Bayesian Kernel Mixtures for Counts

Antonio Canale* & David B. Dunson†

Abstract

Although Bayesian nonparametric mixture models for continuous data are well developed, there is a limited literature on related approaches for count data. A common strategy is to use a mixture of Poissons, which unfortunately is quite restrictive in not accounting for distributions having variance less than the mean. Other approaches include mixing multinomials, which requires finite support, and using a Dirichlet process prior with a Poisson base measure, which does not allow smooth deviations from the Poisson. As a broad class of alternative models, we propose to use nonparametric mixtures of rounded continuous kernels. We provide sufficient conditions on the kernels and prior for the mixing measure under which almost all count distributions fall within the Kullback-Leibler support. This is shown to imply both weak and strong posterior consistency. An efficient Gibbs sampler is developed for posterior computation, and a simulation study is performed to assess performance. Focusing on the rounded Gaussian case, we generalize the modeling framework to account for multivariate count data, joint modeling with continuous and categorical variables, and other complications. The methods are illustrated through an application to marketing data.

Keywords: Bayesian nonparametrics; Dirichlet process mixtures; Kullback-Leibler condition; Large support; Multivariate count data; Posterior consistency; Rounded Gaussian distribution

*Dip. Scienze Statistiche, Università di Padova, 35121 Padova, Italy (*canale@stat.unipd.it*)

†Dept. Statistical Science, Duke University, Durham, NC 27708, USA (*dunson@stat.duke.edu*)

1. INTRODUCTION

Nonparametric methods for estimation of continuous densities are well developed in the literature from both a Bayesian and frequentist perspective. For example, for Bayesian density estimation, one can use a Dirichlet process (DP) (Ferguson 1974, 1973) mixture of Gaussians kernels (Lo 1984; Escobar and West 1995) to obtain a prior for the unknown density. Such a prior can be chosen to have dense support on the set of densities with respect to Lebesgue measure. Ghosal et al. (1999) show that the posterior probability assigned to neighborhoods of the true density converges to one exponentially fast as the sample size increases, so that consistent estimates are obtained. Similar results can be obtained for nonparametric mixtures of various non-Gaussian kernels using tools developed in Wu and Ghosal (2008).

In this article our focus is on nonparametric Bayesian modeling of counts using related nonparametric kernel mixture priors to those developed for estimation of continuous densities. There are several strategies that have been proposed in the literature for nonparametric modeling of count distributions having support on the non-negative integers $\mathcal{N} = \{0, \ldots, \infty\}$. The first is to use a mixture of Poissons, so that one could let

$$\Pr(Y = j \mid P) = \int \mathrm{Poi}(j; \lambda) dP(\lambda), \quad j \in \mathcal{N}, \tag{1}$$

with $\mathrm{Poi}(j; \lambda) = \lambda^j \exp(-\lambda)/j!$ and P a mixture distribution. When P is chosen to correspond to a $\mathrm{Ga}(\phi, \phi)$ distribution on the Poisson rate parameter, one induces a negative-binomial distribution, which accounts for over-dispersion with the variance greater than the mean. Such parametric mixtures of Poissons are widely used in richer hierarchical models including predictors and random effects to account for dependence. A good review of the properties of Poisson mixtures is provided in Karlis and Xekalaki (2005).

As a more flexible nonparametric approach, one can instead choose a DP mixture of Poissons by letting $P \sim \mathrm{DP}(\alpha P_0)$, with α the DP precision parameter and P_0 the base measure. As the DP prior implies that P is almost surely discrete, we obtain

$$\Pr(Y = j | \pi, \lambda) = \sum_{h=1}^{\infty} \pi_h \mathrm{Poi}(j; \lambda_h), \quad \lambda_h \sim P_0, \tag{2}$$

2

with $\pi = \{\pi_h\} \sim \text{Stick}(\alpha)$ denoting that the π are random weights drawn from the stick-breaking process of Sethuraman (1994). Krnjajic et al. (2008) recently considered a related approach motivated by a case control study. Dunson (2005) proposed an approach for nonparametric estimation of a non-decreasing mean function, with the conditional distribution modeled as a DPM of Poissons. Kleinman and Ibrahim (1998) proposed to use a DP prior for the random effects in a generalized linear mixed model, with DPMs of Poisson log-linear models arising as a special case. Guha (2008) recently proposed more efficient computational algorithms for related models. Chen et al. (2002) considered nonparametric random effect distributions in frequentist generalized linear mixed models.

On the surface, model (2) seems extremely flexible and to provide a natural modification of the DPM of Gaussians used for continuous densities. However, as the Poisson kernel used in the mixture has a single parameter corresponding to both the location and scale, the resulting prior on the count distribution is actually quite inflexible. For example, distributions that are under-dispersed cannot be approximated and will not be consistently estimated. One can potentially use mixture of multinomials instead of Poissons, but this requires a bound on the range in the count variable and the multinomial kernel is almost too flexible in being parametrized by a probability vector equal in dimension to the number of support points. Kernel mixture models tend to have the best performance when the effective number of parameters is small. For example, most continuous densities can be accurately approximated using a small number of Gaussian kernels having varying locations and scales. It would be appealing to have such an approach available also for counts.

An alternative nonparametric Bayes approach would be to avoid a mixture specification and instead let $y_i \sim P$ with $P \sim \text{DP}(\alpha P_0)$ and P_0 corresponding to a base parametric distribution, such as a Poisson. Carota and Parmigiani (2002) proposed a generalization of this approach in which they modeled the base distribution as dependent on covariates through a Poisson log-linear model. Although this model is clearly flexible in having large support and seemingly leading to consistent estimation

of the count distribution, there are some major disadvantages in terms of finite sample characteristics. To illustrate the problems that can arise, first note that the posterior distribution of P given iid draws $y^n = (y_1, \ldots, y_n)'$ is simply

$$(P \mid y^n) \sim \mathrm{DP}\left((\alpha + n)\left\{\alpha P_0 + \sum_i \delta_{y_i}\right\}\right),$$

with δ_y a degenerate distribution with all its mass at y. Hence, the posterior is centred on a mixture with weight proportional to α on the Poisson base P_0 and weight proportional to n on the empirical probability mass function. There is no allowance for smooth deviations from the base.

As a motivating application, we consider data from a developmental toxicity study of ethylene glycol in mice conducted by the National Toxicology Program (Price et al. 1985). As in many biological applications in which there are constraints on the range of the counts, the data are underdispersed having mean 12.54 and variance 6.78. A histogram of the raw data is shown in Figure 1 along with a series of estimates of the posterior mean of $\Pr(Y = j)$ assuming $y_i \sim P$ with $P \sim \mathrm{DP}(\alpha P_0)$, $\alpha = 1$ or 5, and $P_0 = \mathrm{Poi}(\overline{y})$ as an empirical Bayes choice. To illustrate the behavior as the sample size increases we take random subsamples of the data of size $n_s \in \{5, 10, 25, 50\}$. As Figure 2 illustrates, the lack of smoothing in the Bayes estimate is unappealing in not allowing borrowing of information about local deviations from P_0. In particular for small sample size as in panels (a) and (b) the posterior mean probability mass function is the base measure with high peaks on the observed y while as the sample size increase as in panels (c) and (d) it is almost just the empirical probability mass function.

