BIOMETRICS , 1–10
December 2015

Robustifying Bayesian nonparametric mixtures for count data

Antonio Canale* **and Igor Prünster****
Department of Economics and Statistics and Collegio Carlo Alberto, University of Torino, Torino, Italy
**email:* antonio.canale@unito.it
***email:* igor.pruenster@unito.it

SUMMARY: Our motivating application stems from surveys of natural populations and is characterized by large spatial heterogeneity in the counts, which makes parametric approaches to modeling local animal abundance too restrictive. We adopt a Bayesian nonparametric approach based on hierarchical mixtures and innovate with respect to popular Dirichlet process mixture of Poisson kernels by increasing the model flexibility at the level both of the kernel and the nonparametric mixing measure. This allows to derive accurate and robust estimates of the distribution of local animal abundance and of the corresponding clusters. The application and an extensive simulation study yield also some general methodological implications. Adding flexibility solely at the level of the mixing measure does not improve inferences since its impact is severely limited by the rigidity of the Poisson kernel with considerable consequences in terms of bias. However, once a kernel more flexible than the Poisson is chosen, inferences can be robustified by choosing a prior more general than the Dirichlet process. Therefore to improve the performance of Bayesian nonparametric mixtures for count data one has to enrich the model simultaneously at both levels, the kernel and the mixing measure.

KEY WORDS: Abundance heterogeneity; Bayesian Nonparametrics; Hierarchical mixture model; Pitman–Yor process; Poisson mixture; Rounded Mixture of Gaussians.

1. Introduction

The Dirichlet process (DP) mixture model, introduced by Lo (1984), currently represents the most popular Bayesian nonparametric model and is widely used for density estimation, clustering, and as key nonparametric ingredient in complex models. See Müller et al. (2015); Hjort et al. (2010) for exhaustive accounts. A recent line of research has explored the possibility of replacing, within hierarchical mixture models, the DP with more general classes of nonparametric priors. It turns out that a more general nonparametric prior can lead to more accurate estimates, especially in terms of the quantification of the mixture components. See, for instance, Ishwaran and James (2001); Lijoi et al. (2005, 2007) and the recent reviews in Barrios et al. (2013); De Blasi et al. (2015). However, up to now all studies have been confined to the case of mixture models for continuous data. Although the case of count data, or discrete data in general, is also important, little is known on the performance of general nonparametric mixtures for their estimation. Here we fill this gap by considering discrete mixtures based on the Pitman–Yor process (Pitman and Yor, 1997), which includes the DP as a special case, and aim at verifying whether the added flexibility is beneficial also in the discrete case.

Our motivating application stems from surveys of natural populations and is characterized by large spatial heterogeneity in the counts, a direct consequence of difference in animal abundance among sample locations. In particular, we focus on a specific dataset consisting of counts of an endangered fish species first analyzed in Dorazio et al. (2008), to be described together with the sampling protocol in Section 2. In their paper Dorazio et al. (2008) nicely show that the data heterogeneity requires a nonparametric approach, which is clearly superior to parametric models. As Bayesian nonparametric model they adopt a DP mixture of Poisson kernels, a natural choice in presence of count data. Poisson parametric and nonparametric mixtures, indeed, played a central role in extending the Poisson distribution for complex situations for their mathematical tractability. See, for instance, Hougaard et al. (1997), Viallefont et al. (2002), Karlis and Xekalaki (2005), Guindani et al. (2006, 2014), Brown and Buckley (2015) and Li et al. (2015). One of the key aspects of the results pointed out in Dorazio et al. (2008) is that the estimation of the mixture components is a difficult task in this context. This observation represents the starting point of our analysis aiming at improving estimation by replacing the DP with a more general nonparametric prior. However, we discover that this is not sufficient to stabilize and improve the estimates of the number of mixture components. In contrast, the difficulty of this estimation problem is even more apparent with a general nonparametric prior. This leads to conjecture that the origin of the problem is actually represented by the Poisson kernel and our findings will confirm it. It is well–known that the standard parametric Poisson model cannot accommodate under- and over-dispersion. However, this lack of flexibility carries over, to a certain extent, to Poisson mixtures regardless of how general the chosen mixing measure is. In fact, the mean-variance structure of Poisson mixtures is still rigid and it is also easy to show that even infinite Poisson mixtures do not contain under-dispersed distributions in their support. Therefore, in order to appropriately tackle the application at hand we also consider kernels more flexible than the Poisson and, in particular, the Rounded Gaussian kernel recently introduced in Canale and Dunson (2011). As will be shown, adding flexibility to both kernel and mixing measure leads then to the envisaged more accurate and robust results.

This paper has been submitted for consideration for publication in *Biometrics*

Given these findings for the population count application, it is then natural to investigate the general validity of the discovered phenomena. This is done by an extensive simulation study. The conveyed evidence is unequivocal in suggesting: (i) to use Poisson mixtures with caution given their lack of flexibility leading to a potentially poor fit in terms of estimation of both the probability mass function and the number of mixture components; (ii) a flexible and at the same time robust mixture model can be achieved by acting at both levels, the kernel and the mixing measure, and modeling process mixtures with Rounded Gaussian kernel appear to be an effective and computationally convenient choice.

Section 2 first describes the dataset to be analyzed together with the sampling protocol, then presents our nonparametric model and computational strategy and finally discusses the results. Section 3 is devoted to a simulation study in which several different scenarios of data generating distribution are considered and the performance of different nonparametric mixtures is compared. Section 4 contains some concluding remarks. The Appendix provides additional details on the prior centering and on the sampling scheme. The Supplementary Materials (SM) display further complementary illustrations concerning both the application and the simulation study.

2. Animal abundance estimation

Surveys of animal populations represent a natural source of count data. See Royle and Dorazio (2008) for a recent review. In an important paper by Dorazio et al. (2008) the problem of modeling heterogeneity in abundance of stream fishes among different sampling locations was considered and a dataset consisting of counts of Okaloosa darters (*Etheostoma okaloosae*) in $n = 53$ different locations of a stream in northwest Florida was introduced and analyzed. Here we consider the same dataset as motivating application.

In particular, data were collected with a specific sampling protocol called "removal sampling" (Dorazio et al., 2005). In most animal sampling protocols the actual number of animals Y_i at sites $i = 1, \ldots, n$ cannot be directly observed and is to be estimated. Under the removal sampling protocol Y_i and the capture probability π_i are both estimated using data on successive removal passes. The observed data, for the i-th site, consist in $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ_i})'$, a vector containing the number of animals observed in J_i successive removal passes. The observed counts $\mathbf{z_i}$ are modeled as multinomial outcomes with parameters $(Y_i, \boldsymbol{\pi}_i)$ with $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ_i})'$ and $\pi_{ij} = \pi_i(1-\pi_i)^{j-1}$ for $j = 1, \ldots, J_i$, i.e. their probability mass function (pmf) is

$$\Pr(\mathbf{z}_i | Y_i, \pi_i) = \frac{Y_i!}{(Y_i - z_i)! \prod_{j=1}^{J_i} z_{ij}!} \left(\prod_{j=1}^{J_i} \pi_{ij}^{z_{ij}} \right) \left(1 - \sum_{j=1}^{J_i} \pi_{ij} \right)^{Y_i - z_i} \tag{1}$$

with $z_i = \sum_{j=1}^{J_i} z_{ij}$.

