Bayesian Robustness
IMS Lecture Notes - Monograph Series (1996) Volume 29

MARKOV RANDOM FIELD PRIORS FOR UNIVARIATE DENSITY ESTIMATION[1]

BY ROBERT L. WOLPERT AND MICHAEL LAVINE
Duke University Institute of Statistics & Decision Sciences

We model the unknown distribution function F of a sequence of independent real-valued random variables by partitioning the real line into intervals $\{I_i\}$ and modeling the vector $p = \{p_i\}$ of probabilities assigned to the intervals using Markov random field priors (MRFPs). We argue and illustrate that many commonly-expressed prior opinions about the shape and form of F can be expressed as statements about the joint distribution of neighboring p_i's, leading to simple MRFP expressions for prior beliefs that are awkward to express in other models. In particular, we will show how to model beliefs about continuity, monotonicity, log concavity, and unimodality of a density function f for F. The posterior distributions of the p_i's in our models (and hence the approximate predictive distributions for subsequent observations) are readily computed using Markov chain Monte-Carlo methods.

1. Introduction. We consider the problem of making inferences about or predictions of observations from an unknown probability distribution $F(\cdot)$, on the basis of expressed prior belief or opinion about the nature of $F(\cdot)$ and also of some number $n \geq 0$ of independent observations $\mathbf{X}_n = \{x_1, \ldots, x_n\}$ from the distribution.

1.1. *Conventional Approaches.* Under the assumption that $F(\cdot)$ has a probability density function (pdf) $f(\cdot)$ with respect to Lebesgue measure, with $f \in \{f_\theta : \theta \in \Theta\}$ for some parametric family, predictive inference might be based on the "plug-in" predictive density $f_{\hat{\theta}_n}(x)$, the parametric pdf evaluated at the maximum likelihood estimator $\hat{\theta}_n$; or on a Bayesian predictive distribution $f_\pi(x|\mathbf{X}_n) \propto \int_\Theta f_\theta(x) e^{\ell_n(\theta)} \pi(d\theta)$; even without the parametric assumption, predictive inference might be based on the (degenerate) empirical distribution $\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}(x)$, with mass $1/n$ at each of the n observed points $\{x_i\}$, or a kernel density estimate $\hat{f}_n^\epsilon(x) = \frac{1}{n}\sum_{i=1}^n k_\epsilon(x - x_i) = \hat{f}_n * k_\epsilon(x)$, the convolution of the empirical

[1] Research partially supported by Environmental Protection Agency grant CR822047-01-0 and National Science Foundation grants DMS-9305699 and DMS-9300137.

Key words and phrases. Nonparametric Bayesian; Hammersley-Clifford theorem; Markov chain Monte-Carlo; Metropolis.

distribution with a scaled kernel function $k_\epsilon(x) = \epsilon^{-1}k(x/\epsilon)$ for some positive $k(\cdot)$ with unit integral. But each of these four conventional approaches understates the predictive uncertainty: both $f_{\hat\theta_n}(x)$ and $f_\pi(x|\mathbf{X}_n)$ by failing to reflect uncertainty about the parametric model $f \in \{f_\theta : \theta \in \Theta\}$, and both $\hat f_n(x)$ and $\hat f_n^\epsilon(x)$ for suggesting that each future observation must equal or lie within about ±ϵ of an earlier observation.

1.2. *Nonparametric Bayesian Approaches.* For any partition of $\mathbb{R}$ into $k+1$ intervals $I_i = (b_{i-1}, b_i]$ with boundaries $b_{-1} = -\infty < b_0 < \cdots < b_k = \infty$, the probabilities $p_i = F(b_i) - F(b_{i-1})$ are uncertain, nonnegative, and sum to one:

Partition of $F(\cdot)$ into $k+1$ cells

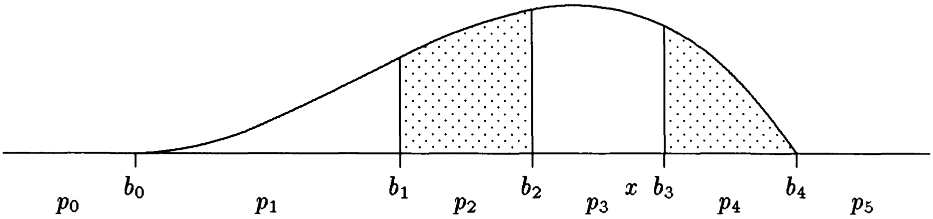


Figure 1

In an early nonparametric Bayesian approach Ferguson (1973) placed Dirichlet prior distributions $p \sim D(\alpha)$ on the vector of cell occupation probabilities $p = \{p_i \equiv F(b_i) - F(b_{i-1})\}$, consistently over all possible partitions, by setting $\alpha_i = \alpha(I_i)$ for some finite positive Borel measure $\alpha(\cdot)$ on $\mathbb{R}$; he called this the Dirichlet Process prior, denoted $F \in DP(\alpha)$. Analysis and inference are particularly simple in this approach, because the posterior distribution is again of the same form, now with $F|\mathbf{X}_n \in DP\big(\alpha + \Sigma\delta_{x_i}(\cdot)\big)$. The predictive distribution of X_{n+1} under prior $F \in DP(\alpha)$ is simply the normalized posterior measure α, *i.e.*,

$$X_{n+1}|\mathbf{X}_n \sim \pi(dx|\mathbf{x}_n) = \frac{1}{\alpha(\mathbb{R}) + n}\Big(\alpha(dx) + \sum_{i=1}^{n} \delta_{x_i}(x)\,dx\Big),$$

a weighted average of the prior predictive distribution $\alpha(dx)/\alpha(\mathbb{R})$ and the degenerate empirical pdf $\hat f_n(x)\,dx$; in particular, under this model the probability of ties among the first n observations converges to one as n increases, contradicting the prior assertion that F should have a density function. More alarming, perhaps, is the *negative* prior correlation between each pair of neighboring cell probabilities p_i and p_{i+1}; indeed the model doesn't distinguish between neighboring cells and distant ones, so the joint prior distribution of p_i and p_j depends only on $\alpha(I_i)$, $\alpha(I_j)$, and $\alpha(\mathbb{R})$, and not at all on $|i - j|$ or the distance between I_i and I_j. Dirichlet process priors offer no avenue for expressing prior belief in the continuity, modality, or even existence of a density function $f(x)$. These difficulties can be ameliorated or at least disguised by introducing measurement-error into the model (the so-called Dirichlet Mixture of Normals model; see Escobar 1994), regarding the observations X_i as sums of independent mean-zero normal deviates and the unobserved sample from F, but this offers only very limited opportunity for expressing prior beliefs about the shape and features of the sampling distribution at a cost of considerable computational complexity.