With this motivation, we propose a general class of kernel mixture models for count data, with the kernels induced through rounding of continuous kernels. In Section 2 we provide theory providing sufficient conditions on the kernels and prior on the mixing measure under which all distributions on $\mathcal{N}$ are in the Kullback-Leibler support of the prior, implying weak posterior consistency under the theory of Schwartz (1965). We also provide theory showing that in the count case weak consistency implies pointwise consistency and strong consistency in the L_1 sense, and hence the Kullback-Leibler condition is sufficient. Methods are developed for efficient posterior computation using
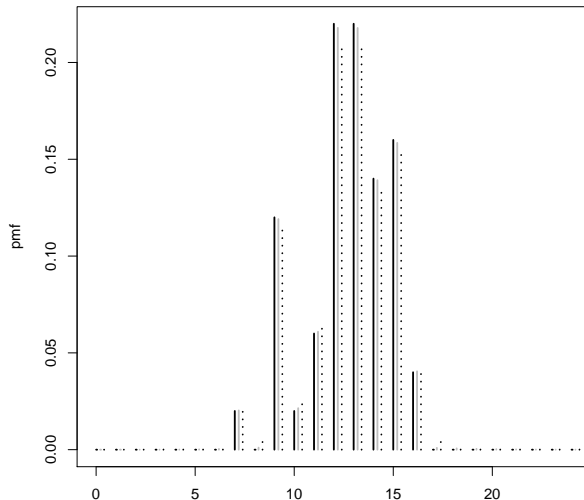
4

Figure 1: Histogram of the number of implantation per pregnant mouse (black line) and posterior mean of $\Pr(Y = j)$ assuming a Dirichlet process prior on the distribution of the number of implants with $\alpha = 1, 5$ (grey and black dotted line respectively) and base measure $P_0 = \mathrm{Poi}(\overline{y})$.

a simple data augmentation Gibbs sampler, which adapts approaches for computation in DPMs of Gaussians. The performance of the method is evaluated through a simulation study. Generalizations to multivariate count data and other complications are outlined in Section 3 along with a simulation study. Section 4 applies the methodology to a marketing problem and Section 5 discusses the results.

## 2. UNIVARIATE ROUNDED KERNEL MIXTURE PRIORS

2.1 Rounding continuous distributions

Let $y \sim p$ with $y \in \mathcal{N}$ and $p \in \mathcal{C}$, the space of probability mass functions on the non negative integers. Introduce an underlying continuous density $f \in \mathcal{L}$, the space of densities with respect to the Lebesgue measure on $\mathcal{Y}$, and a latent variable $y^* \sim f$. The space $\mathcal{Y}$ can be $\mathbb{R}$ or a measurable subset, such as $\mathbb{R}^+$ or a bounded interval $[a, b]$.
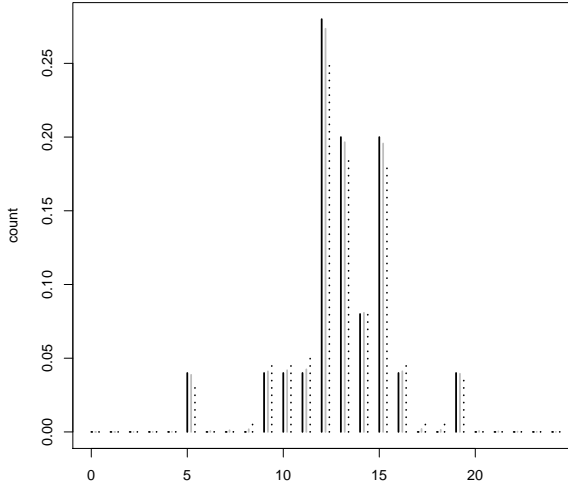
Assume that we observe $y = h(y^*)$, where $h(\cdot)$ is a thresholding function defined
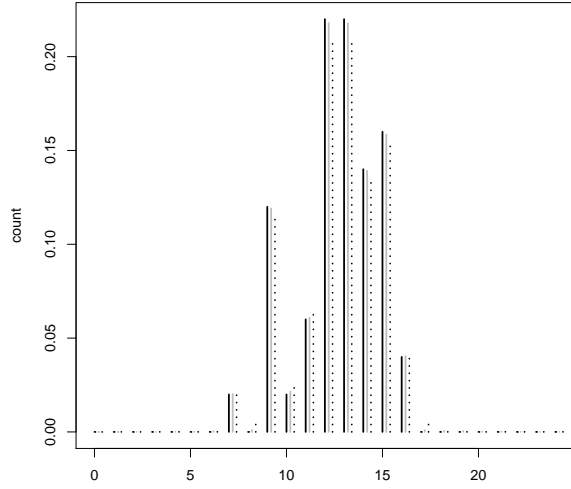
5
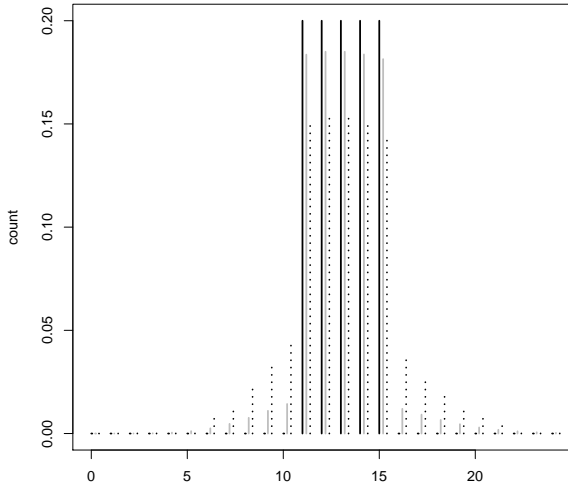
(a)
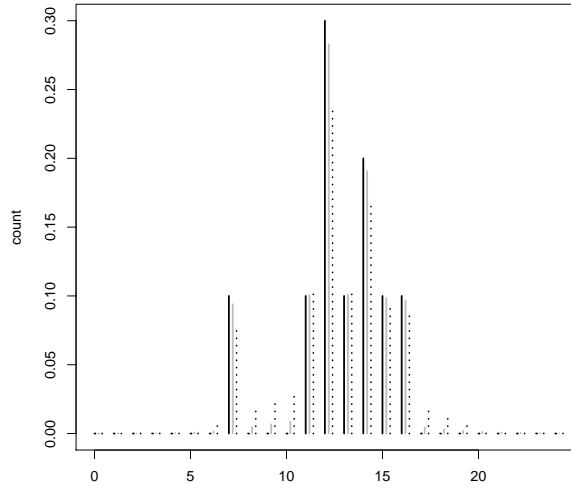
(b)

(c)

(d)

Figure 2: Histogram of subsamples $(n = 5, 10, 25, 50)$ of the data on implantation on mice (black line) and posterior mean of $\Pr(Y = j)$ assuming a Dirichlet process prior on the distribution of the number of implants with $\alpha = 1, 5$ (grey and black dotted line respectively) and base measure $P_0 = \text{Poi}(\bar{y})$.

so that $h(y^*) = j$ if $y^* \in (a_j, a_{j+1}]$, for $j = 0, 1, \ldots, \infty$, with $a_0 < a_1 < \ldots$ an infinite sequence of pre-specified thresholds that defines a disjoint partition of $\mathcal{Y}$. Then the probability mass function p of y is $p = g(f)$ where the mapping function $g(\cdot)$ has the simple form

$$p(j) = g(f)[j] = \int_{a_j}^{a_{j+1}} f(y^*)\mathrm{d}y^* \quad j \in \mathcal{N}. \tag{3}$$

The thresholds a_j are such that $a_0 = \min\{y : y \in \mathcal{Y}\}$, $a_\infty = \max\{y : y \in \mathcal{Y}\}$ and hence $\int_{a_0}^{a_\infty} f(y^*)\mathrm{d}y^* = 1$. Examples of $a_0, \ldots, a_\infty$ include $0, 1, 2, \ldots, \infty$ for an f defined on $\mathcal{Y} = \mathbb{R}^+$ and $0, 1/2, \ldots, 1 - 1/2^h, \ldots$ for an f defined on $\mathcal{Y} = [0, 1]$.

Relating ordered categorical data to underlying continuous variables is quite common in the literature. For example, Albert and Chib (1993) proposed a very widely used class of data augmentation Gibbs sampling algorithms for probit models. In such settings, one typically lets $a_0 = -\infty$ and $a_1 = 0$, while estimating the remaining $k - 2$ thresholds, with k denoting the number of levels of the categorical variable. A number of authors have relaxed the assumption of the probit link function through the use of nonparametric mixing. For example, Kottas et al. (2005) generalized the multivariate probit model by using a mixture of normals in place of a single multivariate normal for the underlying scores, with Jara et al. (2007) proposing a related approach for correlated binary data. Gill and Casella (2009) instead used Dirichlet process mixture priors for the random effects in an ordered probit model.