In the considered dataset, the number of fishes observed in the first pass has mean 40.34 and standard deviation 39.48 suggesting substantial heterogeneity in local abundances. Also, the total number of removal passes varies from site to site and ranges between one and three. In those sites where multiple passes were taken, lower removal counts were registered in successive passes suggesting effectiveness of the sampling protocol in depleting the local populations of animals.

As in Dorazio et al. (2008) we assume independent priors for π_i while the site-specific abundances y_i are modeled via a nonparametric mixture. Dorazio et al. (2008) clearly show that the flexibility conveyed by a nonparametric approach is necessary in this context overcoming drawbacks inherent to a parametric modeling. We innovate on their approach in two dimensions. First, we consider mixing measures more general than the DP, namely the Pitman–Yor process, to further improve the flexibility. Second, by considering kernels more flexible than the Poisson. The results both for this dataset and for data generated in an extensive simulation study, reported in Section 3, show the benefit of our proposed innovations and have interesting general methodological implications.

2.1 *Model and prior specification*

As far as the nonparametric component is concerned, we propose to use a Pitman–Yor (PY) process, which represents probably most tractable generalization of the DP. Such a nonparametric prior has already found many successful applications in various areas including imagine reconstruction, linguistics, networks, and species sampling, among others. See, e.g., Hjort et al. (2010) and De Blasi et al. (2015) for recent accounts. Like for the DP, any sample $X_1, \ldots, X_n$ drawn from a PY process will feature ties with positive probability, therefore generating $K_n \leqslant n$ distinct observations $X_1^*, \ldots, X_{K_n}^*$ with frequencies $n_1, \ldots, n_{K_n}$ such that $\sum_{i=1}^{K_n} n_i = n$. The PY can be defined in terms of its predictive distributions, which take on a particularly simple form and uniquely characterize it. Let (σ, θ) be parameters such that $\sigma \in [0,1]$ and $\theta > -\sigma$ and P_0 be a probability distribution on $\mathbb{X}$. The associated predictive distributions are then of the form

$$\Pr(X_{n+1} \in \cdot \,| X_1, \ldots, X_n) = \frac{\theta + \sigma K_n}{\theta + n} P_0(\cdot) + \frac{1}{\theta + n} \sum_{j=1}^{K_n} (n_j - \sigma) \delta_{X_j^*}(\cdot) \tag{2}$$

with δ_a indicating a point mass at a. In symbols a PY process will be denoted by $\text{PY}(\theta, \sigma; P_0)$. For $\sigma = 0$, the predictive distributions (2) clearly reduce to the well–known DP case. Note that (2) represent a key ingredient of the sampling scheme detailed in the next section.

Given the nonparametric prior to be used, we propose to model the abundance distribution p via PY mixture priors, i.e.

$$p(\cdot) = \int k(\cdot; x) \tilde{P}(dx), \qquad \tilde{P} \sim \text{PY}(\theta, \sigma; P_0). \tag{3}$$

As for the kernel k, we compare the results of two different choices. Following Dorazio et al. (2008), the first specification corresponds to k being a Poisson kernel with mean parameter $\lambda = \exp(\phi)$ and we also set the base measure P_0 to be normal with mean α and variance ω^2. Hence, the hierarchical representation of the model is

$$\begin{aligned} Y_i \mid \phi_i &\sim \text{Poi}(\exp(\phi_i)), \\ \phi_i \mid \tilde{P} &\sim \tilde{P} \\ \tilde{P} \mid \sigma, \theta, N(\alpha, \omega^2) &\sim \text{PY}(\sigma, \theta; N(\alpha, \omega^2)). \end{aligned} \tag{4}$$

The second specification relies on a flexible rounded Gaussian (RG) kernel. The general idea is that a discrete kernel k_r can be obtained by thresholding the domain of a continuous kernel k via a prespecified sequence a_j such that $k_r(j; x) = \int_{a_j}^{a_{j+1}} k(y^*; x) dy^*$. For instance, if $k(\cdot, x)$ is defined on $\mathbb{R}^+$, one can set $a_j = j$ for $j = 0, 1, 2, \ldots$, whereas if $k(\cdot; x)$ is defined on $[0,1]$, one can set $a_j = 0, 1/2, \ldots, 1 - 1/2^h, \ldots$. Henceforth we consider the

following RG mixture

$$\begin{aligned} Y_i \mid \phi_i &\sim \text{RG}(\mu_i, \tau_i^{-1}), \\ (\mu_i, \tau_i) \mid \tilde{P} &\sim \tilde{P} \\ \tilde{P} \mid \sigma, \theta, P_0 &\sim \text{PY}(\sigma, \theta; P_0). \end{aligned} \tag{5}$$

where $\text{RG}(\cdot; \mu, \tau^{-1})$ denotes a RG kernel with location μ and precision τ and thresholds $a_0 = -\infty$, $a_j = j$ for $j = 1, 2, \ldots, \infty$, i.e.

$$\text{RG}(j; \mu, \tau^{-1}) = \int_{a_j}^{a_{j+1}} N(y^*; \mu, \tau^{-1}) dy^*.$$

For the base measure we adopt standard choices (Escobar and West, 1995) assuming $P_0(\mu, \tau) = N(\mu; \mu_0, \kappa\tau^{-1})\text{Ga}(\tau; \alpha, \beta)$ and a hyperprior on the rate parameter β. Adopting a default empirical Bayes approach, the scale parameter κ is fixed equal to the variance of the observed counts in the first removal pass and the location parameter μ_0 is set equal to $\sum_{j=1}^{3} \bar{z}_j$, where $\bar{z}_j$ is the sample mean of the j-th removal pass, calculated for the locations having at least j removal passes. Typically the location parameter is centered on the sample mean of the observed data, which corresponds to computing the sample mean of the y_i's in our sampling protocol. Since these are not observed, we use the mean of the sum of each removal count, accounting for the fact that different numbers of removal counts are considered for different locations, as a proxy.

As for the parameters (θ, σ) of the PY process, we take different values of σ and fix θ in a way to make the corresponding PY priors comparable. Specifically we consider $\sigma = 0, 0.25, 0.5, 0.75$ and fix θ such that the prior expected number of distinct mixture components, $\mathbb{E}[K_n]$, is equal to a desired value. In this way all PY priors are centered, a priori, on the same number of clusters. In our case we consider prior centerings on 10, 22, 30 and 40 components and the corresponding pairs of (θ, σ) are reported in Table 1. This is achieved in a straightforward way by using the well known expressions for $\mathbb{E}[K_n]$ in PY case (reported in the Appendix A.1).