2. The Markov Random Field Prior approach. We propose a Markov random field prior (MRFP) for the vector p of cell probabilities, because many beliefs about shape can be expressed as statements about the joint distribution of neighboring p_i's. We would *like* to specify the prior by giving the complete conditionals $\{[p_i|p_{-i}]\}$ (where p_{-i} denotes the vector of all p_js *except* the i^{th}); unfortunately, $\Sigma p_i \equiv 1$ and so each $[p_i|p_{-i}]$ is the degenerate point mass at $p_i = 1 - \Sigma_{j \neq i} p_j$.

There are many possible schemes for avoiding this degeneracy while using the Markov random field idea. One that we use throughout this paper is to introduce a new parameter $\lambda_+ > 0$, set $\lambda_i \equiv p_i \lambda_+$, and model explicitly prior beliefs about the nonnegative but otherwise unconstrained vector $\lambda = (\lambda_0, \ldots, \lambda_k)$; of course this induces an implicit prior distribution on the derived quantities $p_i \equiv \lambda_i / \lambda_+$. For constant $\beta_i \equiv \beta$ this scheme also arises upon regarding the number n of observations as the observed value of a random variable N with a Poisson prior distribution; the resulting likelihood function for the $\{p_i\}$ is identical to the usual multinomial one.

Ferguson's Dirichlet Process can now be recovered by assigning independent gamma distributions to the $\{\lambda_i\}$, with arbitrary precision (inverse scale) parameter $\beta > 0$ and shape parameters $\alpha_i = \alpha(I_i)$ for a nonnegative measure $\alpha(\cdot)$ on the Borel sets of $\mathbb{R}$,

$$\lambda_i \sim \frac{\beta^{\alpha_i} \lambda_i^{\alpha_i - 1}}{\Gamma(\alpha_i)} e^{-\beta \lambda_i} \, d\lambda_i;$$

this will be a special case of the class of priors we now introduce.

2.1. *Markov Random Fields.* It is more convenient for us to model the probability distribution of the logarithms $\nu_i \equiv \log \lambda_i$ $(0 \leq i \leq k + 1)$, and recover later the distributions of the λ_i's or p_i's. In an unpublished 1971 manuscript Hammersley and Clifford characterized all Markov probability distributions on the set $\{0, \cdots, k + 1\}$; each has the form $\pi(\nu) =$

$ce^{Q(\nu)}$ where

$$Q(\nu) = \sum_{0\le i\le k} \nu_i G_i(\nu_i) + \sum_{0\le i<j\le k} \nu_i\nu_j G_{ij}(\nu_i,\nu_j) +$$
$$+ \sum_{0\le i<j<\ell\le k} \nu_i\nu_j\nu_\ell G_{ij\ell}(\nu_i,\nu_j,\nu_\ell) + \cdots + \nu_0\nu_1\cdots\nu_k G_{01\ldots k}(\nu_0,\nu_1,\ldots,\nu_k)$$

for arbitrary functions $G_{ij\ldots\ell}$ subject to the constraint that $G_S \ne 0$ only if S is a *clique*, *i.e.*, a set of indices each pair (i,j) of which are *neighbors* in the sense that the conditional distribution $[\nu_i|\nu_{-i}]$ depends on ν_j and $[\nu_j|\nu_{-j}]$ on ν_i; see Besag (1974) for a proof and discussion. The simplest nontrivial examples are the "auto" models in which cliques are singletons $\{i\}$ and neighboring pairs $\{i,i+1\}$, so, in one dimension,

$$Q(\nu) = \sum_{i=0}^{k} \nu_i G_i(\nu_i) + \sum_{i=0}^{k-1} \nu_i\nu_{i+1} G_{i\,i+1}(\nu_i,\nu_{i+1}).$$

Continuity. Prior belief in the continuity of f would suggest that the average value of $f(x)$ over the interval I_i, $p_i/(b_i - b_{i-1})$, should be close to the average value over I_{i+1}; upon taking logarithms, the requirement is that

$$\nu_i - \nu_{i+1} \approx \rho_i \equiv \log\frac{b_i - b_{i-1}}{b_{i+1} - b_i}.$$

This can be expressed by including a term $-\frac{1}{2}\sum_i \gamma_i(\nu_i - \nu_{i+1} - \rho_i)^2$ in the log prior, for suitable $\gamma_i \ge 0$. In Hammersley-Clifford form, for any nonnegative $\{\alpha_i,\beta_i,\gamma_i\}_{0\le i\le k}$ satisfying $\gamma_k = 0$, we can set $\gamma_{-1} = 0$ and define $e(x) \equiv (e^x - 1)/x$ (and $e(0) = 1$),

$$\begin{aligned} G_i(\nu_i) &= \alpha_i - \beta_i\, e(\nu_i) - \tfrac{1}{2}(\gamma_{i-1} + \gamma_i)\,\nu_i + \gamma_{i-1}\rho_{i-1} - \gamma_i\rho_i \\ G_{i\,i+1}(\nu_i,\nu_{i+1}) &= \gamma_i \end{aligned}$$

to get

$$\pi(\nu) \propto \exp\left(\sum \alpha_i\nu_i - \sum \beta_i e^{\nu_i} - \frac{1}{2}\sum \gamma_i(\nu_i - \nu_{i+1} - \rho_i)^2\right), \quad \text{or}$$

$$\pi(\lambda) \propto \left(\prod \lambda_i^{\alpha_i - 1} e^{-\beta_i\lambda_i}\right) e^{-\frac{1}{2}\sum \gamma_i\left(\log\frac{\lambda_i}{\lambda_{i+1}} - \rho_i\right)^2}. \tag{2.1}$$

This is similar to classes of prior distributions considered by Lenk (1992) and Hjort (1995). The Dirichlet process prior arises from the case of constant $\beta_i \equiv \beta$ and absent $\gamma_i \equiv 0$, but (2.1) allows the precision parameter β_i to vary from cell to cell and includes a penalty $\frac{\gamma_i}{2}\left(\log\frac{\lambda_i}{\lambda_{i+1}} - \rho_i\right)^2 = \frac{\gamma_i}{2}\left(\log\frac{p_i}{p_{i+1}} - \rho_i\right)^2$ for large proportional changes in probability density between neighboring cells; larger values for γ_i express stronger prior beliefs in the continuity of $f(x)$ at or near b_i.