In the setting of count data, instead of estimating the thresholds on the underlying variables, we use a fixed sequence of thresholds and rely on flexibility in nonparametric modeling of f to induce a flexible prior on p. In order to assign a prior Π on the space of count distributions, it is sufficient under this formulation to specify a prior Π^* on the space $\mathcal{L}$. In the following section we describe some properties of the induced prior Π for the unknown count distribution p.

2.2 Properties

The involved mapping functions are both surjective and hence the inverse mapping $g^{-1}(\cdot)$ of a point $p \in \mathcal{C}$ will correspond to an uncountably infinite set of densities in $\mathcal{L}$. Similarly, the inverse mapping $h^{-1}(\cdot)$ of a point $y \in \mathcal{N}$ will correspond to a subset of

$\mathcal{Y}$ containing infinitely many y^*s. The existence of at least one element in $\mathcal{L}$ for every $p \in \mathcal{C}$ is ensured by the following lemma.

Lemma 1. *For every count measure $p_0 \in \mathcal{C}$ and rounding function $g(\cdot)$ defined in (3), there exists at least one $f_0 \in \mathcal{L}$ such that $g(f_0) = p_0$.*

Lemma 2 demonstrates that the mapping $g : \mathcal{L} \to \mathcal{C}$ maintains Kullback-Leibler neighborhoods. As is formalised in Theorem 1, this property implies that the induced prior $p \sim \Pi$ assigns positive probability to all Kullback-Leibler neighbourhoods of any $p_0 \in \mathcal{C}$ if at least one element of the set $g^{-1}(p_0)$ is in the Kullback-Leibler support of the prior Π^*. By using conditions of Wu and Ghosal (2008), the Kullback-Leibler condition becomes straightforward to demonstrate for a broad class of kernel mixture priors Π^*.

Lemma 2. *Assume that the true density of a count random variable is p_0 and choose any f_0 such that $p_0 = g(f_0)$. Let $\mathcal{K}_\epsilon(f_0) = \{f : KL(f_0, f) < \epsilon\}$ be a Kullback-Leibler neighbourhood of size ϵ around f_0. Then the image $g(\mathcal{K}_\epsilon(f_0))$ contains values $p \in \mathcal{C}$ in a Kullback-Leibler neighbourhood of p_0 of at most size ϵ.*

Theorem 1. *Given a prior Π^* on $\mathcal{L}_{\Pi^*} \subseteq \mathcal{L}$ such that all $f \in \mathcal{L}_{\Pi^*}$ are in the Kullback-Leibler support of Π^*, then all $p \in \mathcal{C}_\Pi = g(\mathcal{L}_{\Pi^*})$ are in the Kullback-Leibler support of Π.*

Corollary 1. *From the Schwartz (1965) Theorem, the posterior probability of any weak neighbourhood around the true data-generating distribution $p_0 \in \mathcal{C}_\Pi$ converges to one exponentially fast as $n \to \infty$.*

Theorem 2 and Corollary 2 point out that weak consistency implies both pointwise consistency and strong consistency in the L_1 sense. This implies that the Kullback-Leibler condition is sufficient for strong consistency in modeling count distributions.

Theorem 2. *Let p_n denote a sequence of distributions in $\mathcal{C}$. If $p_n \to p_0$ weakly, then $p_n \to p_0$ also pointwise.*

Corollary 2. *Let p_n denote a sequence of distributions in $\mathcal{C}$. If $p_n \to p_0$ weakly, then $p_n \to p_0$ in L_1.*

2.3 Some examples of rounded kernel mixture prior

For appropriate choices of kernel, it is well known that kernel mixtures can accurately approximate a rich variety of densities, with Dirichlet process mixtures of Gaussians forming a standard choice for densities on $\Re$. Hence, in our setting a natural choice of prior for the underlying continuous density corresponds to

$$f(y^*; P) = \int N(y^*; \mu, \tau^{-1}) dP(\mu, \tau),$$

$$P \sim \Pi^*, \tag{4}$$

where $N(y; \mu, \tau^{-1})$ is a normal kernel having mean μ and precision τ and Π^* is a prior on the mixing measure P, with a convenient choice corresponding to the Dirichlet process $DP(\alpha P_0)$, with P_0 chosen to be Normal-Gamma. Let Π denote the prior on p induced through (3)-(4) via the prior Π^* on P. Using $a_0 = -\infty$ and $a_h = h$ for $h \in \{1, 2, \dots\}$ we obtain that sufficiently large Kullback-Leibler support of the prior (4) on f implies full Kullback-Leibler support and weak and strong posterior consistency at any true p_0.

Other choices can be made for the prior on the underlying continuous density, such as mixtures of log-normal, gamma or Weibull densities with $a_h = h$. In general is trivial to show that for all the kernels studied in Wu and Ghosal (2008) and suitable thresholds, we obtain induced priors Π with large Kullback-Leibler support, and hence weak and strong posterior consistency follow for any true p_0 in a large subset of $\mathcal{C}$. In the following, we will focus on the class of DP mixtures of rounded Gaussian kernels for computational convenience and because there is no clear reason to prefer an alternative choice of kernel or mixing prior from an applied or theoretical perspective. As described in the following subsection, regardless of the choice of kernel the thresholds can be calibrated to center the prior on an initial guess for p.

2.4 Eliciting the thresholds

In the presence of prior information on the random p, one can define the sequence of a_j iteratively in order to let $E\{p(j)\} = q(j)$, where q is an initial guess for the probability mass function, defined for all j. Let G denote the prior on p induced through the

mapping (3) and G^*, the prior on a continuous f with CDF $F(\cdot)$. Clearly

$$E\{p(j)\} = E\left\{\int_{a_j}^{a_{j+1}} f(y)dy\right\} = E\{F(a_{j+1})\} - E\{F(a_j)\}.$$

One can express the expected value of $F(\cdot)$ as the integral

$$\int F(a_j)d\Pi = \int\int_{-\infty}^{a_j} f(z)dz\,dG$$

and then marginalize out. The variance can be also computed in a similar way considering

$$Var\{p(j)\} = Var\{F(a_{j+1})\} + Var\{F(a_j)\}$$

$$= \sum_{l=j,j+1} E\{F(a_l)^2\} + E\{F(a_l)\}^2.$$

Assuming as a prior for the underling continuous distribution the Dirichlet process mixture of Gaussians kernels in (4), we have

$$E\{F(a_j)\} = \int_{-\infty}^{a_j} t_\nu(z;\mu_0,\kappa+1) \tag{5}$$

that leads to define the thresholds iteratively as

$$a_0 = -\infty$$

$$a_1 = \mathcal{T}_\nu^{-1}(q(0);\mu_0,\kappa+1)$$

$$a_2 = \mathcal{T}_\nu^{-1}(q(0)+q(1);\mu_0,\kappa+1)$$

$$\cdots$$

$$a_j = \mathcal{T}_\nu^{-1}(\sum_{i=0}^{j-1} q(i);\mu_0,\kappa+1)$$

where $\mathcal{T}_\lambda(y;\xi,\omega)$ is the CDF of a non central Student-t distribution with λ degrees of freedom, location ξ and scale ω. The variance is

$$Var\{p(j)\} = \sum_{l=j,j+1} q(l)^2 + E\left\{\Phi(a_l;\mu,\tau^{-1})^2\right\}$$

where the second expectation can be computed numerically for fixed a_j. The derivations are outlined in the Appendix.

2.5 A Gibbs sampling algorithm

For posterior computation, we can trivially adapted any existing MCMC algorithm developed for DPMs of Gaussians with a simple data augmentation step for imputing the underlying variables. For simplicity in describing the details, we focus on the blocked Gibbs sampler of Ishwaran and James (2001), with $f(y^*) = \sum_{h=1}^{N} \pi_h N(y^*; \mu_h, \tau_h^{-1})$ with $\pi_1 = V_1$, $\pi_h = V_h \prod_{l<h}(1 - V_l)$, V_h independent Beta(1,α) and $V_N = 1$. Modifications to avoid truncation can be applied using slice sampling as described in Walker (2007) and Yau et al. (2010). The blocked Gibbs sampling steps are as follows:

Step 1 Generate each y_i^* from the full conditional posterior

  Step 1a Generate $u_i \sim U\left(\Phi(a_{y_i}; \mu_{S_i}, \tau_{S_i}^{-1}), \Phi(a_{y_i+1}; \mu_{S_i}, \tau_{S_i}^{-1})\right)$

  Step 1b Let $y_i^* = \Phi^{-1}(u_i; \mu_{S_i}, \tau_{S_i}^{-1})$

Step 2 Update S_i from its multinomial conditional posterior with

$$\Pr(S_i = h|-) = \frac{\pi_h p(y_i|\mu_h, \tau_h^{-1})}{\sum_{l=1}^{N} \pi_l p(y_i|\mu_l, \tau_l^{-1})},$$

where $p(j|\mu_h, \tau_h^{-1}) = \Phi(a_{j+1}|\mu_h, \tau_h^{-1}) - \Phi(a_j|\mu_h, \tau_h^{-1})$.