Table 1

PY prior centering for the $n = 53$ Okaloosa darters dataset: values of θ corresponding to various choices of σ such that the prior expected number of components is equal to a desired number.

$\mathbb{E}[K_n]$	$\sigma = 0$	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$
10	3.38	1.60	0.21	-0.60
22	13.59	8.24	3.46	0.07
30	27.82	18.24	9.20	1.72
40	72.55	50.95	29.75	9.83

In both cases we assume π_i to be fixed for each location and variability in detectability among sites is modeled with independent beta priors $\pi_i \sim \text{Be}(a, b)$ with a and b equal to the posterior means obtained by Dorazio et al. (2008).

2.2 *Posterior computation*

Posterior samples from the models discussed in Section 2.1 are obtained by using Markov chain Monte Carlo (MCMC) algorithms. For the nonparametric Poisson mixtures, the algorithm detailed in the Supplementary Materials of Dorazio et al. (2008) has been used with the appropriate modifications to extend it to PY processes. Instead, for model (5), our algorithm is obtained by suitably adapting the one set forth in Canale and Dunson (2011). According to their proposal, a first data augmentation step is required to simulate latent continuous Y_i^*'s. Then, conditionally on the Y_i^*'s, the algorithm relies on any existing MCMC algorithm developed for nonparametric mixtures of Gaussians. However, in this particular application also the Y_i's are unobserved and need to be estimated from the observed removal counts. In Dorazio et al. (2008) the full conditionals of the Y_i's have a simple Poisson specification and thus the simulation of the Y_i's can be done easily. In contrast, for the RG case the conditional posterior of Y_i is not in closed form and a Metropolis-Hastings step needs to be introduced to simulate y_i. However, we are able to mitigate this issue by merging the steps to simulate Y_i and Y_i^* in a single step, directly simulating Y_i^* from its full conditional posterior distribution via Metropolis-Hastings. Details are reported in Appendix A.2.

Conditionally on Y_i^*, each observation is assigned to a cluster S_i with $S_i = 1, \ldots, K_n$ with $K_n \leqslant n$. The posterior cluster allocation is performed via a generalized Pólya-Urn sampler based on the predictive distributions (2). In particular, the modification of Algorithm 8 of Neal (2000) reported in the Appendix is employed. A further reshuffling step that, conditionally on such cluster allocations, draws new values for the kernel's parameters is also performed following Bush and MacEachern (1996).

Finally, the conditional posterior distribution of π_i and the probability of animal detection at site i in a single removal have the same simple closed form as in Dorazio et al. (2008).

2.3 *Results and discussion*

In discussing the results we focus on the two key quantities of inferential interest, namely the estimation of the pmf of the local abundance, which given the heterogeneity in population distribution can be thought of as mixture, and the number of components such a pmf is made of.

First we focus on the estimation of the pmf of local abundance, which according to the adopted Bayesian nonparametric approach, is obtained as posterior expected value of (3) or, in other terms, as the predictive distribution. Figure 1 displays the corresponding estimates for the Poisson and RG mixture models. In terms of prior specification, the plots correspond to the intermediate case of $\sigma = 0.25$ and prior expected number of components $\mathbb{E}[K_n] = 22$. It is important to remark that there is no significant difference in the pmf estimates for each model when varying σ or the prior centering of $\mathbb{E}[K_n]$. The pmf estimates corresponding to the Poisson mixture essentially coincides with one obtained in Dorazio et al. (2008) for the special case of a Dirichlet process, i.e. a PY process with $\sigma = 0$. However, only by comparing the Poisson mixture to the RG mixture one realizes that the Poisson mixture is quite inflexible and tends to over-smooth under-dispersed mixture components, especially away from 0. In fact, considering Poisson mixtures is a natural and at first glance highly flexible choice. However, the well-know rigidity of the Poisson kernel (due to a single parameter controlling location and scale) carries over to the mixture model even when the mixing measure is a highly flexible nonparametric prior such as the PY process. In contrast, as apparent from Figure 1, the RG model is able to capture over- and under-dispersed components. Another appealing aspect of the RG model, if compared to the Poisson model, is the ability to naturally detect zero-inflated pmf. Indeed, the estimated mass in zero for the RG mixture is 0.088 while for the Poisson mixture it is just 0.043. Note that the proportion of zero counts in the sample is 0.094.

Things become even more interesting when looking at the second key aspect, the posterior distribution of the number of mixture com-

