Bayesian Analysis (2019) **14**, Number 3, pp. 677–702

Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models

Yushu Shi*, Michael Martens†, Anjishnu Banerjee‡, and Purushottam Laud§

Abstract. Dirichlet process mixture (DPM) models provide flexible modeling for distributions of data as an infinite mixture of distributions from a chosen collection. Specifying priors for these models in individual data contexts can be challenging. In this paper, we introduce a scheme which requires the investigator to specify only simple scaling information. This is used to transform the data to a fixed scale on which a low information prior is constructed. Samples from the posterior with the rescaled data are transformed back for inference on the original scale. The low information prior is selected to provide a wide variety of components for the DPM to generate flexible distributions for the data on the fixed scale. The method can be applied to all DPM models with kernel functions closed under a suitable scaling transformation. Construction of the low information prior, however, is kernel dependent. Using DPM-of-Gaussians and DPM-of-Weibulls models as examples, we show that the method provides accurate estimates of a diverse collection of distributions that includes skewed, multimodal, and highly dispersed members. With the recommended priors, repeated data simulations show performance comparable to that of standard empirical estimates. Finally, we show weak convergence of posteriors with the proposed priors for both kernels considered.

Keywords: Bayesian nonparametric methods, density estimation, survival analysis, low-information prior, Dirichlet process mixture model.

1 Introduction

The Dirichlet process mixture (DPM) model was first proposed by Lo (1984). The marginal distribution of a DPM is a convolution of a kernel density function and a Dirichlet process, $g(y) = \int f(y|G)DP(dG)$. This model uses the Dirichlet process (DP) of Ferguson (1973) effectively to estimate density functions even though the DP almost surely generates discrete distributions. The DPM model can be written also as:

$$\begin{aligned} y_i|\theta_i &\sim f(\cdot|\theta_i) \\ \theta_i|G &\sim G \\ G|G_0, \nu &\sim DP(G_0, \nu). \end{aligned}$$

Here each observation y_i arises from a density function $f(\cdot|\theta_i)$ with corresponding parameter θ_i, which in turn arises from a discrete distribution G. The distribution G is

*University of Texas M. D. Anderson Cancer Center, yshi7@mdanderson.org
†The Emmes Corporation, mmartens@emmes.com
‡Division of Biostatistics, Medical College of Wisconsin, abanerjee@mcw.edu
§Division of Biostatistics, Medical College of Wisconsin, laud@mcw.edu

 https://doi.org/10.1214/18-BA1119

randomly generated from a DP with baseline distribution G_0 and concentration parameter ν. The choice of kernel density $f(\cdot|\theta)$ determines the mixture components to use in a DPM; for example, if $f(\cdot|\theta)$ is a normal kernel, then this DPM is a mixture of Gaussians.

The Gaussian kernel was employed and computationally implemented by Escobar and West (1995). Kottas (2006) considered a mixture of Weibulls model for positive valued survival data. In contrast with much development of the DPM model itself in various directions, the prior specification for it is often undertaken in an ad-hoc fashion with little formal guidance available in the literature. The method proposed here attempts to address this gap in cases where prior information is scant or intentionally avoided in the analysis.

Using as input some simple scaling information for the data to be analyzed, we transform the data to a common axis on which we construct the prior. This transformation and axis depends on the kernel chosen for the DPM, as does the method of construction of the prior. Once constructed, the LIO prior is fully specified and can be used a black-box for the DPM with this kernel. Inference on the original data scale is recovered by a back transformation of the posterior samples. In the case of the Gaussian DPM, we do this construction for univariate as well as multivariate data.

The paper is organized as follows. Section 2 introduces general guiding objectives in constructing low information priors. Sections 3 and 4 apply these notions to the construction of particular prior specifications for Gaussian and Weibull DPMs, illustrating the priors' use through implementation on real and simulated data sets. Section 5 conducts sensitivity analysis and compares repeated data simulation results from the Gaussian and Weibull DPMs using the proposed priors with those from empirical methods. The comparison is intended primarily to demonstrate the low information nature of the prior, not to claim performance superiority. Any advantages over empirical methods accrue from well-recognized aspects of Bayesian nonparametric models, such as distributional flexibility and easy estimation of functionals of the underlying distribution (e.g., density and hazard rate) along with attendant uncertainty quantification for each. Finally, Section 6 establishes posterior weak convergence properties with the priors, while Section 7 concludes the paper with a brief discussion.

2 Rationale and Construction Outline for Low-information Omnibus (LIO) Priors

When applying a DPM model to data, the base distribution G_0 should be specified with care, as G_0 represents prior knowledge about the distribution of the data in an intricate combination with ν. One's first instinct might be to use a vague G_0. However, it is well known that this is not advisable. For example, the authors of Chapter 23 of Gelman et al. (2014) point out that using such a choice of G_0 places *"a heavy penalty on the introduction of new clusters"*. In effect, a highly dispersed choice of G_0 is highly informative, as it implies that all data points belong to a common cluster in the posterior predictive distribution. They recommend standardizing the data and using *"an*

informative G_0 that places high probability on introducing clusters near the support of the data". Similar uses of data scaling and low information prior can be seen in parametric Bayesian data analysis. Gelman et al. (2008) suggested specific scaling and a low information prior that is *"vague enough to be used as a default in routine applied work"* instead of aiming for a no-information prior. The latter pursuit can be challenging both theoretically and computationally.

With this rationale, we propose a specific data-scaling that depends on the DPM kernel and a particular hierarchical specification of the prior on G_0 for the scaled observations, which jointly serve as a "black box" for various data contexts. The prior elicitation requires minimal scale-related information (such as a high percentile of the population distribution for the mixture-of-Weibulls model; and the median and 95th percentile for the mixture-of-Gaussians model) from the investigator knowledgeable in the subject matter. While fitting a particular DPM with an already constructed LIO prior might be seen as a black box method, deriving the prior for the kernel chosen requires much care, and is certainly not a black box. For some kernels, even a trial and error process with visual inspection might be needed (as for the Weibull kernel below) but only once per kernel.

In using the prior, there are three simple steps:

1. With the scaling information provided by the investigator, transform the data to a suitable fixed scale.
2. Apply the recommended LIO prior to the fixed-scale data and obtain posterior samples using established computational methods. The prior specification is aimed at providing a variety of mixture components rich enough to allow flexible modeling of observations on the fixed scale. We thus find a set of hyperparameters capable of generating such components. The process of finding these hyperparameters is discussed for two specific DPMs in the sequel.
3. Transform back the samples representing posterior inference to obtain originally targeted inference.

Currently, we have two such distinct black-box implementations: one is for the mixture-of-Gaussians model that is well-suited to modeling real-valued and vector-valued data. The other is for the mixture-of-Weibulls model, which is more appropriate for time-to-event data as the Weibull distribution has a positive domain and convenient mathematical forms for interpretable functions such as the survival and hazard functions.

When considering the kernel density components needed for fixed-scale data, we keep a modest goal in sight: give a reasonable and rich variety of components a fair chance to be selected by the data. The DPM model itself is robust in that the information in the data will be dominant when the prior is sufficiently flexible. Specifics of prior construction are given in the next two sections. In implementing inference with the proposed priors, for all computational results reported here, we used the 8th algorithm of Neal (2000).

3 LIO Prior for DPM of Gaussian Distributions

A Dirichlet process mixture of Gaussian distributions is versatile in that it applies straightforwardly to univariate as well as multivariate data. Below, after establishing notation, we develop LIO prior specifications; first for univariate and then for multivariate data.