Step 3 Update the stick-breaking weights using

$$V_h \sim \text{Be}\left(1 + n_h, \alpha + \sum_{l=h+1}^{N} n_l\right)$$

Step 4 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 $\hat{\mu}_h = \hat{\kappa}_h(\kappa^{-1}\mu_0 + n_h \bar{y}_h^*)$.

2.6 Simulation study

To assess the performance of the proposed approach, we conducted a simulation study. Four different approaches for estimating the probability mass function were compared: our proposed rounded mixture of Gaussians (RMG), the empirical probability mass function (E), a Bayesian nonparametric approach assuming a Dirichlet process with

11

a Poisson base measure (DP), and maximum likelihood estimation under a Poisson model (MLE). Several simulations have been run under different simulation settings leading to qualitatively similar results. In what follows we report the results for two scenarios. The first simulation case assumed the data were simulated as the floor of draws from the following mixture of Gaussians:

$$0.4N(25, 1.5) + 0.15N(20, 1) + 0.25N(24, 1) + 0.2N(21, 2)$$

while the second case assumed a simple Poisson model with mean 12.

For each case, we generated sample sizes of $n = 10$ and $n = 25$, as the nonparametric approaches are expected to perform similarly in large samples. Each of the four analysis approaches were applied to $R = 1,000$ replicated data sets under each scenario. The methods were compared based on a Monte Carlo approximation to the mean Bhattacharya distance (BCD) and Kullback-Leibler divergence (KLD) calculated as

$$\text{BCD} = \frac{1}{R} \sum_{r=1}^{R} \left(\sum_{j=\max(0,\min(y)-B)}^{\max(y)+B} - \log\left(\sqrt{p(j)\hat{p}_r(j)}\right) \right),$$

$$\text{KLD} = \frac{1}{R} \sum_{r=1}^{R} \left(\sum_{j=\max(0,\min(y)-B)}^{\max(y)+B} p(j) \log\left(p(j)/\hat{p}_r(j)\right) \right).$$

where we take the sums across the range of the observed data $\pm$ a buffer of 10.

In implementing the blocked Gibbs sampler for the rounded mixture of Gaussians, the first $1,000$ iterations were discarded as a burn-in and the next $10,000$ samples were used to calculate the posterior mean of $\hat{p}(h)$. For the hyperparameters, as a default empirical Bayes approach, we chose $\mu_0 = \overline{y}$, the sample mean, and $\kappa = s^2$, the sample variance, and $a_\tau = b_\tau = 1$. The precision parameter of the DP prior was set equal to one as a commonly used default and the truncation level N is set to be equal to the sample size of each sample. We also tried reasonable alternative choices of prior, such as placing a gamma hyperprior on the DP precision, for smaller numbers of simulations and obtained similar results. The values of $p(j)$ for a wide variety of js were monitored to gauge rates of apparent convergence and mixing. The trace plots showed excellent mixing, and the Geweke (1992) diagnostic suggested very rapid convergence.

The DP-Poisson approach used $\text{Poi}(\bar{y})$ as the base measure, with $\alpha = 1$ or $\alpha \sim \text{Ga}(1,1)$ considered as alternatives. For fixed α, the posterior is available in closed form, while for $\alpha \sim \text{Ga}(1,1)$ we implemented a Metropolis-Hastings normal random walk to update $\log \alpha$, with the algorithm run for $10,000$ iterations with a $1,000$ iterations burn-in.

Table 1: Bhattacharya coefficient and Kullback-Leibler divergence from the true distribution for samples from first scenario (a) and second scenario (b)

	(a)				(b)		
n	Method	BCD	KLD	n	Method	BCD	KLD
10	RMG	0.04	0.16	10	RMG	0.04	0.17
	E	0.24	∞		E	0.35	∞
	DP ($\alpha = 1$)	0.14	0.68		DP ($\alpha = 1$)	0.19	0.90
	DP ($\alpha \sim Ga(1,1)$)	0.11	0.47		DP ($\alpha \sim Ga(1,1)$)	0.11	0.49
	MLE	0.13	0.37		MLE	0.01	0.05
25	RMG	0.02	0.08	25	RMG	0.02	0.08
	E	0.09	∞		E	0.14	∞
	DP ($\alpha = 1$)	0.07	0.34		DP ($\alpha = 1$)	0.10	0.57
	DP ($\alpha \sim Ga(1,1)$)	0.06	0.24		DP ($\alpha \sim Ga(1,1)$)	0.06	0.29
	MLE	0.13	0.36		MLE	0.01	0.02

The results of the simulation are reported in Table 1. The proposed method performs better, in terms of BCD and KLD, than the other methods when the truth is underdispersed and clearly not Poisson, as in the first scenario. Moreover, even when the truth is Poisson, RMG is still competitive particularly for moderate sample size ($n = 25$) compared with DP and MLE. The ∞ recorded for the empirical estimation is due to the presence of $p(j)$ exactly equal to zero if we do not observe any $y = j$.

We also calculated the empirical coverage of 95% credible intervals for the $p(j)$s. These intervals were estimated as the 2.5th to 97.5th percentiles of the samples collected

after burn-in for each $p(j)$, with a small buffer of $\pm 1e - 08$ added to accommodate numerical approximation error. The plots in Figure 3 report the results with j on the x-axis. The effective coverage of the credible intervals for $p(j)$ for the RMG fluctuates around the nominal value for all the scenarios and the sample sizes. However using the Dirichlet process prior we get an effective coverage that is either strongly less than the nominal levels, or much too high, due to too wide credible intervals.

3. MULTIVARIATE ROUNDED KERNEL MIXTURE PRIORS

3.1 Multivariate counts

Multivariate count data are quite common in a broad class of disciplines, such as marketing, epidemiology and industrial statistics among others. Most multivariate methods for count data rely on multivariate Poisson models (Johnson et al. 1997) which have the unpleasant characteristic of not allowing negative correlation.

Mixtures of Poissons have been proposed to allow more flexibility in modeling multivariate counts (Meligkotsidou 2007). A common alternative strategy is to use a random effects model, which incorporates shared latent factors in Poisson log-linear models for each individual count (Moustaki and Knott 2000; Dunson 2000, 2003). A broad class of latent factor models for counts is considered by Wedel et al. (2003).