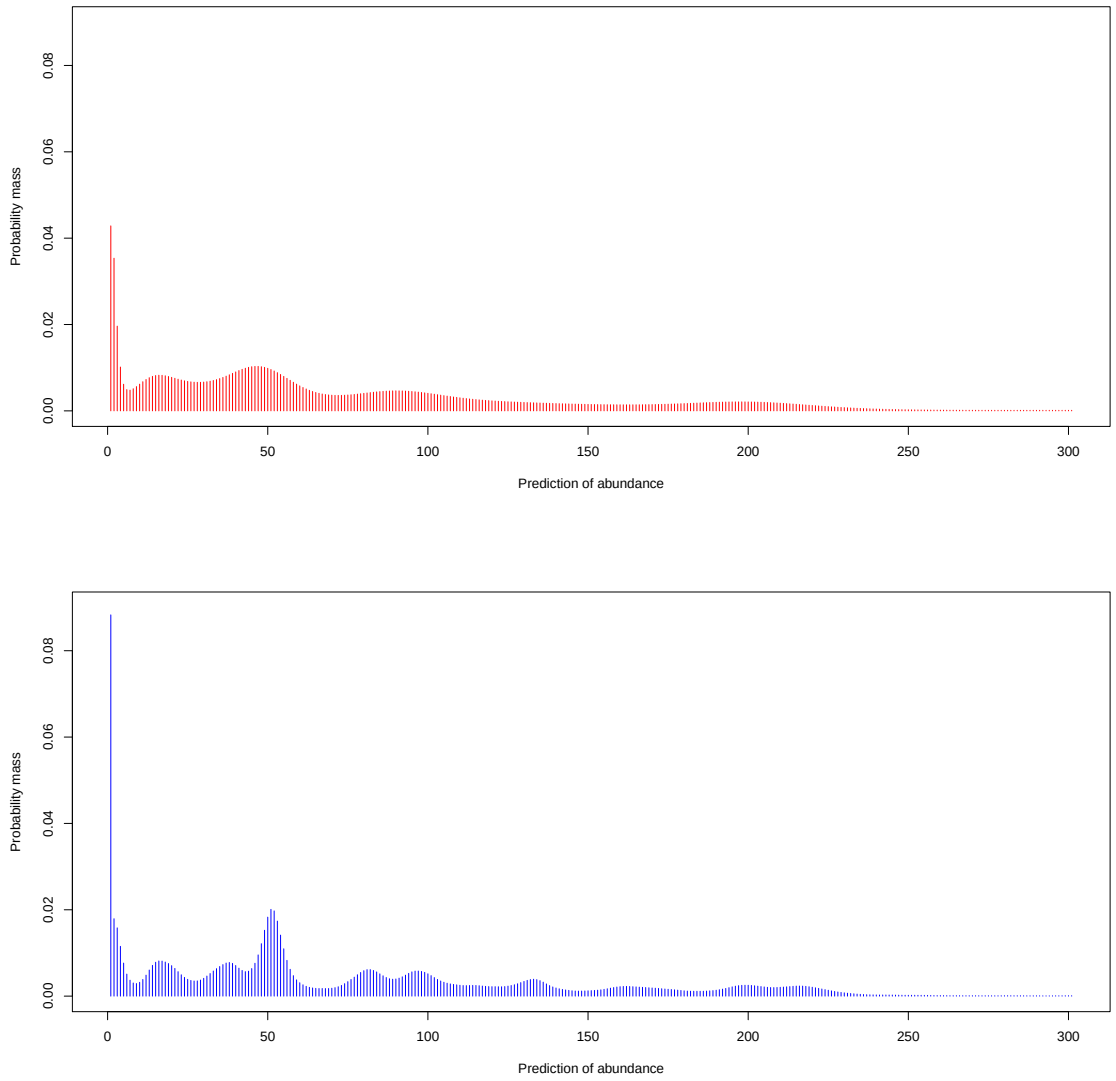

Figure 1. Posterior estimates of local abundance Y for the Okaloosa darters dataset: Poisson mixture (upper panel) and RG mixture (lower panel) with $\sigma = 0.25$ and $\mathbb{E}[K_n] = 22$.

ponents. This is quite a delicate point as already noted in Dorazio et al. (2008), where the authors remark that with a Dirichlet process mixture based on a Poisson kernel the inferential output is heavily dependent on the specification of the total mass parameter θ of the Dirichlet process. This undesirable feature cannot be removed even by putting a prior on θ since the results would then depend on the chosen prior. Dorazio et al. (2008) were the first to highlight this crucial unpleasant behavior, which probably went unnoticed because Dirichlet process mixtures are typically used with continuous kernels and most often mixtures with a small number of components are considered. The authors then circumvented the problem by adopting the empirical Bayes procedure of McAuliffe et al. (2006) to estimate θ and obtained reasonably stable results with a posterior estimate of 22 mixture components.

Here we have a closer look at this important phenomenon. Figure 2 displays the posterior mean number of components used to fit the data by the Poisson and RG mixture models as σ varies and with the 4 different prior specifications $\mathbb{E}[K_n] = 10, 22, 30, 40$. First consider Poisson mixtures. For the Dirichlet process case ($\sigma = 0$) and prior centering $\mathbb{E}[K_n] = 22$, one obtains the same estimate as Dorazio et al. (2008) for the number of mixture components. However, still with $\sigma = 0$, by varying the prior centering and considering $10, 30, 40$, the mean number of components differ significantly and only slightly moves towards the desired 22 components. The unpleasant influence of prior misspecifications on the estimated number of components is well–known in the case of continuous mixtures, where it can be fixed by employing a nonparametric prior more flexible than the Dirichlet process (see Lijoi et al., 2007). In our case this means allowing σ to be different from 0 or, in other terms, using the full range admitted by the PY process. One would then expect

that this should fix the problem also for discrete mixtures. Figure 2 shows that this is not the case and that allowing σ to vary results only in increasing the number of estimated mixture components as σ increases. This is clearly due to the inflexibility of the Poisson kernel, which is not able to benefit from the greater flexibility at the level of the mixing measure and uses it only to increase the number of components resulting in almost erratic behavior. Hence, with a Poisson mixture, the task of estimating the mixture components of the present dataset in a robust way is essentially an impossible task. This discovery and its methodological implications will be explored in depth in Section 3. Turning to the RG mixtures one notes the usual sensitivity w.r.t. the prior specification of $\mathbb{E}[K_n]$ for the Dirichlet process case ($\sigma = 0$). However, for larger σ the estimates shrink closer to each other and, regardless of the prior centering of $\mathbb{E}[K_n]$, essentially agree on about 30 components for $\sigma = 0.75$. The path and tendency to overall stability of the estimates is neat and means that, with a more flexible kernel like the RG, the mixture model is able to make a good use of the added flexibility at the mixing measure level. This phenomenon will be further investigated through a large simulation study in Section 3. Finally, further evidence of the described behaviors can be deduced from Figures 1 and 2 in the SM, where the corresponding posterior distributions are depicted.

Summing up, we discovered that the local abundance distribution of Okaloosa darters dataset is a highly complex mixture with zero-inflation and over- and under-dispersed components. This leads to overwhelming evidence of the limitations of Poisson mixture models in terms of estimation of both the pmf and the number of mixture components. These limitations cannot be remedied by using a more flexible mixing measure, which merely results in further highlighting these. Instead, once a sufficiently flexible kernel, such as the RG, is chosen, the benefit of a general nonparametric component is apparent and inferences can be robustified by choosing a prior more general than the DP.

3. Simulation study

By means of an extensive simulation study we now further investigate the behavior of both Poisson and RG mixtures driven by a PY process. In order to exclude a possible influence of the sampling protocol on the inferential outcome, we assume to directly observe the count data Y. As the results will show, the behaviors emerged in the application do not depend on it and are confirmed by the simulation study. Our goal is to compare the performance of the two competing mixture models in terms of performance in estimating the pmf of count data and providing robust estimates of the number of components of the data generating distribution. Three types of data generating distributions are considered: RG mixtures, Poisson mixtures and complex mixtures made of components belonging to different distributions. The corresponding pmf, from which the data are generated, are displayed in Figure 3 of the SM. The first (second) type serves to test the RG (Poisson) mixture in the most favorable situation, i.e. when data are drawn from a mixture made of the same kernels, and verifies whether it detects the correct number of components. Note that outside this favorable scenario, one cannot expect to detect the correct number of components. In fact, when fitting a mixture of kernels k_a with a mixture of kernels k_b, the number of kernels k_b needed is different, and typically larger, than the correct number of kernels k_a. However, it is crucial that the estimate of the number components is robust w.r.t. the prior specification leading to consistently stable estimates. Such a robustness should also hold w.r.t. to increases in the sample size, although some moderate increase in the estimated components as the sample size increases is reasonable. In fact, a larger sample implies potentially more components in nonparametric model and it is natural that when using kernels k_b to fit an k_a kernel mixture some of these potential new components will be used to produce a better fit.