Monotonicity. A non-linear bi-clique term G_{ij} also allows one to express the prior belief that $f(x)$ should be monotonically increasing or decreasing at a particular point or over some interval, without specifying or even suggesting a functional form or slope. Monotonic increase of the density function at b_i requires only that $p_i/(b_i - b_{i-1}) < p_{i+1}/(b_{i+1} - b_i)$ or, equivalently, that

$$\nu_i - \nu_{i+1} < \rho_i \equiv \log \frac{b_i - b_{i-1}}{b_{i+1} - b_i}$$

(in the equal-width case this simplifies to $\nu_i < \nu_{i+1}$, of course); this can be encouraged or nearly enforced by including in $Q(\nu) = c + \log \pi(\nu)$ a term of the form $-\sum_i e^{\delta_i(\nu_i - \nu_{i+1} - \rho_i)}$ for suitable $\delta_i > 0$. The Hammersley-Clifford form requires terms $G_{ij} = ... + \delta_i^2 e^{-\delta_i \rho_i} e(\delta_i \nu_i) e(-\delta_i \nu_{i+1})$ for $j = i + 1$ and $G_i = ... + \delta_i e^{\delta_i \rho_i} e(-\delta_i \nu_i) - \delta_{i-1} e^{\delta_{i-1} \rho_{i-1}} e(\delta_{i-1} \nu_i)$. Monotone *decrease* would follow from $\delta_i < 0$; monotonicity in either direction can be imposed at a single point b_i (for example, in the presence of the log concavity restriction below, to express conviction that the function is unimodal with a mode to the right of b_i) or over an interval, either finite or not (to express the conviction that the density function is monotonic over that interval).

Log Concavity. Introducing a non-vanishing tri-clique term $G_{ij\ell}$ allows one to express the prior belief that $f(x)$ should be log concave (and, in particular, unimodal) without restricting the location of the mode. Log concavity of the density function at the i^{th} cell requires that the slope of the log density be decreasing,

$$\frac{\log\left(\frac{p_i}{b_i - b_{i-1}}\right) - \log\left(\frac{p_{i-1}}{b_{i-1} - b_{i-2}}\right)}{b_i - b_{i-2}} > \frac{\log\left(\frac{p_{i+1}}{b_{i+1} - b_i}\right) - \log\left(\frac{p_i}{b_i - b_{i-1}}\right)}{b_{i+1} - b_{i-1}},$$

or equivalently that

$$0 < (b_{i+1} - b_{i-1})(\nu_i - \nu_{i-1} + \rho_{i-1}) - (b_i - b_{i-2})(\nu_{i+1} - \nu_i + \rho_i).$$

In the (common) case of equal-width cells the requirement simplifies to $\nu_i > (\nu_{i-1} + \nu_{i+1})/2$; in any case it is a simple restriction on consecutive ν_i's, a lower bound on ν_i given by a convex affine combination $\nu_i > (l_i \nu_{i-1} + m_i + r_i \nu_{i+1})$ of its left and right neighbors (with $l_i + r_i = 1$), which can be encouraged if not enforced by including in the log-density $Q(\nu)$ a term that might take a simple form such as:

$$G_{ij\ell}(\nu_i, \nu_j, \nu_\ell) = \begin{cases} -\epsilon_j \nu_i \nu_j \nu_\ell & \text{if } i = j - 1,\ \ell = j + 1, \text{ and} \\ & \quad \nu_j < (l_j \nu_i + m_j + r_j \nu_\ell); \\ 0 & \text{otherwise} \end{cases}$$

for suitably large ϵ_j (the product $\nu_i\nu_j\nu_\ell$ merely ensures that the term $\nu_i\nu_j\nu_\ell G_{ij\ell}(\nu_i,\nu_j,\nu_\ell)$ will be negative, even for λ_i less than one). Or, more elaborately, we might choose numbers $\epsilon_j \geq 0$ and set $\pi(\nu) \propto e^{Q(\nu)}$, with

$$\begin{aligned}(2.2) \qquad Q(\nu) = c + &\sum \alpha_i\nu_i - \sum \beta_i e^{\nu_i} - \sum \frac{\gamma_i}{2}(\nu_i - \nu_{i+1} - \rho_i)^2 \\ &- \sum \exp\left(\delta_i(\nu_i - \nu_{i+1} - \rho_i)\right) \\ &- \sum \exp\left(\epsilon_i(l_i\nu_{i-1} + m_i - \nu_i + r_i\nu_{i+1})\right),\end{aligned}$$

virtually forcing $\nu_i > (l_i\nu_{i-1} + m_i + r_i\nu_{i+1})$ if ϵ_i is large but otherwise acting exactly like (2.1). Calculating the Hammersley-Clifford terms $G_i(\nu_i)$, $G_{ij}(\nu_i,\nu_j)$, and $G_{ij\ell}(\nu_i,\nu_j,\nu_\ell)$ from (2.2) is tedious but straightforward.

3. Computations. We have described in detail our *prior* distributions for the vector λ of cell rates (and, implicitly, for the vector p of probabilities assigned to the cells I_i by the unknown distribution F); upon observing $X_1 = x_1, \ldots, X_n = x_n$ we approximate the posterior distribution by regarding the "evidence" of $\mathbf{X}_n$ to be only the counts $n_0, \ldots, n_k$ of observations in the $(k + 1)$ cells. Our priors are conjugate for data of this form, leading only to an increase of each α_i by the corresponding $n_i = \sum_{m=1}^{n} \mathbf{1}_{I_i}(x_m)$, and β_i by one. Thus the problems of integrating our prior and posterior distributions are identical.