Copula models are an alternative approach to model the dependence among multivariate data. A p-variate copula $C(u_1, \ldots, u_p)$ is a p-variate distribution defined on the p-dimensional unit cube such that every marginal distribution is uniform on $[0, 1]$. Hence if F_j is the CDF of a univariate random variable Y_j, then $C(F_1(y_1), \ldots, F_p(y_p))$ is a p-variate distribution for $Y = (Y_1, \ldots, Y_p)$ with F_js as marginals. A specific copula model for multivariate counts is recently proposed by Nikoloulopoulos and Karlis (2010). Copula models are built for a general probability distribution and can hence be used to model jointly data of diverse type, including counts, binary data and continuous data. A very flexible copula model that considers variables having different measurement scales is proposed by Hoff (2007). This method is focused on modeling the association among variables with the marginals treated as a nuisance.
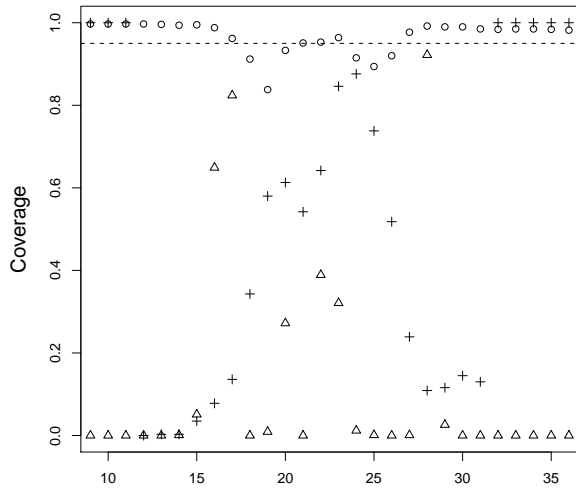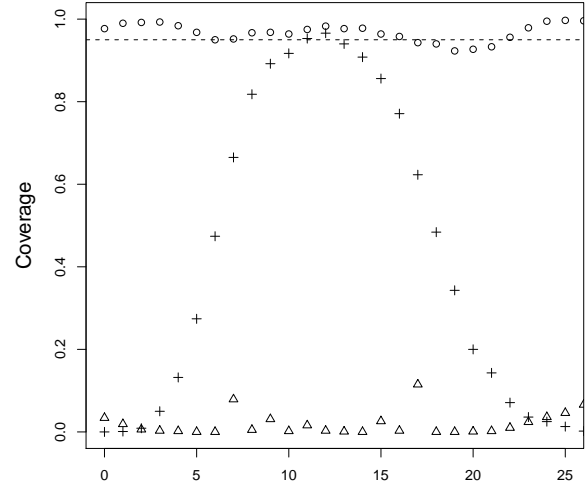
(a)
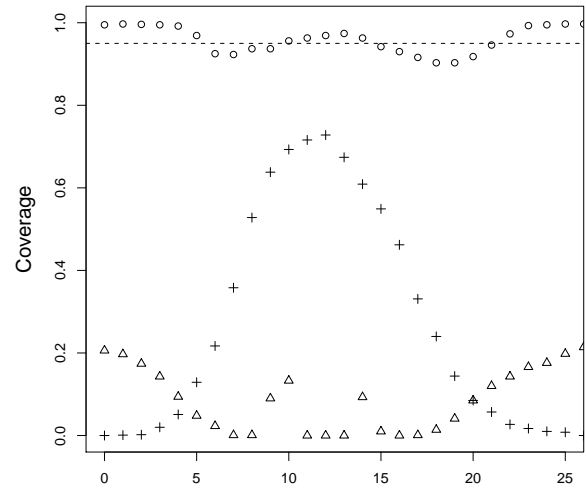
(b)
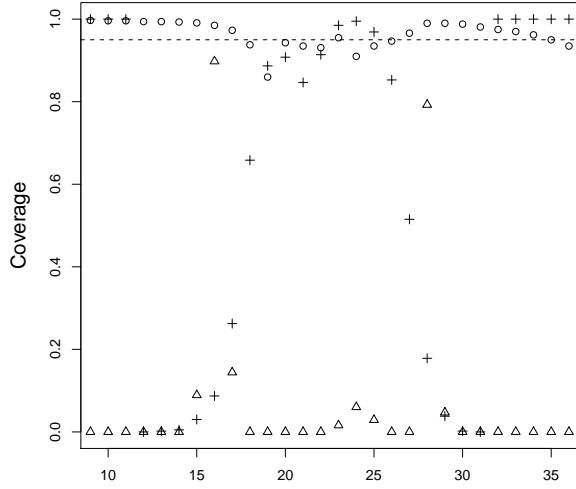
(c)

(d)

Figure 3: Coverage of 95% credible intervals for $p(j)$. Points represent the RMG method, crosses the DP with $\alpha = 1$ and triangles the DP with $\alpha \sim Ga(1,1)$. Panels (a) and (c) are for the samples of size $n = 10$ and $n = 25$ respectively, from the first scenario. Panels (b) and (d) are for samples of size $n = 10$ and $n = 25$ respectively from the second scenario.

We propose a multivariate rounded kernel mixture prior that can flexibly characterize the entire joint distribution including the marginals and dependence structure, while leading to straightforward and efficient computation. The use of underlying Gaussian mixtures easily allows the joint modeling of variables on different measurement scales including continuous variables, categorical and counts. In the past, it was hard to deal with counts jointly using such underlying Gaussian models unless one inappropriately treated counts as either categorical or continuous. In addition we can naturally do inference on the whole multivariate density, on the marginals or on conditional distributions of one variable given the others.

3.2 Multivariate rounded mixture of Gaussians

Each concept of Section 2 can be easily generalized into its multivariate counterpart. First assume that the multivariate count vector $\mathbf{y} = (y_1, \ldots, y_p)$ is the transformation through a threshold mapping function h of a latent continuous vector $\mathbf{y}^*$. In a general setting we have

$$\mathbf{y} = h(\mathbf{y}^*)$$
$$\mathbf{y}^* = (y_1^*, \ldots, y_p^*) \sim f(\mathbf{y}^*) = \int \mathrm{K}_p(\mathbf{y}^*; \theta, \Omega) d\mathrm{P}(\theta, \Omega),$$
$$P \sim \Pi^*, \tag{6}$$

where $\mathrm{K}_p(\cdot; \theta, \Omega)$ is a p-variate kernel with location θ and scale-association matrix Ω and Π^* is a prior for the mixing distribution. The mapping $h(\mathbf{y}^*) = \mathbf{y}$ implies that the probability mass function p of $\mathbf{y}$ is

$$p(y_1 = J_1, \ldots, y_p = J_p) = p(J) = g(f)[J] = \int_{A_J} f(\mathbf{y}^*) d\mathbf{y}^* \quad J \in \mathcal{N}^p \tag{7}$$

where $A_J = \{\mathbf{y}^* : a_{1,J_1} \le y_1^* < a_{1,J_1+1}, \ldots, a_{p,J_p} \le y_p^* < a_{p,J_p+1}\}$ defines a disjoint partition of the sample space. Marginally this formulation is the same of that in (3).

Remark 1. *Lemma 2 and Theorem 1 demonstrate that in the univariate case the mapping $g : \mathcal{L} \to \mathcal{C}$ maintains Kullback-Leibler neighborhoods and hence the induced prior Π assigns positive probability to all Kullback-Leibler neighbourhoods of any $p_0 \in \mathcal{C}$. This property holds also in the multivariate case.*

The true p_0 is in the Kullback-Leibler support of our prior, and hence we obtain weak and strong posterior consistency following the theory of Section 2, as long as there exists at least one multivariate density $f_0 = g^{-1}(p_0)$ that falls in the KL support of the mixture prior for f described in (6). In the sequel, we will assume that K_p corresponds to a multivariate Gaussian kernel and Π^* is $\text{DP}(\alpha P_0)$, with P_0 corresponding to a normal inverse-Wishart base measure. Wu & Ghosal (2008) showed that certain DP location mixtures of Gaussians support all densities f_0 satisfying a mild regularity condition. The size of the KL support of the DP location-scale mixture of Gaussians has not been formalized (to our knowledge), but it is certainly very large, suggesting informally that we will obtain posterior consistency at almost all p_0.

3.3 Out of sample prediction

Focusing on Dirichlet process mixtures of underlying Gaussians, we let the mixing distribution in (6) be $P \sim \text{DP}(\alpha P_0)$ with base measure

$$P_0 = \text{N}_p(\mu; \mu_0, \kappa_0 \Sigma)\text{Inv-W}(\Sigma; \nu_0, S_0). \tag{8}$$

To evaluate the performance, we simulated 100 data sets from the mixture

$$\mathbf{y}_i^* \sim \sum_{h=1}^{3} \pi_h N(\mu_h, \Sigma_h),$$

with $\pi = (\pi_1, \pi_2, \pi_3) = (0.14, 0.40, 0.46)$, $\mu_1 = (35, 82, 95)$, $\mu_2 = (-2, 1, 2.5)$, $\mu_3 = (12, 29, 37)$ and variance-covariance matrices

$$\Sigma_1 = \begin{pmatrix} 3 & -0.6 & 0.25 \\ -0.6 & 3 & 0.7 \\ 0.25 & 0.7 & 2 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0.5 & 0.4 \\ 0.5 & 1 & -0.4 \\ 0.4 & -0.4 & 0.7 \end{pmatrix}, \quad \Sigma_3 = 7.5 \cdot \Sigma_2,$$

with the continuous observation floored and all negative values set equal to zero leading to a multivariate zero-inflated count distribution. The samples were split into training and test subsets containing 50 observations each, with the Gibbs sampler applied to the training data and the results used to predict y_{i1} given y_{i2} and y_{i3} in the test sample. This approach modifies Müller et al. (1996) to accomodate count data.