As for the RG and Poisson mixtures data generating distributions, for each we consider three scenarios with $k_0 = 3, 6, 12$ components and generate datasets of size $n = 50, 100, 200$ on which the models will be tested. In terms of prior specification of the nonparametric models, we vary the key parameter of the PY process σ considering 0, 0.25, 0.5 and 0.75. In addition, to make the models comparable and to check their sensitivity w.r.t. to misspecifications we allow the prior expected number of components, $\mathbb{E}[K_n]$ to be equal to $3, 6, 12, 24$. This means that for each of the 9 samples (as k_0 and n vary) we have 16 estimates (as σ and $\mathbb{E}[K_n]$ vary) allowing to closely inspect the robustness of the model.

Consider first the case of RG mixtures with data generating distribution a RG mixture. This is clearly a benchmark test for the RG mixture model and the posterior mean number of components are reported in Table 2. If the prior expected number of components of the model, $\mathbb{E}[K_n]$ is centered on the correct one k_0 (i.e. 3, 6 or 12 in Table 2), the posterior estimated number of components sticks to the truth with minimal variability as σ varies, hence satisfying this minimal requirement. The key question is then whether the estimated number of components is close to the truth also when the model is misspecified in the sense of centering it on a different number of prior expected components. Table 2 shows that this is the case. For instance, in the case of $k_0 = 6$ true components and $n = 100$, when the prior is centered on 3 components, the posterior estimated number of components increases towards the truth, whereas it decreases towards the truth when centered in 12 or 24. This holds for any value of σ. Moreover, a closer look at the estimates, as σ varies, shows these are significantly better for larger σ implying that a large σ allows to overcome prior misspecifications in a much more effective way. Analogous considerations hold for all other cases. Importantly, from a modeling perspective, this shows that RG mixtures benefit from using a more flexible mixing measure, i.e. with a large σ, to overcome prior misspecifications. This is consistent with the findings in the case of nonparametric mixtures for continuous data. See Lijoi et al. (2007) and De Blasi et al. (2015).

Now consider the case of Poisson mixtures with data generated from a RG mixture. The estimated number of components are also reported in Table 2. If the true data generating distribution is made of 3 RG components, the model behaves relatively well. The estimated number of components stabilizes around 4 components as both the value of σ and the sample size increase. The only exception is the case of $n = 50$ with prior centering on 24 components, where however one can see that the estimate moves in the right direction as σ increases. Recall that the specific estimated value of the number components is not crucial given the data are not generated from a Poisson mixture. What is important is the robustness of the inferential outcome with respect to different prior specifications (and misspecifications). If we move on to considering mixtures made of 6 components, the estimated number of components settles around 13-16 components for $\sigma = 0.75$, but things start to become unstable as σ, n and k_0 vary. This is then apparent for the case of the 12 components data generating mixture where things derail: the added model flexibility connected to larger σ's induces the model to add more and more components rather than to adapt quickly to specific

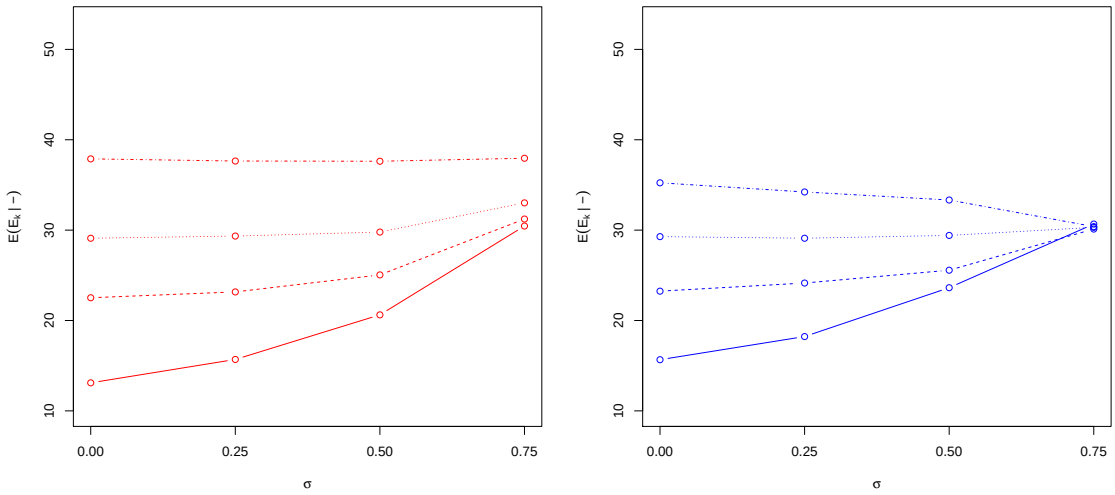

Figure 2. Posterior mean number of distinct clusters $\mathbb{E}[K_n|-]$ for the Okaloosa darters dataset: Poisson mixture (left panel) and RG mixture (right panel) for $\sigma = 0, 0.25, 0.5, 0.75$ and prior expected number of components $\mathbb{E}(K_n)$ equal to 10 (continuous line), 22 (dashed line), 30 (dotted line), and 40 (dash-dot line). Lines are connected for visualization purposes only.

value. In fact, the estimated number of components is increasing in σ, regardless of the prior centering and the sample size, leading for $\sigma = 0.75$ to estimated number of components of about 24, 45 and 60 for samples sizes of 50, 100 and 200, respectively. This means that, with a rigid kernel like the Poisson, adding flexibility to the mixing measure does not bring any benefit and actually adds to the instability. Such phenomena are not known in the case of continuous kernels and seem to be specific to the discrete case; to the authors knowledge this is the the first time such erratic behaviors are reported in the literature. From a methodological point of view the implications are clear: in order to gain the model flexibility required by count data, it is not enough to enrich the mixing measure since this is neutralized by rigid kernels. To gain flexibility both kernel and mixing measure are to be made more flexible at the same time. And, a RG kernel combined with a PY process appear to be sensible and effective choice.

Now consider the second type of data generating distribution, namely that of Poisson mixtures. The full results are reported in Table 1 of the SM. Here we limit ourselves in displaying Figure 3, which depicts the posterior mean number of components for both models estimated on the basis of samples of size $n = 50, 100, 200$ generated from a mixture of $k_0 = 6$ Poisson distributions. In fact, the plot suffices to show the erratic behavior of Poisson mixtures, which are not able to detect the correct number of mixture components (although the data are generated by a Poisson mixture). Moreover, as before, adding flexibility to the mixing measure by increasing σ results in a strong overestimation of the mixture components. For the RG mixture model the behavior is exactly the opposite: the estimated number of components stabilizes around 10, which is reasonable given the data are not generated by a RG mixture, and the larger σ the more the prior misspecification on the components number is overcome.

As far as the estimated pmf are concerned, the plots, corresponding to the two types of data generating distributions considered so far, are reported in Figures 4 and 5 of the SM. The greater flexibility and robustness of RG mixtures are clear as well as the poor fit and rigidity of Poisson mixtures. However, the differences are less apparent at the pmf level given the number of employed components is typically difficult to visualize and, more importantly, the considered data generating distributions have a quite regular structure with components of the same type.