3.1 Model Specification

We use a Gaussian DPM model similar to that employed by the DPdensity function in the R package DPpackage (Jara et al., 2011). Assume $\mathbf{y_1}, \cdots, \mathbf{y_n}$ are conditionally iid vector observations, each of length p. Our approach is to make a location-scale transformation of the data, apply the DPM model to estimate the transformed data's distribution, and then estimate the original data's distribution by transforming back to the original scale. More specifically, we choose some quantities $\mathbf{a} \in \mathbb{R}^p$ and a positive definite $p \times p$ matrix $\mathbf{B}$ to rescale the data as $\mathbf{z}_i = \mathbf{B^{-1}}(\mathbf{y_i} - \mathbf{a})$. Then, the following model is fitted to the transformed data:

$$\begin{aligned}
\mathbf{z}_i|\boldsymbol{\mu}_i, \mathbf{T}_i &\overset{ind}{\sim} No(\boldsymbol{\mu}_i, \mathbf{T}_i), \\
(\boldsymbol{\mu}_i, \mathbf{T}_i)|G &\overset{iid}{\sim} G, \\
G|G_0 &\sim DP(G_0, \nu), \\
\nu &\sim Ga(a, b), \\
G_0|\lambda, \boldsymbol{\Psi} &= NoWi(\mathbf{m}_\mu, \lambda, k_T, \boldsymbol{\Psi}), \\
\lambda &\sim Ga(a_\lambda, b_\lambda), \\
\boldsymbol{\Psi} &\sim Wi(k_\psi, \mathbf{W}_\psi).
\end{aligned}$$

Here $No(\mathbf{m}, \mathbf{U})$ denotes a normal distribution with mean $\mathbf{m}$ and precision matrix $\mathbf{U}$, and $Ga(a, b)$ denotes a Gamma distribution with shape parameter a and rate parameter b. With $Wi(k, \mathbf{W})$ denoting a Wishart distribution with degrees of freedom k and inverse scale matrix $\mathbf{W}$ (expectation $k\mathbf{W}^{-1}$), G_0 has a hierarchical specification, the first level being a normal-Wishart distribution with parameters $\mathbf{m}_\mu$, λ, k_T, and $\boldsymbol{\Psi}$ and the second level having independent Gamma and Wishart distributions for λ and $\boldsymbol{\Psi}$, respectively. To be specific, $(\boldsymbol{\mu}, \mathbf{T}) \sim NoWi(\mathbf{m}, \lambda, k, \boldsymbol{\Psi})$ means $\boldsymbol{\mu}|\mathbf{T}, \lambda \sim No(\mathbf{m}, \lambda\mathbf{T})$ and $\mathbf{T}|\boldsymbol{\Psi}, k \sim Wi(k, \boldsymbol{\Psi})$. Because the support of the Wishart distribution is the set of $p \times p$ positive definite matrices, all $\mathbf{T}_i$ obtained from this model are positive definite. The concentration parameter ν is set to have a $Ga(a, b)$ prior with $a = 1$ and $b = 1$ (Escobar and West, 1995).

As the $\mathbf{z}_i$ arise from an infinite mixture of normal distributions, their cumulative distribution function (CDF) is

$$F_z(\mathbf{z}) = \sum_{i=1}^{\infty} p_i F_{z_i|\mu_i,\tau_i}(\mathbf{z}) = \sum_{i=1}^{\infty} p_i \Phi_p\big[\boldsymbol{\Lambda_i}(\mathbf{z} - \boldsymbol{\mu_i})\big], \qquad \mathbf{z} \in \mathbb{R}^p,$$

where Φ_p is the CDF of a p-variate normal distribution $No(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\Lambda}_i$ resulting from the unique Cholesky decomposition $\mathbf{T}_i = \boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_i'$, and $\sum_{i=1}^{\infty} p_i = 1$. By the correspondence

between the $\mathbf{y}_i$ and $\mathbf{z}_i$, this implies that the original data's distribution is an infinite mixture of normal distributions with CDF

$$F_y(\mathbf{y}) = F_z[\mathbf{B^{-1}}(\mathbf{y} - \mathbf{a})] = \sum_{i=1}^{\infty} p_i \Phi_p\big\{\mathbf{\Lambda_i B^{-1}}[\mathbf{y} - (\mathbf{B}\boldsymbol{\mu}_\mathbf{i} + \mathbf{a})]\big\}, \qquad \mathbf{y} \in \mathbb{R}^p.$$

Thus, fitting this model to the transformed data induces a DPM model on the original data and provides an estimate of its CDF. Through this, one can estimate any functionals of the distribution of the original data through posterior sampling of $\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \mathbf{T}_i)$. We want to transform the data to a common location and dispersion; this will justify applying the same model to all transformed data sets. In our location-scale transformation, $\mathbf{a}$ and $\mathbf{B}$ are measures of the location and scale of the original data that need specification. We employ contextual choices of some quantiles of the data's underlying distribution. The investigator supplies values c_k and d_k that are reasonable pre-data estimates of the median and the 95^{th} percentile of each component y_{1k} of the data vector. These percentiles are natural quantities to consider and should facilitate elicitation based on existing results or expert opinion. The scale (or standard deviation) of the k^{th} component can be estimated roughly by $(d_k - c_k)/2$, so we set $\mathbf{a} = \mathbf{c}$ and $\mathbf{B} = \mathbf{Diag}\{(\mathbf{d} - \mathbf{c})/2\}$. The transformation $\mathbf{z}_i = \mathbf{B^{-1}}(\mathbf{y_i} - \mathbf{a})$, then, is a standardization of the data based on the investigator's input. Finally, we state a theorem – a corollary of Theorem 3 in Ferguson (1973) – that is of use in the next two sections.

Theorem 1. *Let* $p \geq 1$, $(\boldsymbol{\mu}_i, \mathbf{T}_i) \sim G$, $G \sim DP(\alpha, G_0)$. *Take* $(\boldsymbol{\mu}_0, \mathbf{T}_0) \sim G_0$. *Then* $E(\boldsymbol{\mu}_i) = E(\boldsymbol{\mu}_0)$, $E(\boldsymbol{\mu}_i\boldsymbol{\mu}_i') = E(\boldsymbol{\mu}_0\boldsymbol{\mu}_0')$, $E(\mathbf{T}_i) = E(\mathbf{T}_0)$ *and* $E(\mathbf{T}_i^{-1}) = E(\mathbf{T}_0^{-1})$.

3.2 Hyperparameter Selection for Scalar Data

We first consider the scalar data case, where $p = 1$. The DPM model requires choosing 6 scalar hyperparameters of the distribution of G_0 : $m_\mu, k_\tau, a_\lambda, b_\lambda, k_\psi$, and W_ψ. We consider the standardization of the data in choosing values for the prior moments of G. Having specified these moments, which are functions of the hyperparameters, we can solve for the hyperparameters themselves. Given its parameters $\boldsymbol{\theta}_i$, the distribution of a data point z_i is normal with mean μ_i, precision T_i, and variance T_i^{-1}. Because the data are standardized, we expect that, on average, these means are near 0 and variances are near 1. Thus, we set the expectations of μ_i and T_i^{-1} equal to these values:

$$0 = E(\mu_i) = E(\mu_0) = m_\mu \tag{1}$$

and

$$1 = E(T_i^{-1}) = E(T_0^{-1}) = \frac{k_\psi}{(k_T - 2)W_\psi}, \tag{2}$$

provided $k_T > 2$, where $(\mu_0, T_0) \sim G_0$.