The hyperparameters were specified as follows:

$$\mu_0 \sim N_3(\bar{\mathbf{y}}^*, \hat{S}), \ S_0 \sim \text{InvWishart}(4, \Psi_0),$$

$$\Psi_0 = I_3, \ \nu_0 = 4, \ \kappa_0 \sim \text{Gamma}(0.01, 0.01), \ \alpha = 1 \tag{9}$$

with $\bar{y}^* = (1 - \hat{p}_0)\bar{y}_+ - \bar{p}_0\bar{y}_+$, $\hat{p}_0$ the proportion of zeros in the training sample, $\bar{y}_+$ the mean of the non-zero values and $\hat{S} = \text{diag}(s_1^2, s_2^2, s_3^3)$ with s_j the empirical variance of y_{ij}, $i = 1, \ldots, n$. The Gibbs sampler reported in the Appendix was run for $10,000$ iterations with the first $4,000$ discarded. We assessed predictive performance using the absolute deviation loss, which is more natural than squared error loss for count data. Under absolute deviation loss, the optimal predictive value for y_{i1} corresponded to the median of the posterior predictive distribution.

We compare our approach with prediction under an oracle based on the true model, Poisson log-linear regression fit with maximum likelihood and a generalized additive model (GAM) (Hastie et al. 2001) with spline smoothing function. Our proposed approach provided more accurate predictions than the other methods with the exception of the oracle. The mean absolute deviation errors (MAD) are 2.32, 1.4, 2.72 and 5.35 for our method, the oracle, the GAM and the Poisson regression respectively.

An additional gain of our approach is a flexible characterization of the whole predictive distribution of y_{i1} given y_{i2}, y_{i3} and not just the point prediction $\hat{y}_{i1}$. In addition to median predictions, it is often of interest in applications to predict subjects having zero counts or counts higher than a given threshold q. Based on our results, we obtained much more accurate predictions of both $y_{i1} = 0$ and $y_{i1} > q$ than either the log-linear Poisson model or the GAM approach. As an additional competitor for predicting $y_{i1} = 0$ and $y_{i1} > q$, we also considered logistic regression and logistic GAM fitted to the appropriate dichotomized data. Based on a 0-1 loss function that classified $y_{i1} = 0$ if the probability (posterior for our Bayes method and fitted estimate for the logistic GAM) exceeded 0.5, we compute the misclassification rate out-of-sample in Table 2.

18

Table 2: Misclassification rate out-of-sample based on the proposed method, generalized linear regressions, GAM and the oracle.

	RMG		GAM		GLM		Oracle
	Median^a	0-1 Loss^b	Poisson	Logistic	Poisson	Logistic	-
$y_{i1} = 0$	0.02	0.08	0.42	0.14	0.42	0.20	0.00
$y_{i1} > 20$	0.02	0.10	0.02	0.40	0.08	0.30	0.00
$y_{i1} > 25$	0.04	0.02	0.04	0.06	0.06	0.08	0.02
$y_{i1} > 35$	0.06	0.06	0.06	0.08	0.06	0.14	0.06

a = prediction based on posterior median, b = prediction based on 0-1 loss

4. APPLICATION TO MARKETING DATA

4.1 Data and motivation

Telecommunication companies every day store plenty of information about their customer behaviour and services usage. Mobile operators, for example, can store the daily usage stream such as the duration of the calls or the number of text and multimedia messages sent. Companies are often interested in profiling both customers with high usage and customers with very low usage. Suppose that at each activation a customer is asked to simply state how many text messages (SMS), multimedia messages (MMS) and calls they anticipate making on average in a month and the company wants to predict the future usage of each new customer.

We focus on data from $2,050$ SIM cards from customers having a prepayed contract, with a multivariate $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$ available representing usage in a month for card i. Specifically, we have the number of outgoing calls to fixed numbers (y_{i1}), to mobile numbers of competing operators (y_{i2}) and to mobile numbers of the same operator (y_{i3}), as well as the total number of MMS (y_{i4}) and SMS (y_{i5}) sent. Jointly modeling the probability distribution $f(\cdot)$ of the multivariate $\mathbf{y}$ using a Bayesian mixture and assuming an underlying continuous variable for the counts, we focus on the forecast of y_{i1}, using data on $y_{i2}, \ldots, y_{i5}$. Some descriptive statistics of the dataset show the presence of a lot of zeros for our response variable y_1. Such zero-inflation

is automatically accommodated by our method through using thresholds that assign negative underlying y_{ij}^* values to $y_{ij} = 0$ as described in Section 2.3. Excess mass at zero is induced through Gaussian kernels located at negative values.

4.2 Model estimation and results

We can model the data assuming the model in (6) with hyperparameters specified as in (9) and computation implemented as in Section 3.3. A training and test set of equal size are chosen randomly. Trace plots of y_{i1} for different individuals exhibit excellent rates of convergence and mixing, with the Geweke (1992) diagnostic providing no evidence of lack of convergence.

Our method is compared with Poisson and logistic GAMs as in Section 3.3. The out-of-sample median absolute deviation (MAD) value was 8.08 for our method, which is lower than the 8.76 obtained for the best competing method (Poisson GAM). These results were similar for multiple randomly chosen training-test splits. Suppose the interest is in predicting customers with no outgoing calls and highly profitable customers. We predict such customers using Bayes optimal prediction under a 0-1 loss function. Using optimal prediction of zero-traffic customers, we obtained lower out-of-sample misclassification rates than the Poisson GAM, but had comparable results to logistic GAM as illustrated in the ROC curve in Figure 5 (a). Our expectation is that the logistic GAM will have good performance when the proportion of individuals in the subgroup of interest is $\approx 50\%$, but will degrade relative to our approach as the proportion gets closer to 0% or 100%. In this application, the proportion of zeros was 69% and the sample size was not small, so logistic GAM did well. The results for predicting highly profitable customers having more than 40 calls per month are consistent with this, as illustrated in Figure 5 (b). It is clear that our approach had dramatically better predictive performance.

5.   DISCUSSION

The usual parametric models for count data lack flexibility in several key ways, and nonparametric alternatives have clear disadvantages. Our proposed class of Bayesian

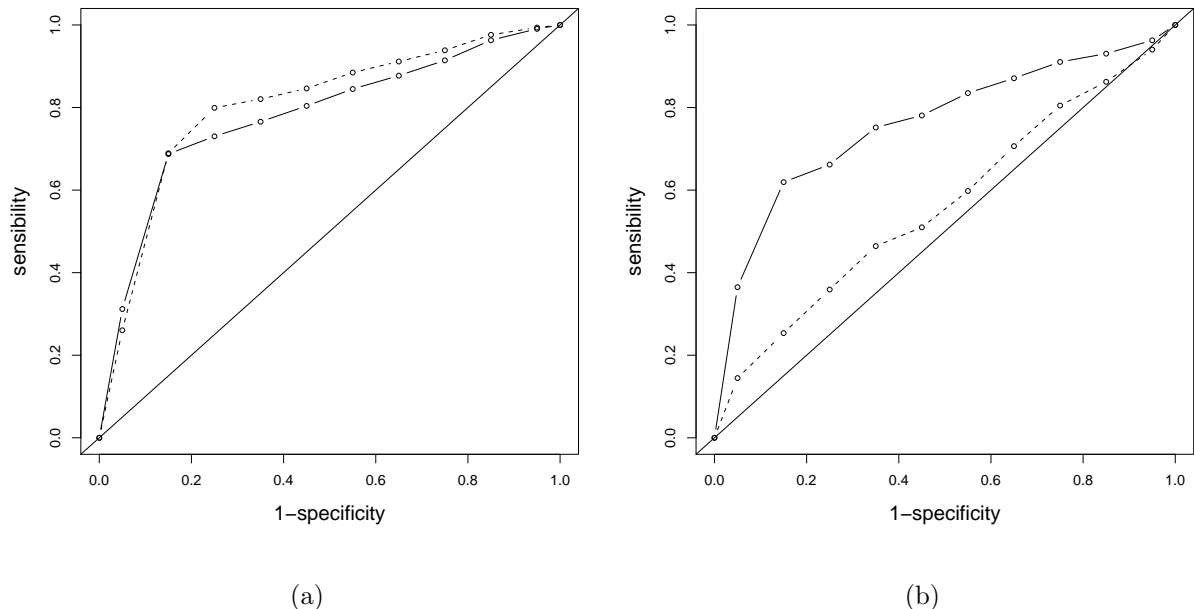|	
(a)	(b)