Things change dramatically when considering more complex data generating distributions with components of different shape. As we will see the rigidity of the Poisson mixture emerges strikingly also at the pmf estimation level. In particular, the first scenario we consider corresponds to a data generating distribution with 6 component pmf of the form

$$.05\delta_0(\,\cdot\,)+.2\text{Poi}(\cdot;10)+.1\text{B}(\cdot;100,.6)+.15\text{B}(\cdot;100,.6)+.2\text{R-Poi}(\cdot;40,9)+.3 \tag{6}$$

where $\text{B}(\cdot;n,\pi)$ is a binomial with $n \in \mathbb{N}$ and $\pi \in [0,1]$, and R-Poi$(\cdot;m,\lambda)$ and NC-Poi$(\cdot;m,\lambda)$ are, respectively, a reverse and non-central Poisson, i.e.

$$\text{R-Poi}(j;m,\lambda) \propto \frac{\lambda^{m-j}}{(m-j)!}\exp\{-\lambda\} \text{ for } j = 1,\dots,m$$

$$\text{NC-Poi}(j;m,\lambda) = \frac{\lambda^{j-m}}{(j-m)!}\exp\{-\lambda\} \text{ for } j = m, m+1,\dots.$$

The second scenario corresponds to a 9 component mixture with pmf of the form

$$.1\delta_0(\,\cdot\,)+.05\delta_1(\,\cdot\,)+.3\text{Poi}(\cdot;5)+.05\text{Poi}(\cdot;1)+.15\text{B}(\cdot;25,0.8)$$
$$+.2\text{R-Poi}(\cdot;45,6)+.05\text{R-Poi}(\cdot;40,3)+.05\text{NC-Poi}(\cdot;45,7)+.05\text{NC-Poi}(\cdot;50,$$

Both data generating mixtures are depicted in the third row of Figure 3 of the SM. As far as their estimation is concerned, we compare our two competing models. For samples of sizes $n = 100$, the posterior pmf, corresponding to both the Poisson and RG nonparametric mixtures with $\mathbb{E}[K_n] = 6$ and $\sigma = 0.75$, are depicted in Figure 4. The

Table 2

Posterior mean number of mixture components $\mathbb{E}(K_n|-)$ for the simulated datasets. Data generated from RG mixtures with $k_0 = 3, 6, 12$ components and samples sizes $n = 50, 100, 200$. Results for Poisson mixtures and RG mixtures and for $\sigma = 0, 0.25, 0.50, 0.75$ and prior expected number of components $\mathbb{E}(K_n) = 3, 6, 12, 24$.

			Mixture of Poissons				Mixture of Rounded Gaussians			
			σ							
k_0	n	$\mathbb{E}[K_n]$	0	0.25	0.50	0.75	0	0.25	0.50	0.75
3	50	3	2.77	3.11	3.26	4.95	2.80	2.85	3.02	3.11
		6	4.56	4.63	4.83	4.94	4.22	3.66	3.33	3.10
		12	8.22	7.38	6.28	5.57	6.99	5.65	4.22	3.40
		24	16.26	14.38	12.67	10.16	12.66	10.43	7.58	4.48
	100	3	2.76	2.97	3.40	4.28	2.77	2.84	3.05	3.18
		6	4.16	4.26	4.38	4.42	4.22	3.56	3.26	3.21
		12	8.26	6.53	4.79	4.48	7.00	5.13	3.82	3.34
		24	17.16	14.72	10.57	5.28	11.86	8.89	5.58	3.70
	200	3	2.66	2.79	3.22	3.66	2.88	2.86	3.00	3.15
		6	3.99	4.04	4.07	4.11	3.93	3.30	3.07	3.05
		12	7.45	5.38	4.06	4.01	6.59	4.62	3.41	3.25
		24	15.52	11.51	6.47	4.77	11.52	7.45	4.45	3.31
6	50	3	4.02	4.88	7.30	12.93	4.24	5.13	5.79	6.09
		6	6.23	6.61	8.75	13.17	5.85	5.90	5.97	6.05
		12	11.15	11.26	11.53	14.82	8.51	7.59	6.82	6.17
		24	23.18	22.84	21.83	19.65	12.91	11.26	9.38	7.00
	100	3	3.97	5.16	8.14	14.31	4.27	5.33	5.88	6.14
		6	5.93	6.80	8.68	13.78	5.92	5.99	6.15	6.20
		12	11.13	11.16	11.24	14.49	8.90	7.83	6.78	6.40
		24	21.78	20.74	17.79	15.51	13.24	11.06	8.26	6.50
	200	3	3.78	4.99	7.62	14.24	3.96	5.02	5.53	5.95
		6	5.86	7.04	9.03	15.93	5.45	5.49	5.61	5.90
		12	10.08	10.22	10.29	16.38	8.51	6.91	5.97	5.90
		24	21.74	18.99	16.07	17.08	13.50	9.94	7.05	6.01
12	50	3	6.27	9.28	14.59	23.80	6.36	7.80	8.69	9.38
		6	8.33	10.55	15.17	23.86	7.61	8.13	8.61	9.53
		12	13.13	14.35	16.85	24.13	9.17	9.24	9.37	9.59
		24	23.25	23.38	23.84	26.40	13.20	12.17	11.07	10.02
	100	3	6.78	11.98	22.84	44.77	6.53	10.21	11.76	12.26
		6	9.38	13.54	23.96	45.63	9.15	10.92	11.65	11.87
		12	14.19	17.25	25.23	45.50	12.65	12.48	12.19	12.08
		24	24.96	26.05	29.82	45.89	17.22	15.06	13.52	12.12
	200	3	7.02	13.14	27.98	57.54	6.02	10.25	11.30	11.96
		6	9.36	14.69	27.96	60.07	9.19	10.80	11.97	12.19
		12	14.20	17.90	30.19	57.56	12.39	12.26	12.23	12.02
		24	24.24	26.12	33.27	62.82	17.53	15.48	13.13	12.37