Next, we desire for the μ_i drawn from the prior distribution to lie near any of the standardized data points. That is, we choose the prior variance of μ_i to be large enough so that the spread of the μ_i's matches, a priori, the spread of the standardized data. Let $v = SD(\mu_i)$; since $E(\mu_i) = E(\mu_0) = 0$, Theorem 1 implies

$$v^2 = Var(\mu_i) = Var(\mu_0) = Var[E(\mu_0|T_0, \lambda)] + E[Var(\mu_0|T_0, \lambda)]$$

$$= Var(m_\mu) + E(\lambda^{-1}T_0^{-1}) = 0 + E(\lambda^{-1})E(T_0^{-1}) = \frac{b_\lambda}{a_\lambda - 1} \tag{3}$$

provided $a_\lambda > 1$, using (1), (2), and prior independence of λ and T_0.

To choose v, we appeal to Chebyshev's inequality. Since we are concerned with the spread of the standardized data, we apply this inequality to its empirical distribution, which has mean 0 and variance $(n-1)/n$. This gives

$$\frac{1}{n}\sum_{i=1}^{n} I(|z_i| \leq c) \geq 1 - \frac{n-1}{nc^2} \quad \text{for any } c > 0.$$

Suppose we require that the left hand side is at least a proportion $\pi \in (0,1)$. Chebyshev's inequality implies that choosing $c = \sqrt{(n-1)/[n(1-\pi)]}$ satisfies this condition that the proportion π of the z_i will fall in $[-c, c]$. Now, $\mu_0|T_0, \lambda$ has a normal distribution, so we expect that π of its density lies within $d = z_{1-(1-\pi)/2}$ standard deviations of its mean, $m_\mu = 0$. From (3), we have

$$E[Var(\mu_0|\lambda, T_0)] = v^2,$$

so we expect π of the area under the probability density of $\mu_0|T_0, \lambda$ to lie in $[-dv, dv]$. Matching this range of values of μ_0 to the range of data points, $[-c, c]$, gives $v = \sqrt{(n-1)/[nz^2_{1-(1-\pi)/2}(1-\pi)]}$. To capture most of the data in this range and to ensure that newly sampled μ_i's lie near these data points, we would choose π to be large, say 95% or 99%. Experimentation suggests that $\pi = 99\%$ works well for a wide range of data distributions, so we recommend this value. Also, the factor $(n-1)/n$ can be replaced by 1 as this inflates v by a small amount for most practical sample sizes.

Now, we have three equations (1)–(3) and two constraints, namely $k_T > 2$ and $a_\lambda > 1$, for the six hyperparameters $m_\mu, k_\psi, k_T, W_\psi, , a_\lambda, b_\lambda$. While (1) yields $m_\mu = 0$, it is unclear how to choose others exactly. However, smaller values of k_ψ, k_T and a_λ give less informative priors for the corresponding Gamma and Wishart distributed parameters. A choice of $a_\lambda = 3/2$ implies λ has a scaled χ^2 distribution with 3 degrees of freedom, the minimal integer degrees that give $a_\lambda > 1$. Similarly, in the case $p = 1, Wi(k, W)$ is a scaled χ^2 distribution with k degrees of freedom. Then 3 is the minimal integer degrees of freedom that will satisfy the constraint $k_T > 2$, so we set $k_T = 3$ and $k_\psi = 1$. With these choices and v as chosen above, (1)–(3) give unique values for m_μ, b_λ, and W_ψ, completing the prior specification.

3.3 Hyperparameter Selection for Vector Data

Here again, we need to specify 6 hyperparameters; the only changes are that $\mathbf{m}_\mu$ is a vector and $\mathbf{W}_\psi$ is a matrix. Similar to the univariate case, on the average we expect $\mathbf{z_i}|\boldsymbol{\mu}_i, \mathbf{T}_i$ has mean close to $\mathbf{0}$ and covariance matrix close to $\mathbf{I}$, since the data is standardized. This implies

$$\mathbf{0} = E(\boldsymbol{\mu}_i) = E(\boldsymbol{\mu}_0) = \mathbf{m}_\mu \tag{4}$$

and

$$\mathbf{I} = E(\mathbf{T}_i^{-1}) = E(\mathbf{T}_0^{-1}) = \frac{k_\psi}{k_T - p - 1}\mathbf{W}_\psi^{-1}, \tag{5}$$

provided $k_T > p + 1$. Standardization also implies that setting $Var(\boldsymbol{\mu}_i) = v^2\mathbf{I}$ for some $v > 0$ is sensible. Since $E(\boldsymbol{\mu}_i) = E(\boldsymbol{\mu}_0) = \mathbf{0}$, then

$$v^2\mathbf{I} = Var(\boldsymbol{\mu}_i) = Var(\boldsymbol{\mu}_0) = Var[E(\boldsymbol{\mu}_0|\mathbf{T}_0, \lambda)] + E[Var(\boldsymbol{\mu}_0|\mathbf{T}_0, \lambda)] = \frac{b_\lambda}{a_\lambda - 1}\mathbf{I} \tag{6}$$

provided $a_\lambda > 1$ and using (4), (5), and prior independence of λ and $\mathbf{T}_0$.

The empirical distribution of the standardized data has mean $\mathbf{0}$ and covariance matrix $\mathbf{I}(n-1)/n$. Applying a multivariate version of Chebyshev's Inequality (Chen, 2007) to the empirical distribution, we get

$$\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{z}_i^T\mathbf{z}_i \le c^2) \ge 1 - \frac{p(n-1)}{nc^2} \quad \text{for any } c > 0.$$

To ensure that the Euclidean length of the $\mathbf{z}_i$ is within c units of the origin for a proportion π of the data, we set $c = \sqrt{p(n-1)/[n(1-\pi)]}$. Also, $\boldsymbol{\mu}_0|\mathbf{T}_0, \lambda$ is normally distributed with mean $\mathbf{m}_\mu = \mathbf{0}$ and $E[Var(\boldsymbol{\mu}_0|\mathbf{T}_0, \lambda)] = v^2\mathbf{I}$, so we expect π of the volume under the density of $\boldsymbol{\mu}_0$ to lie within Euclidean distance dv of the origin for some $d > 0$. Then, on the average, $Var(\boldsymbol{\mu}_0|\mathbf{T}_0, \lambda)$ is close to $v^2\mathbf{I}$, so

$$\begin{aligned}\pi &= P(\boldsymbol{\mu}_0^T\boldsymbol{\mu}_0 \le d^2v^2|\mathbf{T}_0, \lambda) = P(\boldsymbol{\mu}_0^T(v^2I)^{-1}\boldsymbol{\mu}_0 \le d^2|\mathbf{T}_0, \lambda) \\ &\approx P(\chi_p^2 \le d^2).\end{aligned}$$

Therefore, we set $d = \sqrt{\chi^2_{p,\pi}}$. Setting $dv = c$ as before, we get $v = \sqrt{\frac{p(n-1)}{n\chi^2_{p,\pi}(1-\pi)}}$.

Similar to the univariate case, we set $a_\lambda = 3/2$ and $k_T = p + 2$ and $k_\psi = p$, the minimal integer degrees of freedom that satisfy $k_T > p + 1$ as required. Then we can obtain $\mathbf{m}_\mu, b_\lambda$, and $\mathbf{W}_\psi$ from equations (4)–(6). Using the fact that $\chi^2_{1,\pi} = z^2_{1-(1-\pi)/2}$ for any π, it is easy to see that the choice of hyperparameters for the vector data case reduces to the scalar case when $p = 1$.

