m_k = \#\{\xi_i = \zeta_k\}$ in the sequence $(\xi_1,\ldots,\xi_N)$. A countably infinite mixture of the above form, which fulfils the definition of the Dirichlet Process, can be constructed by the stick-breaking random measure, as shown by Sethuraman (1994).

Stick breaking construction It is used to construct the infinite number of weights in (33). Let

$$\nu_k \sim \text{Beta}(1,\alpha), \quad k = 1,2,\ldots \tag{34}$$

be an infinite sequence of beta distributed random variables. Set $\pi_1 = \nu_1$ and construct the remaining π_k via

$$\pi_k = \nu_k \prod_{l=1}^{k-1}(1-\nu_l), \quad k = 2,3,\ldots . \tag{35}$$

Let $\zeta_k \sim G_0$ and $G = \sum_{k=1}^{\infty} \pi_k \delta(\zeta_k)$; then $G \sim DP(\alpha, G_0)$. This constructive scheme implies that, just as the finite case in (6)–(8), G has now a clear definition as a random measure, since

$$\sum_{k=1}^{\infty} \pi_k = 1 \ \text{ wp } 1. \tag{36}$$

To see this note that

$$1 - \sum_{k=1}^{K} = 1 - \nu_1 - \nu_2(1-\nu_1) - \nu_3(1-\nu_2)(1-\nu_1) - \cdots \tag{37}$$

$$= (1-\nu_1)(1-\nu_2 - \nu_3(1-\nu_2) - \cdots) \tag{38}$$

$$= \prod_{k=1}^{K}(1-\nu_k) \tag{39}$$

$$\rightarrow 0 \ \text{ wp } 1 \text{ as } K \rightarrow \infty. \tag{40}$$

Estimation via truncated Dirichlet process The stick breaking construction suggests an approximate sampling strategy for posterior DP inference. Choose a truncation value T for K, and set $\nu_T = 1$ to ensure that weights do sum to one. Then we have a finite

representation of the infinite mixture of points:

$$G = \sum_{k=1}^{T} \pi_k \delta_{\zeta_k}, \tag{41}$$

where $\pi_k = 0$ for $k > T$. More details are given by Ishwaran and James (2001) and and Ishwaran and Zarepour (2002). This approximation yields good approximations even with low values for T, and is computationally tractable and can be implemented in available general purpose Bayesian inference packages such as JAGS, WinBUGS or PyMC. Discussions of other, more sophisticated sampling strategies (which require tailored code) are given in Escobar and West (1995), MacEachern and Müller (1998), Neal (2000), and Kyung et al. (2011).

In any 'real-life' political science application, one should check if the truncation threshold T was chosen large enough. A straightforward way is to sample from a model where T is set at twice the size, and investigate if the posterior samples of K – the sampled number of subclusters – are larger for this model. Figure 1 shows a histogram of the posterior distribution of K from just such a model run, where I set $T = 40$. It indicates that even with a higher truncation thresholds, the Dirichlet process never created more than 20 subclusters (the maximum sampled value of K is 17). Thus, the truncation level used in the main part of the paper is a good approximation.[3]

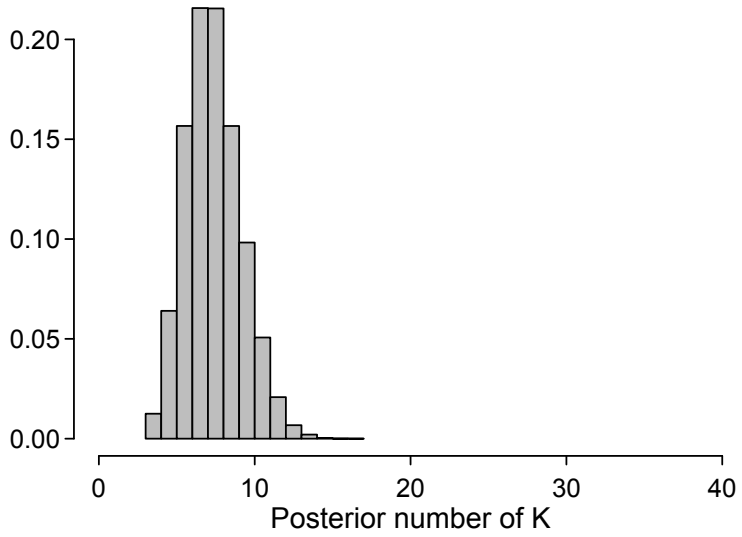

Figure 1: Posterior number of subclusters sampled from TDP(α, G_0, $T = 40$)

[3]Furthermore, inspection of parameter estimates reveled no differences to a model with $T = 20$.