Figure 4: ROC curves for predicting customers having outgoing calls to fixed numbers equal to zero (a) or more than 40 (b). The continuous line is for our proposed approach and the dotted lines are for the logistic GAM. Both classifications are based on a 0-1 loss function that classify $y_{i1} = 0$ or $y_{i1} > 40$ if the posterior (estimated) probability is greater than 1/2.

nonparametric mixtures of rounded continuous kernels provide a useful new approach that can be easily implemented in a broad variety of applications. We have demonstrated some practically appealing properties including simplicity of the formulation, ease of computation through trivial modifications of MCMC algorithms for continuous densities, large Kullback-Leibler support, weak and strong posterior consistency, and straightforward joint modeling of counts, categorical and continuous variables from which is it possible to infer conditional distributions of response variables given predictors as well as marginal and joint distributions. The proposed class of conditional distribution models allows a count response distribution to change flexibly with multiple categorical, count and continuous predictors.

Our approach has been applied to a marketing application using a DP mixture of multivariate rounded Gaussians. The use of an underlying Gaussian formulation is

quite appealing in allowing straightforward generalizations in several interesting directions. For example, for high-dimensional data instead of using an unstructured mixture of underlying Gaussians, we could consider a mixture of factor analyzers (Gorur and Rasmussen 2009). An alternative is to mix generalized latent trait models, which induce dependence through incorporating shared latent variables in generalized linear models for each response type. However, this strategy would rely on mixtures of Poisson log-linear models for count data, which restrict the marginals to be over-dispersed and can lead to a restrictive dependence structure. It also becomes straightforward to accommodate time series and spatial dependence structures through mixtures of Gaussian dynamic or spatially dependent models.

ACKNOWLEDGMENTS

APPENDIX

Proof of Lemma 1. The lemma is trivially proved by defining f_0 as a step function of the form

$$f_0(x) = \frac{p_0(0)}{a_1 - b} \mathbb{I}_{[b,a_1)}(x) + \sum_{h=1}^{\infty} \frac{p_0(h)}{a_{h+1} - a_h} \mathbb{I}_{[a_h, a_{h+1})}(x),$$

where $\mathbb{I}_A(x)$ is 1 iff $x \in A$ and b is an arbitrary number such that (b, a_1) is in the domain of f. $\qquad\square$

Proof of Lemma 2. Let f a general element of $\mathcal{K}_\epsilon(f_0)$ and denote $p = g(f)$ its image on $\mathcal{C}$, hence

$$KL(f_0, f) = \int_{a_0}^{a_\infty} f_0(x) \log\left(\frac{f_0(x)}{f(x)}\right) dx < \epsilon. \tag{10}$$

If we discretizise the integral (10) in the infinite sum of integrals on disjoint subset of the domain of f we have

$$\sum_{h=0}^{\infty} \int_{a_h}^{a_{h+1}} f_0(t) \log\left(\frac{f_0(t)}{f(t)}\right) dt < \epsilon.$$

22

Using the condition (see Theorem 1.1 of Ghurye (1968))

$$\int_A g_1(t)\mathrm{d}t \times \log\left(\frac{\int_A g_1(t)\mathrm{d}t}{\int_A g_2(t)\mathrm{d}t}\right) \leq \int_A g_1(t)\log\left(\frac{g_1(t)}{g_2(t)}\right)\mathrm{d}t$$

for each $A \in \mathcal{A}$, countable family of disjoint measurable sets of $\mathcal{Y}$ and $g_1, g_2 \in \mathcal{L}$, we get

$$p_0(j)\log\frac{p_0(j)}{p(j)} \leq \int_{a_j}^{a_{j+1}} f_0(t)\log\left(\frac{f_0(t)}{f(t)}\right)\mathrm{d}t$$

and hence

$$\sum_{j=0}^{\infty} p_0(j)\log\frac{p_0(j)}{p(j)} \leq \int_{a_0}^{a_\infty} f_0(x)\log\left(\frac{f_0(x)}{f(x)}\right)\mathrm{d}x < \epsilon,$$

that gives the result. $\qquad\square$

Proof of Theorem 1. For every $f \in \mathcal{L}_{\Pi^*}$ by Lemma 2 we have

$$\Pi(\mathcal{K}_\epsilon(p)) \geq \Pi(g(\mathcal{K}_\epsilon(f))) = \Pi^*(\mathcal{K}_\epsilon(f)) > 0.$$

$\qquad\square$

Proof of Theorem 2. Weak convergence of $p^n \to p_0$ is expressed as

$$\int \phi(y)p^n(y)dy \to \int \phi(y)p_0(y)dy$$

for all bounded and continuous functions ϕ.

Now, since $p^n(y)$ and p_0 are discrete with masses on the non negative integers, we obtain

$$\sum_{y=0}^{\infty} \phi(y)p^n(y) \to \sum_{y=0}^{\infty} \phi(y)p_0(y).$$

Since all functions on the non-negative integers are continuous, take $\phi \in \{\mathbb{1}_j(x) : j = 0, 1, \ldots\}$ and we obtain the result. $\qquad\square$

Proof of Lemma 2. Strong convergence is implied by the Shur's property that holds in the space of absolutely summable sequences. $\qquad\square$

Algebraic details to center Rounded mixture of Gaussians prior

Assuming a Dirichlet process mixture of Gaussians kernels we have

$$
\begin{aligned}
E\{F(a_j)\} &= E\{\Phi(a_j; \mu, \tau^{-1})\} \\
&= \int_{R \times R^+} \Phi(a_j; \mu, \tau^{-1}) N(\mu; \mu_0, \tau^{-1}\kappa) Ga(\tau; \nu/2, \nu/2) \, d\mu \, d\tau \\
&= \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^{a_j} N(z; \mu, \tau^{-1}) N(\mu; \mu_0, \tau^{-1}\kappa) Ga(\tau; \nu/2, \nu/2) \, d\mu \, d\tau \, dz. \quad (11)
\end{aligned}
$$

Marginalizing out μ from (11) we get

$$
E\{F(a_j)\} = \int_{-\infty}^{a_j} \int_0^\infty N(z; \mu_0, (\kappa+1)/\tau) Ga(\tau; \nu/2, \nu/2) \, d\tau \, dz.
$$

while marginalizing out τ we obtain

$$
E\{F(a_j)\} = \int_{-\infty}^{a_j} t_\nu(z; \mu_0, \kappa+1).
$$

Similarly we get the expression for the variance. The second moment of $F_0(y)$ is

$$
\begin{aligned}
E\{(F(y))^2\} &= E\left\{\left(\sum_{h=1}^N \pi_h \Phi(y; \mu_h, \tau_h^{-1})\right)^2\right\} \\
&= \sum_{h=1}^N E\left\{(\pi_h \Phi(y; \mu_h, \tau_h^{-1}))^2\right\} + 2\sum_{k \neq l} E\left\{\pi_k \pi_l \Phi(y; \mu_k, \tau_k^{-1})\Phi(y; \mu_l, \tau_l^{-1})\right\}.
\end{aligned}
$$