3.4 A Different View: Prior Specification on Mixture Components

In the preceding, we derived a prior for G by placing constraints on moments of its distribution. This, in turn, places a prior on the $\boldsymbol{\theta}_i$, since $\boldsymbol{\theta}_i|G \sim G$. From another viewpoint, we have specified a prior for the normal mixture components $f(\cdot|\boldsymbol{\theta}_i)$. We wish to have mixture components that are suitable for density estimation of the standardized data. Because the majority of data points will lie near $\mathbf{0}$, we set $E(\boldsymbol{\mu}_i) = \mathbf{0}$ and $Var(\boldsymbol{\mu}_i) = v^2\mathbf{I}$ in order to ensure that, a priori, most mixture components are centered near $\mathbf{0}$. Setting $E(T_i^{-1}) = \mathbf{I}$ places a constraint on how dispersed the components are, providing mixture components that are, on the average, neither extremely dispersed nor extremely concentrated.

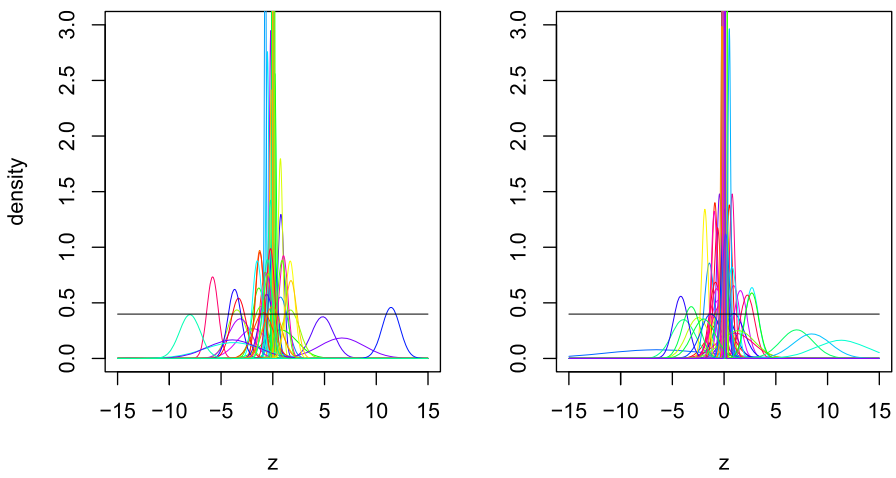

Figure 1: A Hundred Gaussian DPM Mixture Components from LIO Prior; Section 3.4.

Figure 1 shows two sets of randomly generated mixture components from our prior in the scalar data case; each plot contains 50 components. To obtain the components, we generated a sample of $\boldsymbol{\theta}_i$ from G using the stick breaking procedure in Sethuraman (1994). The black line shows the height of a standard normal density at 0 and is included as a benchmark. We see many mixture components centered near the origin, in the range $[-5, 5]$; this includes both sharply peaked and more diffuse curves. By Chebyshev's inequality, 96% of standardized data points will lie in $[-5, 5]$, so this set of components will be useful for estimating density at points near the origin. A few curves, including sharply peaked ones, are centered outside of the range $[-5, 5]$ and help to estimate the density at outliers. Our specification intends for 99% of mixture components to be centered in $[-10, 10]$; in these plots, 98% of the components are centered there.

As a result of specifying a prior on the mixture components, we have also specified a prior for the infinite mixture of these components. In Figure 2, we show 20 prior predictive densities, 10 in each plot. Though the majority of these curves are centered near 0, we do see densities centered outside of $[-1, 1]$. Moreover, the sample includes skewed, multimodal, heavy-tailed, and sharply peaked densities. This permits the model to accommodate many data distributions and shows that, though we expect the transformed data to be centered at 0 with unit scale, the LIO prior does not strictly enforce these conditions.

3.5 Examples

In the first example, we test this prior with 200 points generated from a univariate standard Cauchy distribution. In Figure 3, we see the estimates and 95% pointwise credible intervals (CI) for the density of this distribution along with the true density curve. A rug plot is included; 25 points fell outside the range [-6,6] and are not shown.

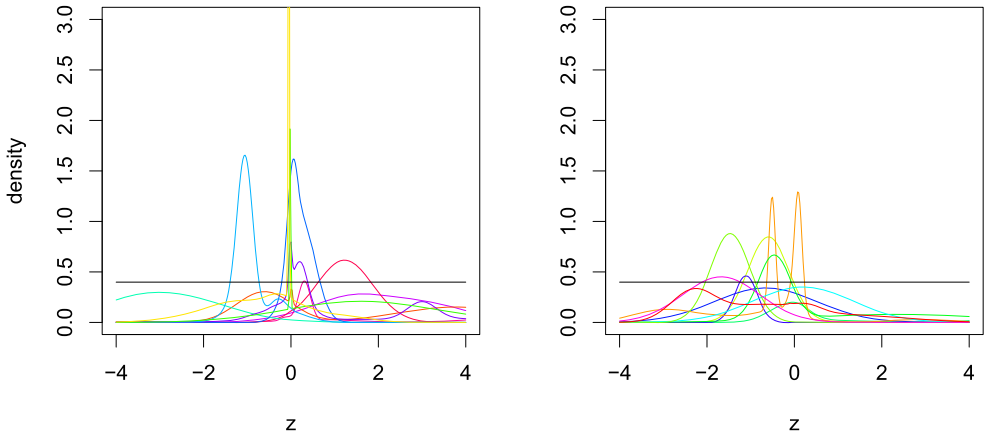

Figure 2: Twenty Prior Predictive Densities from Gaussian DPM with LIO Prior; Section 3.4.

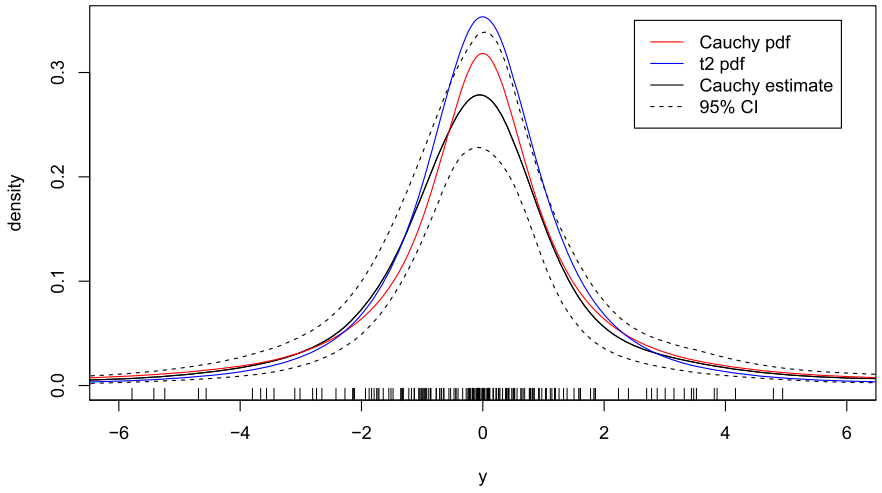

Figure 3: Density estimation of Cauchy distribution; Example 1, Section 3.5.