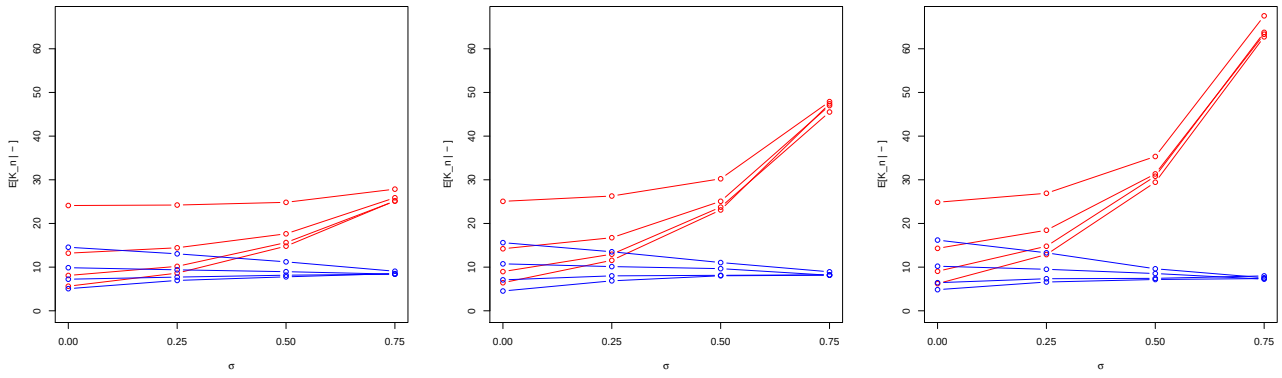

Figure 3. Posterior mean number of mixture components $\mathbb{E}[K_n|-]$ for simulated datasets. Data generated from a $k_0 = 6$ Poisson mixture and results reported for samples sizes $n = 50, 100, 200$ (from left to right). Each plot depicts the posterior mean for both Poisson (red) and RG (blue) models for $\sigma = 0, 0.25, 0.50, 0.75$ and $\mathbb{E}[K_n] = 3, 6, 12, 24$. Lines are connected for visualization purposes only.

evidence concerning the lack of flexibility of the Poisson mixture model is indisputable and clearly shows its inability to fit under-dispersed components and overly smooth different components with locations far from zero. On the other hand the RG mixture model has a completely satisfactory performance being able to closely resemble the data generating distributions. Analogous behaviors arise for different prior specifications and sample sizes. For instance, for the case (7), their performance as the sample sizes varies is shown in Figure 6 of the SM. The results concerning the posterior mean number of components are reported in Table 2 of the SM. Given (6) and (7) have little in common with both Poisson and RG mixtures and the components are irregular, it is not surprising that the nonparametric model uses more components than the actual ones. However, exactly as in the cases considered before, the RG nonparametric mixture stabilizes around the used number of components as σ increases being able to overcome prior misspecifications. The Poisson nonparametric mixture, instead, is again characterized by instability and erratic behavior.

The considered scenarios are not particular cases and are confirmed by several other simulation studies not reported here. Although there may be cases in which also a Poisson mixture well approximates true pmf with the correct number of mixture components, practitioners are to warned to using nonparametric Poisson mixtures with caution.

4. Concluding remarks

We considered an application concerning surveys of natural populations of animals with significant spatial heterogeneity in the corresponding counts. Given the need for nonparametric modeling in such contexts as proven in Dorazio et al. (2008), we adopted a Bayesian nonparametric approach and innovated previous studies by considering mixture models with more flexible both kernel and mixing measure. This leads to more accurate estimation of the pmf of local abundance and to a more robust quantification of its components. Starting from these findings, we enlarged the goal to deduce general methodological implications via an extensive simulation study. We discovered that adding flexibility to a Poisson mixture model by generalizing the nonparametric mixing measure is severely limited by the rigidity of the Poisson kernel and leads to a full display of the instability of Poisson mixtures in estimating the number of mixture components. In contrast, if a sufficiently flexible kernel, such as the RG, is chosen, inferences become more accurate and robust by choosing a prior more general than the DP. Overall inferences for count data are improved when simultaneously selecting both kernel and mixing measure more general than the standard DP mixture with Poisson kernel.

Appendix

A.1 Prior elicitation for the Pitman–Yor process

Given a sample $X_1, \ldots, X_n$ generated by a $\text{PY}(\theta, \sigma; P_0)$ process the expected number of distinct values, K_n, is equal to

$$\mathbb{E}[K_n] = \sum_{i=1}^{n} \frac{(\theta+\sigma)_{i+1}}{(\theta+1)_{i+1}} = \begin{cases} \sum_{i=1}^{n} \frac{\theta}{\theta+i-1} & \text{if } \sigma = 0 \\ \frac{(\theta+\sigma)_n}{\sigma(\theta+1)_{n-1}} - \frac{\theta}{\sigma} & \text{if } \sigma > 0 \end{cases}$$

where $(a)_n = \Gamma(a+n)/\Gamma(a)$ is the ascending factorial coefficient. See Pitman (2006).

The previous relations can be readily used to the identify θ such that $\mathbb{E}[K_n]$ is equal to a desired value for any given σ and sample size n with straightforward numerical methods.

A.2 Gibbs sampling algorithm for Section 2

The Gibbs sampling algorithm set forth in Section 2.2 iterates the following steps.

(1) For each $i = 1, \ldots, n$

- generate a candidate $\tilde{Y}_i^*$ from $N_{A(x_i)}(\mu_i, \tau_i^{-1})$, were $A(x_i) = \{Y^* : Y^* \geqslant a_{x_i}\}$;
- let $\tilde{Y}_i = s$ if $a_s \leqslant \tilde{Y}_i^* < a_{s+1}$
- keep the proposed value with probability

$$\min\left\{1, \frac{\tilde{Y}_i!(Y_i - x_i)!}{Y_i!(\tilde{Y} - x_i)!}(1-\pi_i)^{J_i(\tilde{Y}_i - y_i)}\right\}$$

(2) Let $S_1, \ldots, S_n$ be the current cluster allocation. For $i = 1, \ldots, n$ let $H_{\setminus i}$ the set of distinct values of S_j for $j \neq i$ with $k_{\setminus i}$ its cardinality. Then allocate the i-th observation to one existing cluster $h \in H_{\setminus i}$ or to a new cluster h^* with the following probability

$$\Pr(S_i = h|-) \propto \begin{cases} (n_h - \sigma)N(Y_i^*; \mu_h, \tau_h^{-1}) & \text{for } h \in H_{\setminus i} \\ (\theta + k_{\setminus i}\sigma)N(Y_i^*; \mu_*, \tau_*^{-1}) & \text{for } h = h^* \end{cases}$$

where n_h is the cluster size (excluding the i-th observation), and (μ_*, τ_*) are a new draw from P_0.

(3) Update (μ_h, τ_h) from its conditional posterior

$$(\mu_h, \tau_h^{-1}) \sim \text{N}(\hat{\mu_h}, \hat{\kappa}_h\tau_h^{-1})\text{Ga}(\hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

with $\hat{a}_{\tau_h} = a_\tau + n_h/2$, $\hat{b}_{\tau_h} = b_\tau + 1/2(\sum_{i:S_i=h}(Y_i^* - \bar{Y}_h^*) + n_h/(1+\kappa n_h)(\bar{Y}_h^* - \mu_0)^2)$, $\hat{\kappa}_h = (\kappa^{-1} + n_h)^{-1}$ and $\hat{\mu}_h = \hat{\kappa}_h(\kappa^{-1}\mu_0 + n_h\bar{Y}_h^*)$.