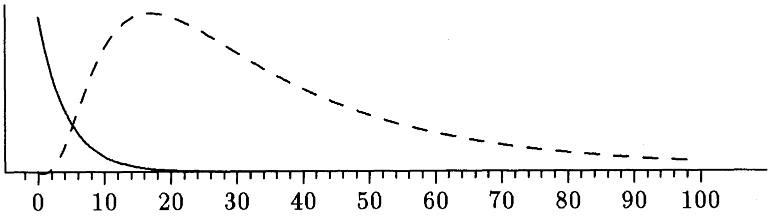
The Gibbs sampling algorithm, in which successive λ_i's (or, equivalently, $\nu_i \equiv \log \lambda_i$) are drawn successively from their complete conditional distributions $[\lambda_i|\lambda_{-i}]$, is simple to implement for our Markov priors and posteriors, because only members of i's cliques appear in ν_i's complete conditionals; for the usual case of bounded cliques, where i and j are never in a common clique if $|i - j| \geq b$ for some integer $b > 0$, there is even a simple parallel implementation, since for each fixed $0 \leq a < b$, all the ν_{a+bi} (for $i \in \mathrm{N}$) have conditionally independent complete conditional distributions and may be sampled simultaneously (on a parallel or vector computer, for example), for successive values of a. Unfortunately the Gibbs scheme's convergence is glacially slow, in our examples, making it virtually useless for problems in which the number of cells is large and the degree of dependence among neighboring cells non-negligible—to move a segment containing many ν_i's takes a very large number of individually very unlikely Gibbs steps.

Fortunately a Metropolis scheme is available. We think of the state as a *function*, not merely a vector of log intensities ν_i, and select proposed Metropolis moves from a mixture distribution intended to take us rapidly around the space of possible intensity functions. One component of our mixture draws randomly two indices i and k, and adjusts the intensity between b_i and b_k (*i.e.*, all the ν_j's for $i \leq j \leq k$) by a randomly-drawn multiplicative factor (with mean one); whole segments of the graph can move up or down in one step. We also include Gibbs draws of randomly-chosen ν_i's from their complete conditionals, and other steps as well. The magnitude of our steps is chosen to maintain a moderate acceptance rate of 20–50% for each of our kinds of moves, using the Hastings/Metropolis acceptance criterion. Convergence remains slow (for our 52-cell example), but adequate.

4. Example. Let $F(\cdot)$ be the distribution of personal income of people at least eighteen years old who were members of the U.S. civilian noninstitutional population in March 1989; we wish to make inferences about and predictions from F, on the basis of sample data from the March 1989 Current Population Survey (reproduced in Freedman *et al.*, 1991).

We use fifty-two cells with $b_0{=}0,\ldots,b_{50}{=}100$ (in thousands of dollars) and construct our prior probabilities $\{p_i\ :\ i\ =\ 0,\ldots,51\}$ for the cells by thinking separately about workers and nonworkers, guessing that about $^1/_3$ of the population would be employed. Our prior guess at the distribution of employed persons' incomes is lognormal $(10.3, .75)$ (with median about $\$30,000$ and quartiles around $\$18,000$ and $\$50,000$; see dashed curve in Figure 2). For nonworkers our guess is an exponential distribution with mean $(3000/\log 2) \approx \$4,300$, making the fraction of incomes that exceed x fall by half for each $\$3,000$ increase in x (solid curve, Figure 2). We also expect about 0.1% of the population to have (otherwise unspecified) negative incomes. These distributions and their mixture are displayed in Figures 2 and 3, respectively (unemployed individuals are represented by the solid line in Figure 2 and the lower portion of each bar in Figure 3); both mixture portions were elicited before considering the data, by fitting conventional distributional forms to hand-drawn curves. Probabilities for the two extremal cells I_0 and I_{51}, each a half-infinite interval, are (somewhat misleadingly) represented in Figure 3 by thin solid bars at $x\ =\ -1$ and $x = 101$, respectively.

Prior Mean Income Distributions by Employment Status



Income (thousands)
Figure 2

We believe that F is fairly smooth, but we don't believe strongly in log concavity, so we choose a prior distribution of form (2.1). With δ_i =

Mixture Prior Mean for Income Distribution, 1980

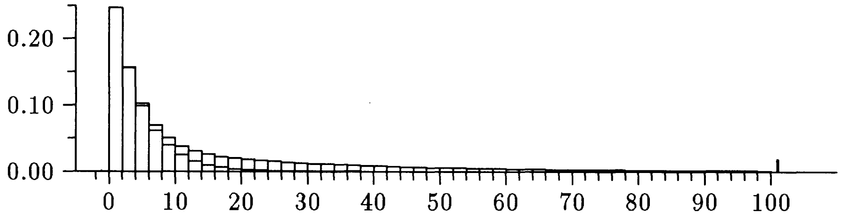


Figure 3

$\epsilon_i = 0$ for all i, we need only assess $\alpha = (\alpha_0, \ldots, \alpha_{51})$, $\beta = (\beta_0, \ldots, \beta_{51})$ and $\gamma = (\gamma_0, \ldots, \gamma_{50})$.

Our approach is to use rough heuristic arguments to find trial values for the α_i, β_i and γ_i; use those trial values in a Metropolis chain to generate draws from our prior; and see whether those draws reflect accurately our prior beliefs. If they don't, we adjust the parameters and try again. We think of (2.1) as the product of a gamma term, expressing beliefs about individual λ_i's, and a lognormal term, expressing beliefs about interactions among neighboring λ_i's.

Let m_i denote the mass of the i^{th} cell according to the mixture described above. To assign a prior probability of $1/2$ to densities within a factor of about 2 of the mixture, we use gamma distributions with shape parameters $\alpha_i = 2$ for $i = 0, \ldots, 51,$. For sample-size N we expect about Nm_i observations to fall in the i^{th} cell, so we choose a prior with mean $\alpha_i/\beta_i = Nm_i$ by setting $\beta_i = \alpha_i/(Nm_i)$.

That makes the mixture distribution equal to the mean of the gamma part of our prior, and implicitly expresses the prior expectation $\mathsf{E}(\lambda_+) = \sum_i \alpha_i/\beta_i$ for the sample-size $N = \sum n_i$. If the $\{\beta_i\}$ are not all constant (as in this example), then λ_+ and $p_i = \lambda_i/\lambda_+$ are *not* independent, so casual misspecification of prior beliefs about λ_+ would (and did, in our early explorations) affect inference about the probabilities $\{p_i\}$. Possible ways of accommodating this include:

(1) Selecting a prior distribution tailored to a specific sample-size, N, as above;

(2) Reparametrizing to $p = \{p_0, \ldots, p_k\}$ so the issue does not arise; and

(3) Selecting a prior distribution as above for an individual observation (*i.e.*, for $N = 1$, setting $\beta_i = \alpha_i/m_i$), and updating it for each successive observation.

For the present example we implemented the third alternative.

For the interaction (*i.e.*, Markov or spatial) part of the prior we set cells 0 and 51 in cliques of their own by setting $\gamma_0 = \gamma_{50} = 0.$; it remains to determine $\gamma_1, \ldots, \gamma_{49}$. For incomes above about $20,000 we held strong beliefs in monotonicity. In fact, we believed with probability 0.9 that the density would decrease monotonically above \$20,000. There are 39 cells in this range. If each cell had a probability of 0.0026 of increase, independently of one another, then the probability of decrease (and hence no mode) beyond \$20,000 would be $0.997^{39} \approx 0.90$.