The credible intervals contain all of the true density, showing that this model performs well even with such "badly behaved" data. The plot also shows the density of a t distribution with 2 degrees of freedom. The credible intervals exclude the t_2 density for the range $[-0.5, 1.0]$. This demonstrates that the Gaussian DPM with our LIO prior can adeptly estimate a Cauchy distribution and, furthermore, is sensitive enough to discriminate between Cauchy/t_1 and t_2 distributions. In this simulated example, we used the true median and 95th percentile of the Cauchy distribution as scaling input. Sensitivity to such choices is considered in Section 5.

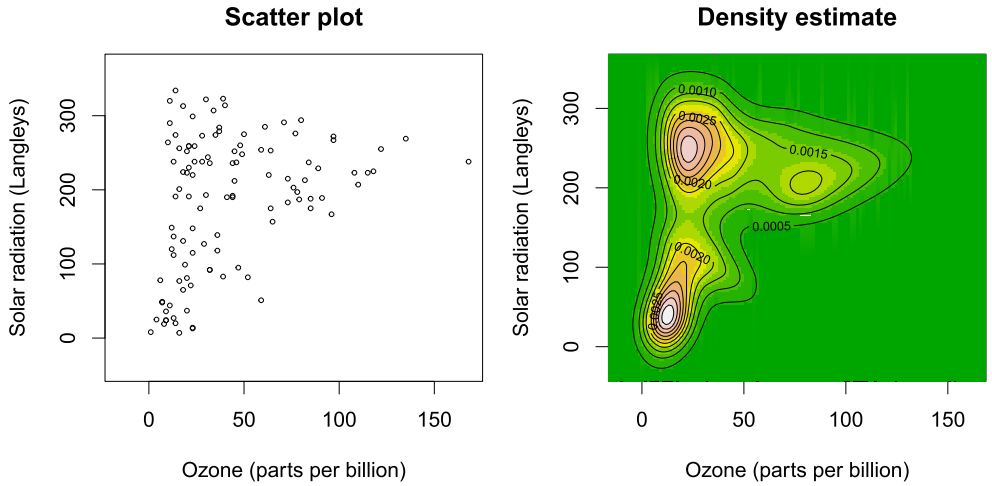

Figure 4: Plots from air quality data; Example 2, Section 3.5.

The next example uses data from air quality measurements in New York, from May to September 1973, contained in the R dataset "airquality". We estimate the bivariate distribution of ozone and solar radiation levels from 111 pairs of measurements in this set. Figure 4 has a scatter plot of the data and the density estimate. The estimate appears to fit the data quite well. Because the ozone and radiation levels only take on positive values, however, some density is placed outside the possible range of values. Using a log transformation of the levels before fitting might give even better estimation while ensuring that all density is placed within the possible range of values. In the absence of external information, for illustrative purposes, we used needed scaling percentiles from the data.

Example 3 illustrates density estimation using 400 data vectors from a bivariate mixture distribution, $F = 0.5F_1 + 0.5F_2$. Here F_1 is the bivariate t distribution with 5 degrees of freedom and an identity covariance matrix, while F_2 is a bivariate normal with mean $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1/3 & 1/3 \\ 1/3 & 4/3 \end{pmatrix}$. Figure 5 shows four plots: a scatter plot of the data, contour plots of the true and estimated density of the mixture distribution, and a coverage plot. The density was estimated on a 127x127 grid of points. The coverage plot shows whether the true density falls within the 95% pointwise credible interval at each point in the grid, with white squares indicating coverage and red indicating noncoverage. The density estimate is quite similar to the true density. This is impressive, considering that the data's distribution is a mixture of a bivariate normal distribution with positive correlation of 0.5 and a more dispersed, uncorrelated t distribution. Furthermore, the 95% CIs contain the true density at approximately 98% of the grid points.

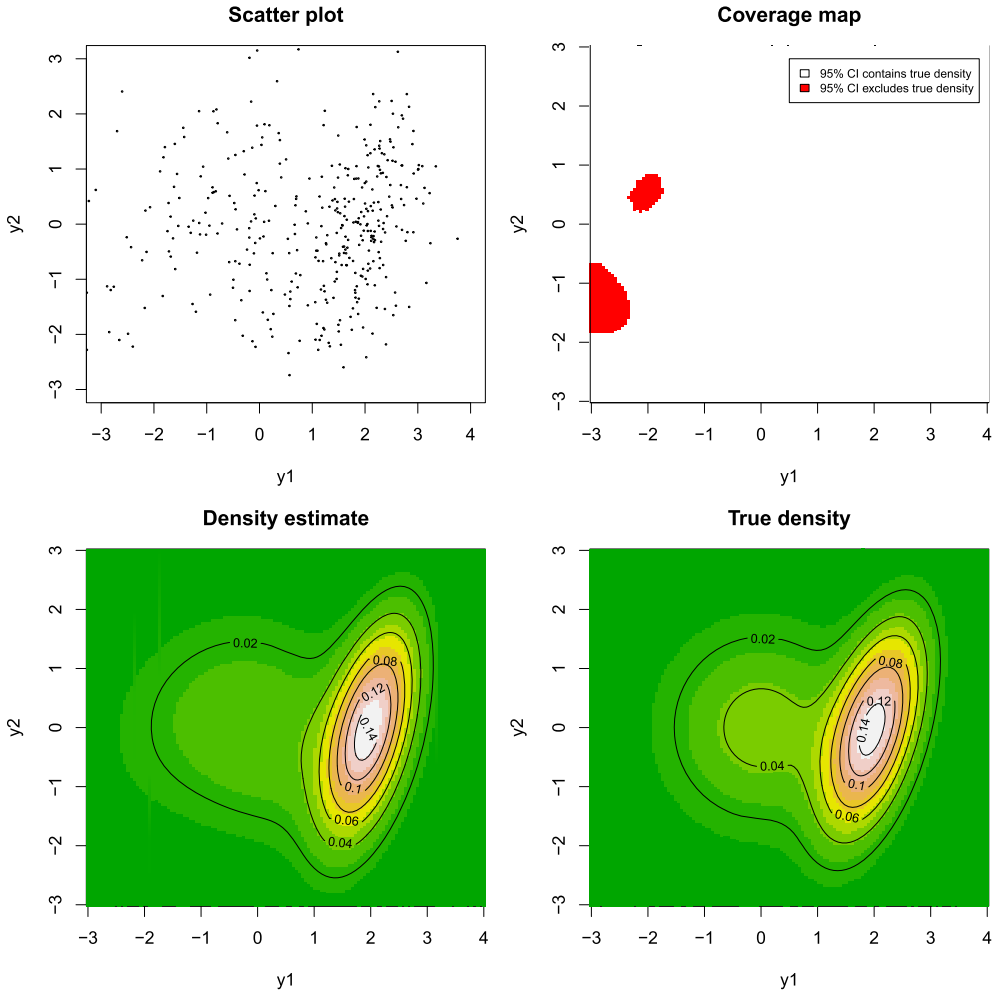

Figure 5: Plots from bivariate t_5 and normal mixture; Example 3, Section 3.5.