C. ELICITATION OF PARAMETERS OF Γ PRIOR FOR α

Kottas et al. (2005) derive an approximation of the mean and variance of the number of subclusters, which can be used to select semi-informative values for the Gamma prior of α. The expected number of subclusters given precision, α, and number of observations, N, is

$$E(k|\alpha, N) = \sum_{i=1}^{N} \frac{\alpha}{\alpha + i - 1} \approx \alpha \log\left(\frac{\alpha + N}{\alpha}\right) \tag{42}$$

with variance

$$\mathrm{Var}(k|\alpha, N) = \sum_{i=1}^{N} \frac{\alpha(i-1)}{(\alpha + i - 1)^2} \approx \alpha \left[\log\left(\frac{\alpha + N}{\alpha}\right) - 1\right]. \tag{43}$$

As a result of my Gamma prior specification $E(\alpha) = a_0/b_0$ and $\mathrm{Var}(\alpha) = a_0/b_0^2$. Some algebra yields the *a priori* expected mean and variance for the number of subclusters (cf. Kottas et al. 2005; Liu 1996: 916):

$$E(k) \approx \frac{a_0}{b_0} \log\left(1 + \frac{n b_0}{a_0}\right) \tag{44}$$

$$\mathrm{Var}(k) \approx \frac{a_0}{b_0} \log\left(1 + \frac{n b_0}{a_0}\right) - \frac{n b_0}{a_0} + \left[\log\left(1 + \frac{n b_0}{a_0}\right) - \frac{n b_0}{a_0 + n b_0}\right]^2 \frac{a_0}{b_0^2} \tag{45}$$

This expressions can be evaluated numerically to obtain reasonable values for a_0 and b_0 given ones prior expectations of the mean number of subclusters.[4]

D. INVERSE-GAMMA VARIANCE PRIORS

As mentioned in the main text there are good reasons to prefer more informative priors for the random effect variance. In this section, I describe the specification (or 'elicitation') of two sets of hyperprior values.

Usually one specifies a prior for the inverse variance, or precision. The Gamma distribution is a popular choice (e.g. Gelman et al. 2004: 579). With given *a-priori* values for the expected mean m_0 and variance v_0 of the random effect precision σ_ξ^{-2}, hyperprior

[4] If researchers feel uncomfortable with choosing values based on expectations about K, they can either rely on priors suggested in the literature such as $\Gamma(1,1)$ or $\Gamma(2,2)$, which prevent very small and large values (Ishwaran and Zarepour 2000).

values for $\Gamma(a_0, b_0)$ are given by:[5]

$$a_0 = m_0^2/v_0 \tag{46}$$
$$b_0 = v_0/m_0 \tag{47}$$

Alternatively, when specifying a prior for the variance directly the inverse gamma distribution can be used. Here hyperprior values for $\Gamma^{-1}(a_0, b_0)$ are given by:

$$a_0 = (m_0^2 + 2v_0)/v_0 \tag{48}$$
$$b_0 = m_0(m_0^2 + v_0)/v_0 \tag{49}$$

A simple random effects ordered probit model fit using a laplace approximation to integrate out the random effects (ignoring the lagged dependent variable, and initial conditions) suggest a variance of the individual effects of ca. 1.392 or a precision of 0.7182. Thus, setting $m_0 = 0.7182$ I choose a two differently 'tight' v_0 values: $v_0 = \{1, 0.25\}$. This leads to hyperprior values of $a_0 = 0.5158, b_0 = 1.3924$, and $a_0 = 2.0632, b_0 = 0.3481$. The resulting prior distributions are illustrated in Figure 2 which plots 10,000 draws from the respective prior distributions.