The complete conditional distribution of ν_i ($\equiv \log \lambda_i$) given $\nu_{-i} \equiv \{\nu_j : j \neq i\}$ depends only on the $j = i \pm 1$ terms. Typically the adjacent γ_i's will be nearly equal, and the ν_i's will be close to their modes $\hat{\nu}_i \equiv \log(N\, m_i) = \alpha_i/\beta_i$; a second-order Taylor expansion of e^x near $\hat{\nu}_i$ gives the approximate conditional log density:

$$\begin{aligned}\log f(\nu_i|\nu_{-i}) &= c_1 + \alpha_i\nu_i - \beta_i e^{\nu_i} - \frac{\gamma_{i-1}}{2}(\nu_i - \nu_{i-1})^2 - \frac{\gamma_i}{2}(\nu_i - \nu_{i+1})^2 \\ &\approx c_2 - \frac{\alpha_i}{2}(\nu_i - \log(Nm_i))^2 - \frac{\gamma_{i-1}}{2}(\nu_i - \nu_{i-1})^2 - \frac{\gamma_i}{2}(\nu_i - \nu_{i+1})^2,\end{aligned}$$

so ν_i has approximately a normal complete conditional $[\nu_i|\nu_{-i}] \sim N(\mu_i, \sigma_i^2)$ with mean and variance

$$\mu_i = \frac{\alpha_i \log(Nm_i) + \gamma_{i-1}\nu_{i-1} + \gamma_i\nu_{i+1}}{\alpha_i + \gamma_{i-1} + \gamma_i} \qquad \sigma_i^2 = \frac{1}{\alpha_i + \gamma_{i-1} + \gamma_i},$$

so (to this approximation)

$$\begin{aligned}\Pr(\nu_i \in (\nu_{i-1}, \nu_{i+1})|\nu_{-i}) &= \Phi\Big(\frac{\alpha_i(\nu_{i-1} - \log(Nm_i)) + \gamma_i(\nu_{i-1} - \nu_{i+1})}{\sqrt{\alpha_i + \gamma_{i-1} + \gamma_i}}\Big) \\ &\quad -\Phi\Big(\frac{\alpha_i(\nu_{i+1} - \log(Nm_i)) + \gamma_{i-1}(\nu_{i+1} - \nu_{i-1})}{\sqrt{\alpha_i + \gamma_{i-1} + \gamma_i}}\Big) \\ &\approx \Phi\Big(\frac{\alpha_i \log(m_{i-1}/m_i) + \gamma_i \log(m_{i-1}/m_{i+1})}{\sqrt{\alpha_i + \gamma_{i-1} + \gamma_i}}\Big) \\ &\quad -\Phi\Big(\frac{-\alpha_i \log(m_i/m_{i+1}) - \gamma_{i-1} \log(m_{i-1}/m_{i+1})}{\sqrt{\alpha_i + \gamma_{i-1} + \gamma_i}}\Big).\end{aligned}$$

We can achieve $\Pr(\nu_i \in (\nu_{i-1}, \nu_{i+1})|\nu_{-i}) \approx (0.90)^{1/39} \approx 0.997$ by arranging that the first argument of $\Phi(z)$ above be $Z_* = \Phi^{-1}(0.997) \approx 2.97$ and the second $-Z_*$; if each α_i is much smaller than $(\gamma_{i-1} + \gamma_i)$, then the requirement is satisfied for

$$\gamma_i \approx \frac{2\, {Z_*}^2}{(\log(m_{i-1}/m_{i+1}))^2} \qquad \text{for } i = 11, \ldots, 49,$$

while for non-negligible but constant $\alpha_j = \alpha$ the computation is tedious but routine. We reasoned similarly for incomes less than \$20,000 but with a weaker belief in monotonicity. The result was

$$\gamma_i \approx \frac{8}{(\log(m_{i-1}/m_{i+1}))^2} \qquad \text{for } i = 1, \ldots, 10.$$

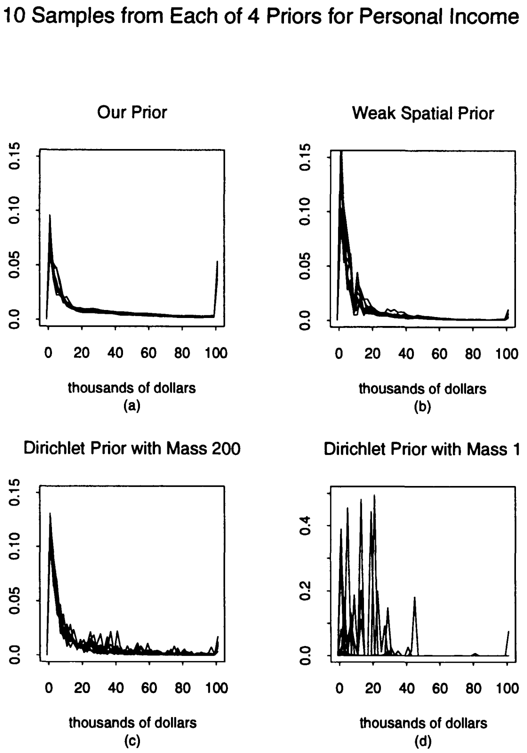


Figure 4

Figure 4 (a) displays 10 draws from our prior. Each draw is represented as a density linearly interpolated between cells. For comparison we also examined three other priors. Figure 4 (b) shows 10 draws from a prior using our original α_i and β_i but in which γ_i has been divided by 20. As expected, draws from this prior are a bit more wiggly than draws from our prior. Figures 4 (c) and (d) show, respectively, draws from Dirichlet priors whose means are equal to the mixture distribution displayed in Figure 3. The prior in Figure 4 (c) has total mass parameter 200; the prior in Figure 4 (d) has total mass 1. As we hoped, realizations from the MRFP's have fewer bumps than realizations from the Dirichlet priors. Note that the vertical scale of Figure 4 (d) had to be enlarged to accommodate large spikes because the Dirichlet prior with mass one expresses a belief that much of the mass is concentrated in only a few points. Note also that the points at $x = -1$ and $x = 101$, for all four priors, are a bit misleading; they represent mass in semi-infinite intervals.