3.6 Prior's Effect on Number of Clusters

Although the main focus of this article is to construct a widely usable low information prior for the purpose of density estimation, the DPM model has also been used for clustering observations in many applications. See, for example, Dorazio et al. (2008), Canale and Prünster (2017) and the references therein. It is well known (Antoniak, 1974) that the so-called total mass parameter ν in the DPM of Section 3.1 strongly controls the prior distribution of the number of components in a DPM. This prior distribution also depends on the sample size n. To mitigate the issue Escobar and West (1995) first recommended a prior on ν, a gamma prior. It has been standard practice since to use such a prior, and we have chosen this to be $Ga(1, 1)$. Here we describe simulations intended

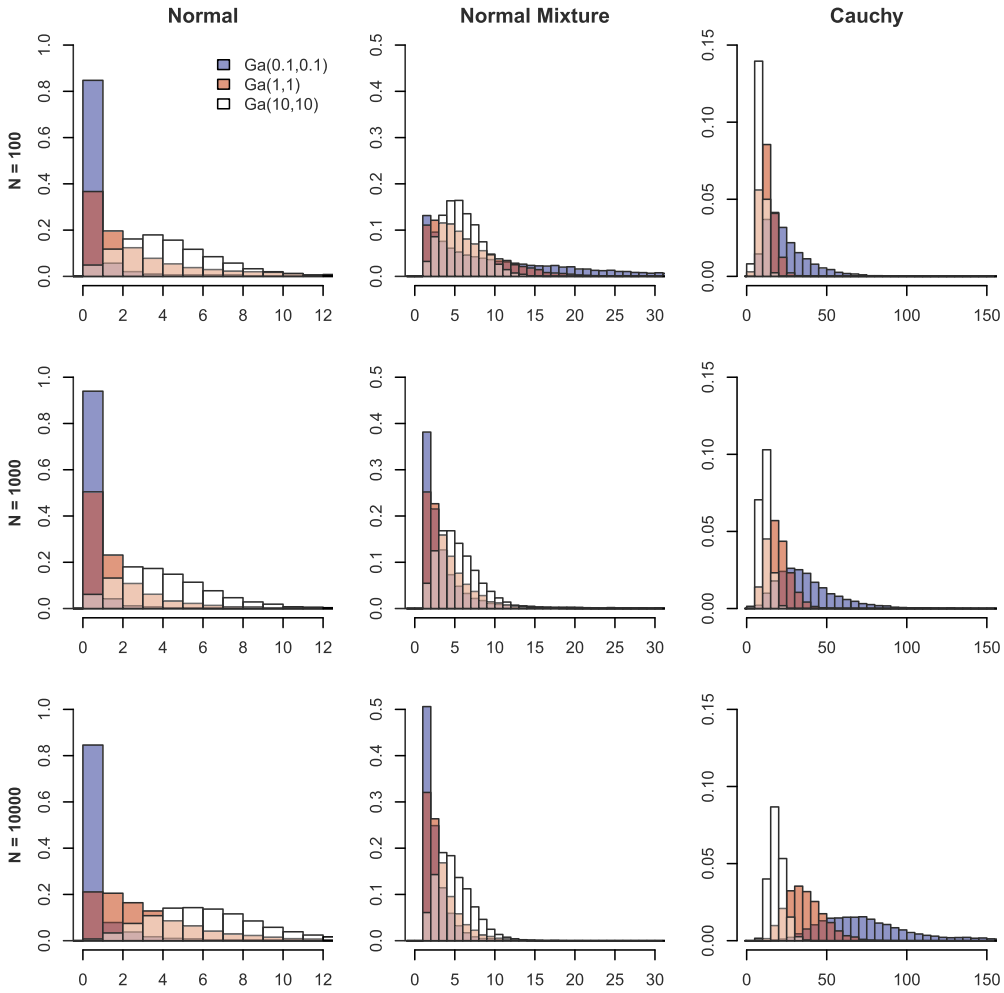

Figure 6: Posterior distributions of number of clusters; Section 3.6.

to shed some light on how the posterior distribution of the number of components, using the LIO prior, responds to the data generating distribution and the sample size.

We explored dependence of the posterior distribution of clusters under three choices of prior for ν: $Ga(0.1, 0.1)$, $Ga(1, 1)$, and $Ga(10, 10)$, all having mean 1 and with variances 10, 1, and 0.1, respectively. Sample sizes of 100, 1000, and 10000 were simulated, 200 times each, from three possible distributions: $No(5, 1/4)$, a mixture of two normals $(0.8No(0, 4)+0.2No(2, 25))$, and a standard Cauchy distribution, which can be viewed as an infinite mixture of normals. We expect to see larger numbers of clusters appear more frequently as the number of true normal mixture components of the data generating distribution increases. Figure 6 displays averaged distributions of the posterior number of clusters, obtained from histograms produced by retaining 10000 mcmc iterations after

burnin under each sample size, data distribution, and prior for ν considered. Under the $Ga(0.1, 0.1)$ and $Ga(1, 1)$ priors, this posterior distribution is quite responsive to the number of true mixture components of the data, with the bulk of the densities placed on cluster sizes of 5 or less for the normal and normal mixture and much larger cluster sizes being prevalent for the Cauchy distribution. Under the $Ga(10, 10)$ prior for ν, the posterior number of components are much more similar across data distributions.

While the results seem to indicate reasonable behavior with the recommended prior, we caution that for posterior number of clusters these are early investigations, and more work is warranted. Other possibilities, using different models, may have better promise as mentioned in the last paragraph of the Discussion section. On the other hand, we are confident in our recommendation of the prior for inference on functions such as density, cumulative distribution and hazard (Figure 1 in Supplementary Material (Shi et al., 2018)) with the LIO prior parameter settings.

4 DPM of Weibull Distributions

The proposed prior here is designed for the model of Kottas (2006). When both parameters of the Weibull distribution are given a flexible DP prior, this model approximates arbitrarily closely any distribution on the positive real line. The model is especially convenient for time-to-event data as the Weibull distribution offers simple mathematical expressions for the survival, hazard, cumulative hazard and density functions. Moreover, likelihood expressions for right, left and interval censored data remain tractable. After establishing notation for the model, we construct a LIO prior for it. Although the details apply only to the DPM of Weibulls, we note that the method of construction can be adapted to any DPM model with kernel family closed under scale change; for example, the Gamma family.

4.1 Model Specification

We begin with $y_1, \ldots, y_n$ denoting conditionally iid observations modeled with a DPM of Weibulls. As in the Gaussian case, the first step is to rescale the data to a convenient fixed scale. Using a contextually specified value c for the 95th percentile of the data's underlying distribution, we make the transformation $z_i = 10y_i/c$. Then, generally following Kottas (2006), we fit this model:

$$\begin{aligned}
z_i|\alpha_i, \lambda_i &\overset{ind}{\sim} Weib(z_i|\alpha_i, \lambda_i), \quad i = 1, \ldots, n \\
(\alpha_i, \lambda_i)|G &\overset{ind}{\sim} G, \quad i = 1, \ldots, n \\
G &\sim DP(G_0, \nu) \\
G_0 &\propto Ga(\lambda|\alpha_0, \lambda_0)Ga(\alpha|\alpha_\alpha, \lambda_\alpha)I_{(f(\lambda),\infty)}(\alpha) \\
\lambda_0 &\sim Ga(\alpha_{00}, \lambda_{00}) \\
\nu &\sim Ga(a, b).
\end{aligned}$$

Here again, $x \sim Ga(\alpha, \lambda)$ means x has density $\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ and $x \sim Weib(\alpha, \lambda)$ means its density is $\lambda\alpha x^{\alpha-1} e^{-\lambda x^\alpha}$; with, in both cases, $x > 0,\ \alpha > 0,\ \lambda > 0$. As before, the concentration parameter ν is set to have a $Ga(a, b)$ prior with $a = 1$ and $b = 1$ (Escobar and West, 1995).