Re-estimating my main model with these two more informative random effects variance prior choices leads to very similar estimated variances of 0.83 (sd=0.09) and 0.84 (sd=0.09), respectively. Coefficient estimates are virtually indistinguishable at two significant figures.

[5] I use the same notation for shape and scale of the Gamma distribution (a_0, b_0) as in the main text purely for notational convenience.

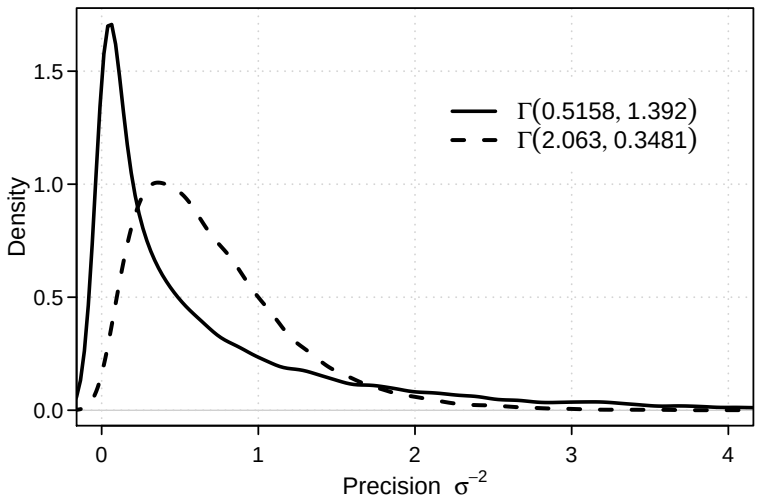

Figure 2: Distribution of variance prior precision under two Gamma prior specifications (based on 10,000 samples from prior distribution).

REFERENCES

Blackwell, David and James B MacQueen. 1973. Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics* 1:353–355.

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

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

Ferguson, Thomas S. 1974. Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics* 2:615–629.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 2004. *Bayesian Data Analysis*. Boca Raton: Chapman & Hall.

Gill, Jeff. 2008. *Bayesian Methods. A Social and Behavioral Sciences Approach*. Boca Raton: Chapman & Hall.

Harris, Mark N, Laszlo Matyas, and Patrick Sevestre. 2008. Dynamic Models for Short Panels. In *The Econometrics of Panel Data. Fundamentals and Recent Developments in Theory and Practice*, edited by Laszlo Matyas and Patrick Sevestre. Berlin: Springer, 249–278.

Heckman, James J. 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. F. Manski and Daniel McFadden. Cambridge: MIT Press, 179–195.

Ishwaran, Hemant and Lancelot F James. 2001. Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association* 96:161–173.

Ishwaran, Hemant and Mahmoud Zarepour. 2000. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* 87:371–390.

Ishwaran, Hemant and Mahmoud Zarepour. 2002. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal Of Statistics* 30:269–283.

Kottas, Athanasios, Peter Müller, and Fernando Quintana. 2005. Nonparametric Bayesian Modeling for Multivariate Ordinal Data. *Journal of Computational and Graphical Statistics* 14:610–625.

Kyung, Minjung, Jeff Gill, and George Casella. 2011. Sampling schemes for generalized linear Dirichlet process random effects models. *Statistical Methods & Applications* 20:259–290.

Liu, Jun S. 1996. Nonparametric Hierarchical Bayes Via Sequential Imputations. *The Annals of Statistics* 24:911–930.

Lo, Albert Y. 1984. On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* 12:351–357.

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

Navarro, Daniel J, Thomas L Griffiths, Mark Steyvers, and Michael D Lee. 2006. Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology* 50:101–122.

Neal, Radford M. 2000. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics* 9:249—-265.

Pudney, Stephen. 2008. The dynamics of perception: modelling subjective wellbeing in a short panel. *Journal of the Royal Statistical Society A* :21–40.

Schervish, Mark J. 1995. *Theory of Statistics*. New York: Springer.

Sethuraman, Jayaram. 1994. A Constructive Definition of Dirichlet Priors. *Statistica Sinica* 4:639–650.