Figure 5 shows the full sample of incomes ($n = 12669$) from Freedman *et al.* (1991) and our subsamples of sizes $n = 50$ and $n = 2000$, from which we compute posteriors. Each histogram is overlayed with a plot of the mixture prior density.

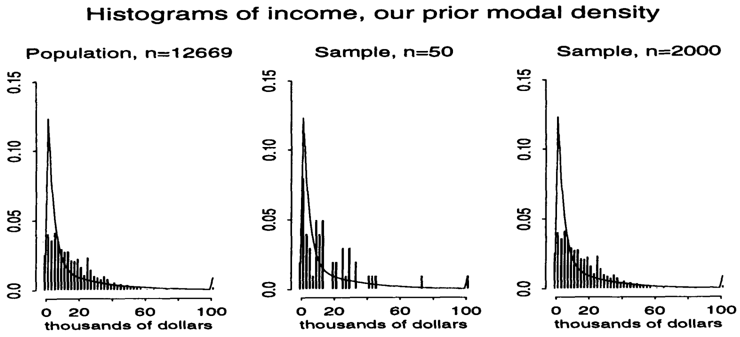


Figure 5

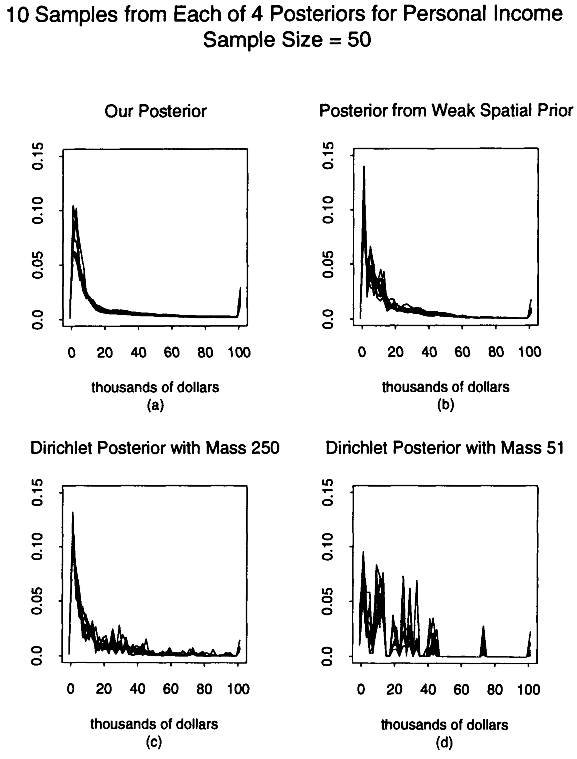


Figure 6

Figure 6 shows draws from the posteriors after a sample of $n = 50$. As hoped, draws from our posterior are smoother than others. The Dirichlet posteriors contain apparently spurious bumps that are especially noticeable at incomes greater than about \$20,000. The Dirichlet posterior with mass 51 is especially egregious. The three posteriors in (b), (c) and (d) show bumps around \$16,000–\$18,000, probably due to the three large histogram bars for those incomes in Figure 5 (b). Draws from our posterior tend not to have modes in that region, but they do diverge, indicating a bit more uncertainty there than elsewhere.

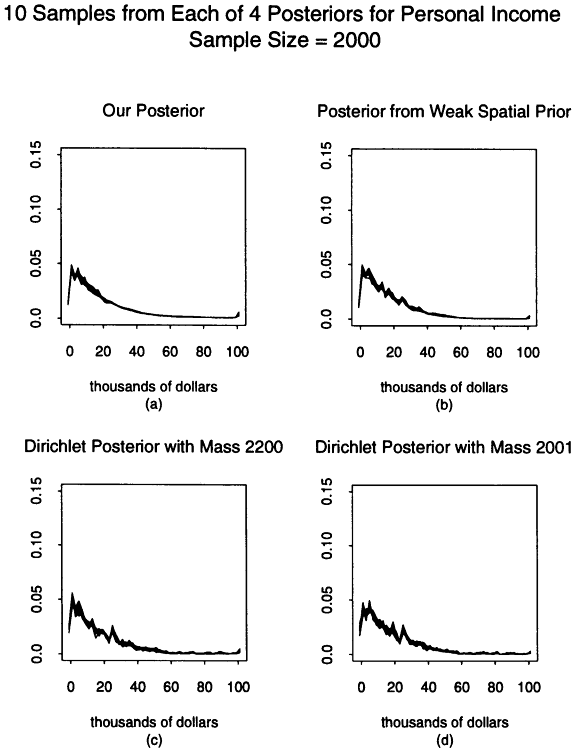


Figure 7

Figure 7 shows draws from the posteriors after a sample of $n = 2000$. Draws from all four posteriors exhibit roughly the same shape. Ours are much smoother, especially above \$20,000 where we expressed a strong prior belief in smoothness. Draws from the other three posteriors all have modes at \$25,000, as does the histogram of the sample, Figure 5 (c); our posterior does not have these modes. Our posterior reflects our beliefs about *incomes*; but the data are *reported incomes.* In retrospect, we made the beginner's mistake of expressing too much certainty in our beliefs—this time, beliefs about smoothness.

5. Discussion. We believe this class of models offers a flexible and easy-to-use method of generating plausible prior distributions in difficult problems. It is far more flexible than Dirichlet process methods or logistic-normal methods, both of which it subsumes, by offering a wide range of local expressions of behavior (via the Hammersley-Clifford $G_{i \dots \ell}$ functions) that can be tailored to common expressions of prior belief (in continuity, monotonicity, log concavity, *etc.*) without forcing any particular form on the unknown distribution. It also subsumes earlier more specific proposals including nearest-neighbor models; Markov-chain, random-walk, and auto-regressive process models; and logistic normal models. The methods generalize easily to higher dimension (where our intervals are replaced by squares, triangles, or hexagons in two dimensions, for example, and appropriate polyhedra in higher dimensions, with corresponding cliques) or to more complex settings (manifolds, graphs, *etc.*)

Even one-dimensional density estimation leads to high-dimensional statistical models, and in particular to high-dimensional prior distributions, and so requires a great deal of experimentation and "exploratory prior analysis." Repeatedly we refined our selection of a representation intended to express our prior beliefs; drew a series of samples from the *prior* and from posteriors with synthetic data; and plotted various features of the results, in an effort to arrive at a representation that *does* express those beliefs, before we began analysis of the actual data set. We believe this approach may be fruitful in other high-dimensional problems in Bayesian analysis.