$$
(12)
$$

For the general h we have

$$
\begin{aligned}
E\left\{(\pi_h \Phi(y; \mu_h, \tau_h^{-1}))^2\right\} &= \int_0^1 \int_0^\infty \int_{-\infty}^\infty (\pi_h \Phi(a_j; \mu_h, \tau_h^{-1}))^2 N(\mu_h; \mu_0, \tau_h^{-1}\kappa) Ga(\tau_h; \nu/2, \nu/2) \times \\
&\quad \times \mathrm{Dir}(\pi_h; \alpha) \, d\mu_h \, d\tau_h \, d\pi_h \\
&= \int_0^\infty \int_{-\infty}^\infty (\Phi(a_j; \mu_h, \tau_h^{-1}))^2 N(\mu_h; \mu_0, \tau_h^{-1}\kappa) Ga(\tau_h; \nu/2, \nu/2) \, d\mu_h \, d\tau_h \times \\
&\quad \times \int_0^1 \pi_h^2 \mathrm{Dir}(\pi_h; \alpha) \, d\pi_h \\
&= E\left\{(\Phi(y; \mu_h, \tau_h^{-1}))^2\right\} E\{\pi_h^2\}.
\end{aligned}
$$

and similarly for the general $k \neq l$

$$
E\left\{\pi_k \pi_l \Phi(y; \mu_k, \tau_k^{-1})\Phi(y; \mu_l, \tau_l^{-1})\right\} = E\left\{\Phi(y; \mu_k, \tau_k^{-1})\Phi(y; \mu_l, \tau_l^{-1})\right\} E\{\pi_k \pi_l\}. \quad (13)
$$

Using the results on the variance and covariance of the Dirichlet distribution

$$E\left\{\pi_h^2\right\} = \frac{N+\alpha}{(\alpha+1)N^2}, \quad E\left\{\pi_h\pi_k\right\} = \frac{\alpha}{(\alpha+1)N^2}.$$

Since the sums of (12) are the sums of respectively N and $1/2N(N-1)$ equal elements,

$$\begin{aligned}
E\{(F(y))^2\} &= \frac{N+\alpha}{(\alpha+1)N}E\left\{\left(\Phi(y;\mu,\tau^{-1})\right)^2\right\} + \frac{(N-1)\alpha}{(\alpha+1)N}E\left\{\left(\Phi(y;\mu,\tau^{-1})\right)^2\right\} \\
&= E\left\{\left(\Phi(y;\mu,\tau^{-1})\right)^2\right\},
\end{aligned}$$

and hence

$$Var\{p(j)\} = \sum_{l=j,j+1} q(l)^2 + E\left\{\left(\Phi(a_l;\mu,\tau^{-1})\right)^2\right\}.$$

Multivariate Gibbs Sampler

For the multivariate rounded mixture of Gaussians we adopt the Gibbs sampler with auxiliary parameters of Neal (2000), and more precisely the Algorithm 8 with $m = 1$. The sampler iterates among the following steps:

Step 1 Generate each $\mathbf{y}_i^*$ from the full conditional posterior

   for j in $1,\ldots,p$

Step 1a Generate $u_{ij} \sim U\left(\Phi(a_{y_{ij}-1};\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}^2), \Phi(a_{y_{ij}};\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}^2)\right)$, where

$$\begin{aligned}
\tilde{\mu}_{i,j} &= \mu_{S_i,j} + \Sigma_{S_i,12}\Sigma_{S_i,22}^{-1}(\mathbf{y}_{-j}^* - \mu_{S_i,-j}) \\
\tilde{\sigma}_{i,j}^2 &= \sigma_{S_i,j}^2 - \Sigma_{S_i,12}\Sigma_{S_i,22}^{-1}\Sigma_{S_i,21}
\end{aligned}$$

are the usual conditional expectation and conditional variance of the multivariate normal.

Step 1b Let $y_{ij}^* = \Phi^{-1}(u_i;\tilde{\mu}_{i,j},\tilde{\sigma}_{i,j}^2)$

Step 2 Update S_i as in Algorithm 8 of Neal (2000) with $m = 1$.

Step 3 Update (μ_h,Σ_h) from their conditional posteriors.

REFERENCES

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Carota, C. and Parmigiani, G. (2002), "Semiparametric Regression for Count Data," *Biometrika*, 89, 265–281.

Chen, J., Zhang, D., and Davidian, M. (2002), "A Monte Carlo EM Algorithm for Generalized Linear Mixed Models with Flexible Random Effects Distribution," *Biostatistics (Oxford)*, 3, 347–360.

Dunson, D. B. (2000), "Bayesian Latent Variable Models for Clustered Mixed Outcomes," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 62, 355–366.

— (2003), "Dynamic Latent Trait Models for Multidimensional Longitudinal Data," *Journal of the American Statistical Association*, 98, 555–563.

— (2005), "Bayesian Semiparametric Isotonic Regression for Count Data," *Journal of the American Statistical Association*, 100, 618–627.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230.

— (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615–629.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments," in *Bayesian Statistics 4*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Oxford: Oxford University Press.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), "Posterior Consistency of Dirichlet Mixtures in Density Estimation," *The Annals of Statistics*, 27, 143–158.

Ghurye, S. G. (1968), "Information and sufficient sub-fields," *The Annals of Mathematical Statistics*, 39, 2056–2066.

Gill, J. and Casella, G. (2009), "Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation," *Journal of the American Statistical Association*, 104, 453–454.

Gorur, D. and Rasmussen, C. E. (2009), "Nonparametric mixtures of factor analyzers," in *17th Annual IEEE Signal Processing and Communications Applications Conference*, vol. 1-2, pp. 922–925.

Guha, S. (2008), "Posterior Simulation in the Generalized Linear Mixed Model With Semiparametric Random Effects," *Journal of Computational and Graphical Statistics*, 17, 410–425.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag Inc.

Hoff, P. D. (2007), "Extending the rank likelihood for semiparametric copula estimation," *Ann. Appl. Statist.*, 1, 265–283.

Ishwaran, H. and James, Lancelot, F. (2001), "Gibbs Sampling Methods for Stick Breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Jara, A., Garcia-Zattera, M., and Lesaffre, E. (2007), "A Dirichlet process mixture model for the analysis of correlated binary responses," *Computational Statistics & Data Analysis*, 51, 5402–5415.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, New York: John Wiley & Sons.

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

Kleinman, K. P. and Ibrahim, J. G. (1998), "A semi-parametric Bayesian approach to generalized linear mixed models," *Statistics in Medicine*, 30, 2579–2596.

Kottas, A., Müller, P., and Quintana, F. (2005), "Nonparametric Bayesian Modeling for Multivariate Ordinal Data," *Journal of Computational and Graphical Statistics*, 14, 610–625.

Krnjajic, M., Kottas, A., and Draper, D. (2008), "Parametric and nonparametric Bayesian model specification: A case study involving models for count data," *Computational Statistics & Data Analysis*, 52, 2110 – 2128.

Lo, A. Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12, 351–357.

Meligkotsidou, L. (2007), "Bayesian Multivariate Poisson Mixtures with an Unknown Number of Components," *Statistics and Computing*, 17, 93–107.

Moustaki, I. and Knott, M. (2000), "Generalized Latent Trait Models," *Psychometrika*, 65, 391–411.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67–79.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Nikoloulopoulos, A. K. and Karlis, D. (2010), "MOdeling multivariate count data using copulas," *Communications in Statistics, Simulation and Computation*, 393, 172–187.

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985), "The developemental toxicity of ethylene glycol in rats and mice," *Toxicological and Applied Pharmacology*, 81, 113–127.

Schwartz, L. (1965), "On Bayes Procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4, 10–26.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Walker, S. G. (2007), "Sampling the Dirichlet Mixture Model with Slices," *Communications in Statistics, Simulation and Computation*, 34, 45–54.

Wedel, M., Böckenholt, U., and Kamakura, W. A. (2003), "Factor Models for Multivariate Count Data," *Journal of Multivariate Analysis*, 87, 356–369.

Wu, Y. and Ghosal, S. (2008), "Kullback Leibler Property of Kernel Mixture Priors in Bayesian Density Estimation," *Electronic Journal of Statistics*, 2, 298–331.

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2010), "Bayesian non parametric Hidden Markov Models with applications in genomics," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, to appear.