(4) For each $i = 1, \ldots, n$, update π_i from

$$\pi_i \sim \text{Beta}(a_\pi + x_i, b_\pi + J_i(Y_i - x_i) - x_i + \sum_{j=1}^{J_i} j x_{ij}).$$

ACKNOWLEDGMENTS

I. Prünster is supported by the European Research Council (ERC) through StG "N-BNP" 306406.

SUPPLEMENTARY MATERIAL

The Online Supplementary Materials report additional plots for the Okaloosa darters dataset analyzed in Section 2 and additional plots and tables for the simulation study of Section 3 and are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

REFERENCES

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science* **28,** 313–334.

Brown, G. O. and Buckley, W. S. (2015). Experience rating with Poisson mixtures. *Annals of Actuarial Science* **9,** 304–321.

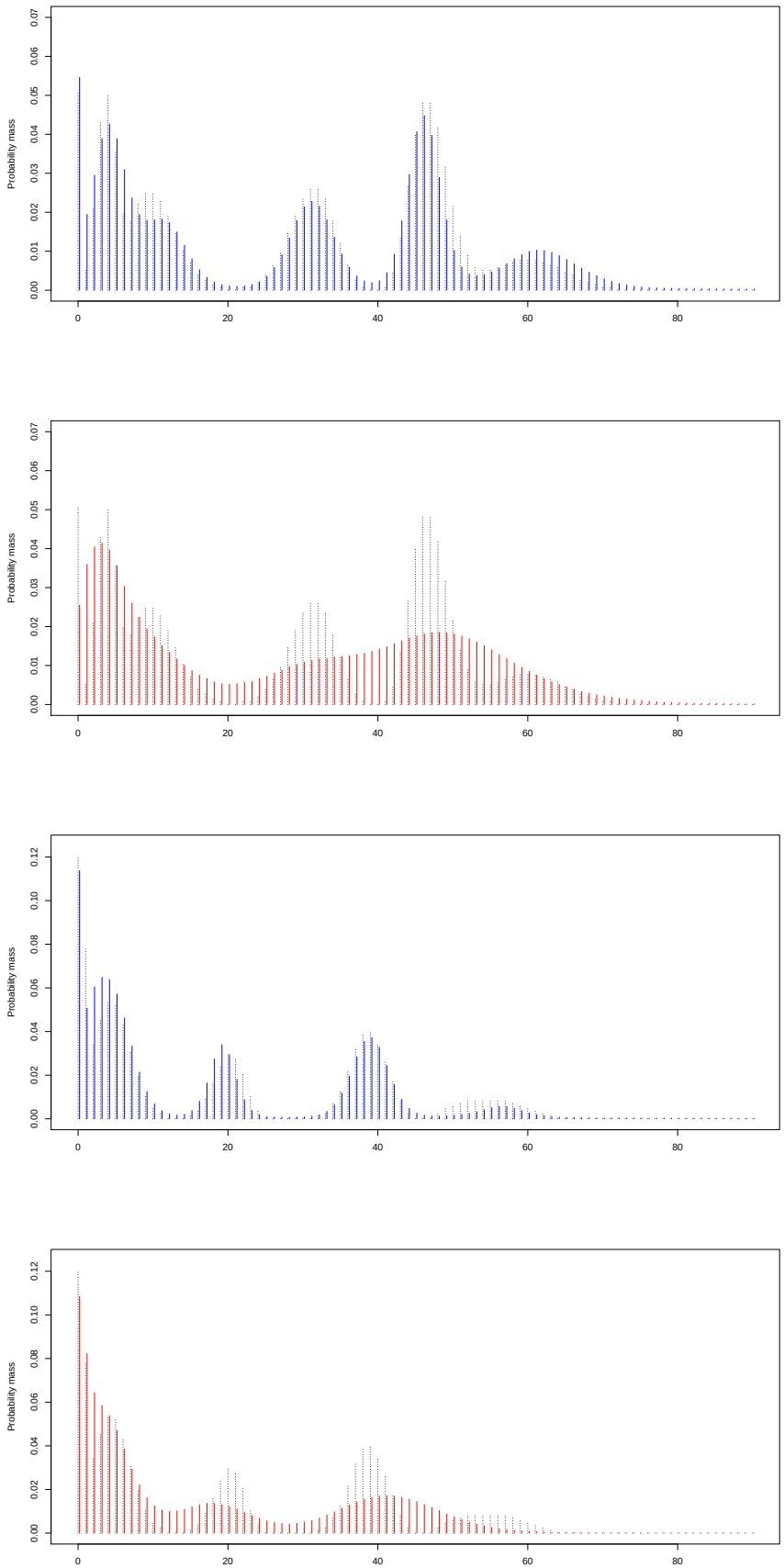

Figure 4. Posterior pmf for the two complex scenarios given $n = 100$ data generated from (6) (upper two plots) and (7) (lower two plots). The first and third figures display the posterior estimates of the RG model (depicted in solid blue), whereas the second and fourth of the Poisson model (depicted in solid red) with $\mathbb{E}[K_n] = 6$ and $\sigma = 0.75$. The true pmf are in dotted grey.

Bush, C. A. and MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83,** 275–285.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association* **106,** 1528–1539.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37,** 212–229.

Dorazio, R. M., Jelks, H. L., and Jordan, F. (2005). Improving removal-based estimates of abundance by sampling a population of spatially distinct subpopulations. *Biometrics* **61,** 1093–1101.

Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., and Jordan, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64,** 635–644.

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

Guindani, M., Do, K., Müller, P., and Morris, J. (2006). Bayesian mixture models for gene expression and protein profiles. In Do, K., Müller, P., and Vannucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*, pages 238–253. Cambridge University Press.

Guindani, M., Sepúlveda, N., Paulino, C. D., and Müller, P. (2014). A bayesian semiparametric approach for the differential analysis of sequence counts data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **63,** 385–404.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hougaard, P., Lee, M.-L. T., and Whitmore, G. A. (1997). Analysis of overdispersed count data by mixtures of Poisson variables and poisson Processes. *Biometrics* **53,** 1225–1238.

Ishwaran, H. and James, Lancelot, F. (2001). Gibbs sampling methods for stick breaking priors. *Journal of the American Statistical Association* **96,** 161–173.

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

Li, Q., Guindani, M., Reich, B., Bondell, H., and Vannucci, M. (2015). A Poisson mixture model for clustering and feature selection of high-dimensional count data under mean constraints. Technical report, Rice University.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100,** 1278–1291.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69,** 715–740.

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

McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing* **16,** 5–14.

Müller, P., Quintana, F., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis*. Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* **9,** 249–265.

Pitman, J. (2006). *Combinatorial stochastic processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875. Springer, New York.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability* **25,** 855–900.

Royle, J. A. and Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press.

Viallefont, V., Richardson, S., and Green, P. J. (2002). Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics* **14,** 181–202.

Received November 2015.