The model here differs slightly from that in Kottas (2006) in one aspect: the form of G_0. The original model of Kottas (2006) uses $\lambda \sim Ga(\cdot, \cdot)$ and an independent Uniform-Pareto distributed α denoted $\alpha \sim UPar(a, b)$ and defined by $\alpha|\phi \sim U(0, \phi),\ \phi \sim Pareto(a, b)$ with density of ϕ given by $ba^b\phi^{-(b+1)} I_{(a,\infty)}(\phi),\ a > 0, b > 0$. We use instead a bivariate prior for (α, λ) employing a product of two gammas with a restriction that keeps G_0's support away from the origin through a choice of $f(\lambda)$ made in Section 4.2 below.

As in Section 3, inference for quantities related to the original data $y_1, \ldots, y_n$ can be recovered from fitting the above model to $z_1, \ldots, z_n$ since $[z] = \sum_{k=1}^{\infty} q_k Weib(\alpha_k, \lambda_k)$ implies $[y] = \sum_{k=1}^{\infty} q_k Weib(\alpha_k, \lambda_k a^{\alpha_k})$, with $a = 10/c$.

4.2 Hyperparameter selection

The approach here is distinct from that for a mixture of Gaussians where we used Chebyshev's inequality and some expectation arguments. Here we work more directly with Weibull distributed mixture components that are deemed desirable with our low-information goals on the pre-fixed data scale. We generate (α, λ) pairs corresponding to such components, inspect these visually, and use heuristics to find parameter specifications that generate similar collections. Details of the process follow.

As two distinct percentiles determine (α, λ) for a Weibull distribution, we began by working with the 5th and 95th percentiles, denoted t_1 and t_2, respectively. We let t_1 range from 0.1 to 24.5 and t_2 from $t_1 + 0.5$ to 25, both by increments of 0.1. We also added a restriction, $t_1/t_2 < 0.95$, to avoid very spikey distributions. This generated the 29487 pairs (α, λ) plotted in the left half of Figure 8.

Working first with the marginal of λ (Figure 7, left half), our goal was to determine α_0, α_{00} and λ_{00} related to λ in the model for $z_1, \ldots, z_n$. Treating the 29487 λ's as data, with the following model and priors:

$$\lambda \sim Ga(\alpha_0, \lambda_0), \quad \lambda_0 \sim Ga(\alpha_{00}, \lambda_{00})$$
$$\alpha_0 \sim UPar(1, 1), \quad \alpha_{00} \sim UPar(1, 1), \quad \lambda_{00} \sim Ga(0.001, 0.001)$$

we used medians of posterior MCMC (Markov chain Monte Carlo) samples to arrive at $\alpha_0 = 0.035,\ \alpha_{00} = 1.354$ and $\lambda_{00} = 7.181$. Using these values in the above hierarchical model for λ, we generated samples which formed the histogram in the right half of Figure 7.

With the marginal of λ in hand, the next task was to specify α_α and λ_α in the prior for α. In the model specification, the lower limit $f(\lambda)$ is intended to avoid near-zero

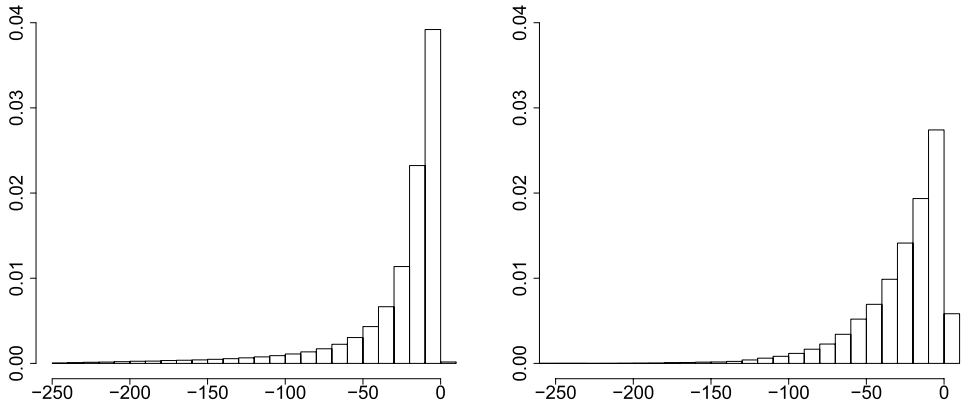

Figure 7: Matching histograms of $\log(\lambda)$; Section 4.2.

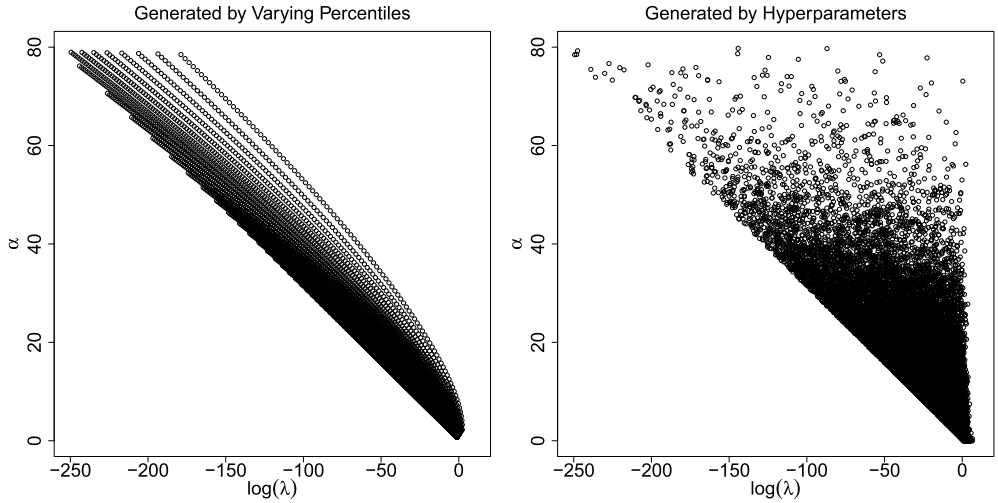

Figure 8: Scatter matching, α and $\log \lambda$; Section 4.2.

values for both α and λ as such values correspond to distributions that have an infinite spike at 0 yet assign substantial probabilities to large values. Since $z_1, \ldots, z_n$ are on a pre-fixed scale not greatly exceeding 10, restricting the 95^{th} percentile to 25 or less is a reasonable specification. This leads to $f(\lambda) = max(0, \log\{\log(20)/\lambda\}/\log(25))$. Using a trial and error process with visual inspections of scatter-plots of data generated under various combinations of $(\alpha_\alpha, \lambda_\alpha)$ resulted in the right half of Figure 8 with $\alpha_\alpha = 0.2$ and $\lambda_\alpha = 0.1$. This completes the hyperparameter selection we recommend for the LIO prior.