MRFPs are typically used for modeling regression functions, where an underlying spatial field is observed with error. For example, in image reconstruction the underlying field emits photons; the data are the numbers of photons captured by an array of detectors, which is approximately the intensity of the field plus error. Our application is different in two ways. First, we don't observe the field plus error; we observe random draws from a density function which we model as a spatial field. Therefore, our data are informative about our field in a way that is fundamentally different from regression data. And second, our field is constrained to integrate to unity. Hence our complete conditionals are degenerate and we can't use the usual Gibbs samplers for MRFPs. We do feel that our computational method, in which a Markov chain is constructed with Hastings/Metropolis move proposals that rescale entire segments of the density function, would be useful in a variety of nonparametric Bayesian analyses (regression, binary regression, hazard analysis), and not only in density estimation.

A criticism that might be addressed to our models is that they require a choice of "cells" or "bins" that is arbitrary, subjective, and can affect inference; it is not alway possible to subdivide or aggregate cells and remain within this class of prior distributions, for example, so the expressions of prior belief in unimodality, continuity, log concavity, *etc.* that they offer may only be meaningful within the context of a particular discretization. The Dirichlet process approach does not suffer from this criticism.

Our method *permits* one to specify prior beliefs in a way that is independent of the discretization (*i.e.*, to specify a consistent way of assigning prior distributions to the discrete approximations in all possible partitions of $\mathbb{R} = \cup(b_i, b_{i+1}])$, but it does not it restrict one to such specifications; it is outside the scope of the present work to offer guidance on how consistent families of Markov priors might be elicited and specified.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Society (Ser. B)* **2**, 192–225.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. American Statist. Assoc.* **89**, 268–277.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

Freedman, D., R. Pisani, R. Purves, & A. Adhikari (1991). Data diskette accompanying instructor's edition of *Statistics, 2nd edn.* Norton Press, New York.

Gilks, W.R. & P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41**, 337–348.

Gilks, W.R., N.G. Best, & K.K.C. Tan (1994). Adaptive rejection Metropolis sampling within Gibbs sampling. *Technical report, Cambridge University.*

Hammersley, J.M., & P. Clifford (1971). Markov fields on finite graphs and lattices. *(unpublished).*

Hjort, N.L. (1995). *Bayesian approaches to non- and semiparametric density estimation.* Bayesian Statistics 5, Oxford University Press.

Lenk, P.J. (1992). A practicable Bayesian nonparametric density estimator. *Technical report, School of Business Administration, The University of Michigan.*

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Markov Random Field Priors for Univariate Density Estimation

discussion by
A.P. Dawid
University College London

It has long been appreciated that the siren appeal of the conjugate Dirichlet Process analysis of nonparametric problems needs to be resisted if we wish to make inferences we can believe in. Attempts at introducing more general prior dependence structures were made by Tom Leonard in the early 70s (see, for example, Leonard (1973) and, for a review, Leonard and Hsu (1994)), but only now do we have the necessary computational tools and power to do it properly. This paper is a welcome contribution to advancing the subject.

The introduction of Poisson parameters $\lambda = (\lambda_i)$, with $p_i = \lambda_i/\lambda_+$, goes back to Fisher, but it is cleverly used here for a new purpose: to break the constraint $\Sigma p_i = 1$, and thus allow non-degenerate MCMC sampling. Lindley (1964), working with Dirichlet distributions, pointed out that, since $\Sigma a_i \log p_i = \Sigma a_i \nu_i$ ($\nu_i = \log \lambda_i$) if $\Sigma a_i = 0$, the distributions of such "log-constrasts" are readily determined from the distribution of ν, and fully determine that of p. This trick might be useful here also.

The full power of the Hammersley-Clifford theorem is not needed for the authors' application. To ensure the Markov property, it is enough to require that the joint density factorise (not, in general, uniquely) as

$$\pi(\nu) = \prod_i \phi_i(\nu_i) \times \prod_{i,j} \phi_{ij}(\nu_i, \nu_j) \times \dots,$$

including only terms which are cliques, and with no need for additional constraints, such as $\phi_{ij}(0, \nu_j) \equiv 1$, as implied by the authors' development. In particular, by allowing 0 values for the ϕ's, it becomes possible to incorporate exactly such "hard constraints" as $\pi = 0$ whenever $\nu_i < l_i \nu_{i-1} + m_i + r_i \nu_{i+1}$.

The authors' replacement of the multinomial model by the Poisson needs to be handled with more care. In terms of $\lambda_+ = \Sigma \lambda_i$ and $p_i = \lambda_i/\lambda_+$, the full Poisson likelihood is

$$\lambda_+^{n_+} e^{-\lambda_+} \times \prod p_i^{n_i}.$$

For multinomial sampling, the former term is absent. Suppose that we use a prior for λ, which we may express as $\pi(p)\pi(\lambda_+|p)$. Then the marginal posterior for p, using the Poisson likelihood, is

$$\pi(p) \times (\Pi p_i^{n_i}) \times \int \lambda_+^{n_+} e^{\lambda_+} \pi(\lambda_+|p) d\lambda_+.$$

This only agrees with the desired multinomial answer if the integral term is independent of p, which requires prior independence of p and λ_+. This holds for the Gamma structure underlying the Dirichlet, but not for the priors introduced in the paper. Hence the analysis is in need of correction.

For example, the correct posterior for λ, based on prior (2.1) and the true multinomial likelihood $\Pi\lambda_i^{n_i}/\lambda_+^{N_+}$, is

$$\pi^*(\lambda) \propto (\Pi\lambda_i^{\alpha_i+n_i-1})e^{-\Sigma\beta_i\lambda_i}e^{-\frac{1}{2}\Sigma\gamma_i(\nu_i-\nu_{i+1}-\rho_i)^2}/\lambda_+^{n_+}.$$

The complete conditional density of λ_i, given the remaining λ's, is thus

$$\pi^*(\lambda_i \mid \lambda_{\setminus i}) \propto \lambda_i^{\alpha_i+n_i-1}e^{-\beta_i\lambda_i}e^{-\frac{1}{2}\{\gamma_{i-1}(\nu_i-\nu_{i-1}-\rho_{i-1})^2+\gamma_i(\nu_i-\nu_{i+1}-\rho_i)^2\}}/(\lambda_i+\sum_{j\neq i}\lambda_j)^{n_+},$$

depending on $\sum_{j\neq i}\lambda_j$, as well as on λ_{i-1} and λ_{i+1}. It should not be difficult to incorporate this correction into a modified MCMC scheme.

I like the authors' interactive approach to specifying prior hyper-parameters, although I was a little puzzled by the details. The expectations of the λ's are not found from the "gamma part" of the prior alone, and I do not see why the mean of the entirely fictitious auxiliary parameter λ_+ need be related to N (nor do I understand the comment about taking $N = 1$). There is no need for the full distribution of λ to match any empirical reality: it is enough that the induced prior for p should appear acceptable.

Finally, an extension of the authors' approach might be useful when the data arrive coarsely grouped, but we wish to model the probabilities for a finer grouping. Again we introduce λ's (for the finer grouping). Then the posterior complete conditional for parameter λ_{ij}, corresponding the the jth fine grouping in coarse grouping i, is proportional to

$$\pi(\lambda_{ij} \mid \lambda_{\setminus(ij)}) \times (\lambda_{ij} + \sum_{s\neq j}\lambda_{is})^{n_i}/(\lambda_{ij} + \sum_{(rs)\neq(ij)}\lambda_{rs})^{n_+}.$$

This thus depends on the prior "neighbours" of λ_{ij}, on $\sum_{s\neq j}\lambda_{is}$, and on $\sum_{(rs)\neq(ij)}\lambda_{rs}$.

REFERENCES

LEONARD, T. (1973). A Bayesian method for histograms. *Biometrika* **60** 297–308.

LEONARD, T. and HSU, J. S. J. (1994). Bayesian analysis of categorical data. In *Aspects of Uncertainty: A Tribute to D. V. Lindley* (P. R. Freeman and A. F. M. Smith, eds.) 283-310. Wiley.

LINDLEY, D. V. (1964). The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35** 1622–1643.

Rejoinder

Robert L. Wolpert and Michael Lavine

We would like to thank Professor Dawid for his insight and suggestions, and especially for correcting our inadvertent omission of reference to the pioneering work of Tom Leonard in this area.

While we agree that the Hammersley-Clifford theorem is overkill for our application, we expressed our prior distributions in this form to emphasize their Markovian nature; we would like to recommend the wider use of Markovian prior distributions in nonparametric Bayesian analysis, and the Hammersley-Clifford theorem gives them a complete characterization. This is also our motivation in employing the well-known technique of embedding our multinomial problem in a Poisson model (*i.e.*, in introducing λ_+ and setting $\lambda_i \equiv p_i\lambda_+$). This representation reveals the Dirichlet process prior as that induced by the *independent-increment* gamma process prior for the Poisson intensities λ, while low-dimensional parametric models lead to *complete dependence* among the Poisson intensities. We recommend the intermediate step of Markovian distributions, able to reflect expressions of strong belief about the relationship in adjacent or nearby cells (unlike the Dirichlet process), without inducing long-range dependence (as do parametric models). As Professor Dawid suggests, our computational techniques would allow us to avoid the Poisson embedding and, with it, the sticky issue of expressing prior belief about a fictitious parameter; probably this would be wise for routine use of these models.

Our assertion that, under the Poisson embedding, "the resulting likelihood function for the $\{p_i\}$ is identical to the usual multinomial one," is of course only correct if the parameters λ_+ and p are independent under the prior distribution—which Professor Dawid rightly questions in our case, apparently fearing that our "Markov" terms (those including coefficients γ_i, δ_i, or ϵ_i in Equation (2.2)) will induce dependence. In fact these terms are benign (each depends only on a log contrast, hence only upon p and not λ_+ in the (p, λ_+) parameterization), but a problem appears already with the gamma (α_i, β_i) part of the prior unless the $\{\beta_i\}$ are constant. If each $\beta_i = \beta$, then in the (λ_+, p) parameterization our prior becomes

$$\begin{aligned}\pi(d\lambda_+\, dp) \propto {\lambda_+}^{\alpha_+ - 1} e^{-\beta\lambda_+}\, d\lambda_+ \\ \times \prod_{i=0}^{k} \left[{p_i}^{\alpha_i - 1}\right] \prod_{i=0}^{k-1} \left[e^{-\frac{\gamma_i}{2}(\log \frac{p_i}{p_{i+1}} - \rho_i)^2}\, e^{-\exp\left(\delta_i(\log \frac{p_i}{p_{i+1}} - \rho_i)\right)}\right] \\ \times \prod_{i=1}^{k-1} \left[e^{-\exp\left(\epsilon_i(l_i \log \frac{p_{i-1}}{p_i} + r_i \log \frac{p_{i+1}}{p_i} + m_i)\right)}\right] d^k p,\end{aligned}$$

with λ_+ and p independent; in this case (as shown in the discussion) the posterior distributions for p and λ_+ are independent, and the Poisson embedding is harmless for either the Poisson or multinomial likelihood.

If however the β_i are *not* constant, then a problem arises even in the absence of our Markovian terms (*i.e.*, even if all the γ_i's, δ_i's, and ϵ_i's vanish)—for inhomogeneous gamma distributions, the joint and marginal posteriors are then

$$\pi^*(d\lambda_+\, dp\,|\mathbf{N} = \mathbf{n}) \propto$$
$$\lambda_+{}^{(\alpha_+ + n_+ - 1)} e^{-\lambda_+ \left(1+\sum_{i=0}^{k} \beta_i\, p_i\right)} \prod_{i=0}^{k} \left[p_i{}^{\alpha_i + n_i - 1}\right] d\lambda_+\, d^k p$$
$$\pi^*(dp\,|\mathbf{N} = \mathbf{n}) \propto$$
$$\left(1 + \sum_{i=0}^{k} \beta_i\, p_i\right)^{-(\alpha_+ + n_+)} \prod_{i=0}^{k} \left[p_i{}^{\alpha_i + n_i - 1}\right] d^k p,$$

unequal to the usual (conjugate) Dirichlet distribution unless the β_i's are constant. The multinomial likelihood (instead of the Poisson, featured here) leads to the same result.

While the problem can be avoided by setting $\beta_i \equiv 1$, there are certainly situations in which it would be preferable to allow unequal β_i's, expressing stronger *a priori* opinions about some λ_i's than others. If the data are censored, for example, including only the counts n_j for certain cells $j \in J \subset \{0, ..., k\}$, then under a Poisson-model posterior, each β_j increases by one while the β_i for $i \notin J$ do not, compelling us to accommodate unequal β_i's. The multinomial model loses its conjugacy and simplicity in this case, leading to mixture distributions for censored data, while the Poisson version remains tractable.