Figure 9 offers an insight into the LIO prior by plotting 100 realizations of survival functions generated from the full prior using the stick-breaking method (Sethuraman,

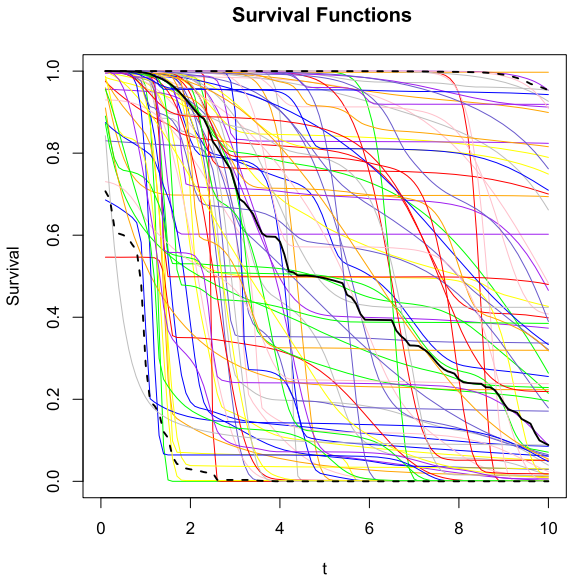

Figure 9: Survival Functions Generated from LIO prior; Section 4.2.

1994). Colored lines show individual random survival functions at a grid of time points. The black solid line is the pointwise median of 10000 such realizations and the dashed black lines represent pointwise 2.5 and 97.5 percentiles. The prior appears to satisfy the low-information goal on the pre-fixed scale.

4.3 Examples

In this section we present inference demonstrations using the LIO prior for survival, density and hazard functions with single datasets of 200 observations each, with 10% right censoring and 10% interval censoring, generated from a mixture of lognormal distributions, $0.8LN(0, 0.25) + 0.2LN(1.2, 0.02)$, which was used in Kottas (2006) (Figure 10). Figure 11 demonstrates the case of heavy right censoring as often occurs at end of study. The Supplementary Material includes additional examples. In all examples the specified 95th percentile equaled the true value. Blue lines are the estimates (solid lines) and 95% pointwise credible intervals (dashed lines) provided by the DPM of Weibulls model with the LIO prior. Red lines show survival, density and hazard functions from which observations were generated. Black lines in the survival plots are the NPMLE (nonparametric maximum likelihood estimate) (Turnbull, 1974) estimates and 95% pointwise confidence intervals for them.

Figure 11 data generation consisted of 2000 observations, 95% right censored at 0.5, generated from the same mixture of log-normals as in the previous figure. It is interesting to see the credible intervals beyond 0.5 immediately reflecting the lack of information there. In practice, elicitation of the 95th percentile may be challenging in

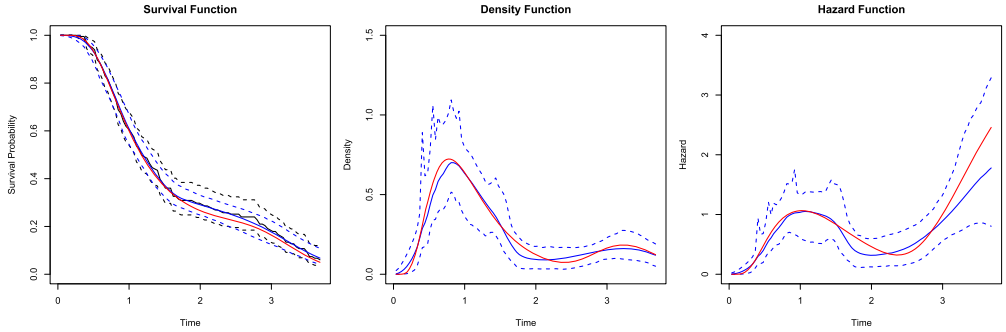

Figure 10: Survival Function, Density Function and Hazard Function Estimates of a mixture of Lognormal distributions; Section 4.3.

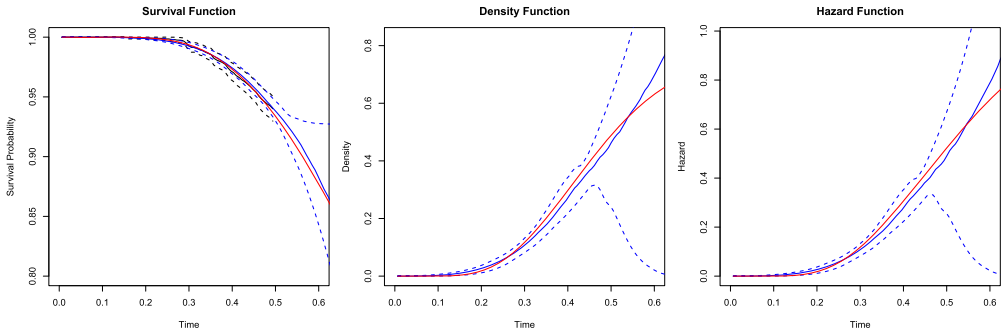

Figure 11: Survival Function Estimates of Heavily Right Censored Data; Section 4.3.

the presence of heavy right censoring at the largest observed time t_{max}. We recommend eliciting the survival probability q at t_{max} and using the prior mean survival (solid black line in Figure 9) to find t_q such that $E(S(t_q)) = q$. Then specify the 95th percentile as $10t_{max}/t_q$.

5 Sensitivity Analysis and Comparison with Empirical Methods

The only information that the LIO prior requires from the investigator is a specification of the scale of the data's underlying distribution, obtained from the median and 95th percentile for the mixture-of-Gaussians model and the 95th percentile for the mixture-of-Weibulls model. A question of interest is how much any misspecification of the scale would affect the results. We address this question through simulations. In addition, we compare the performance of the two DPM models under their respective LIO priors with empirical methods.

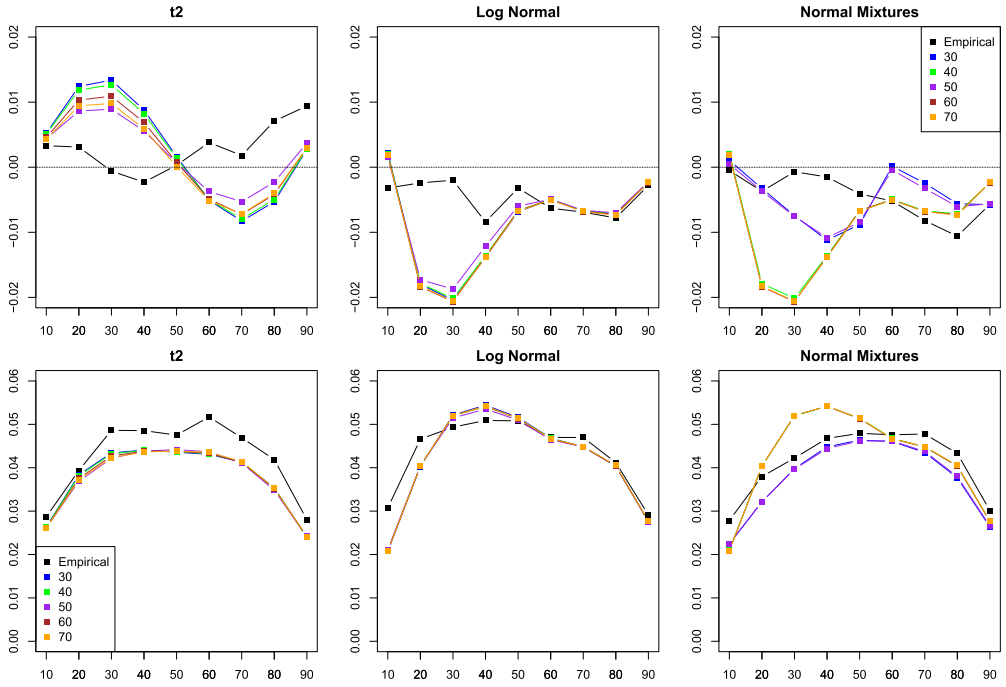

Figure 12: Sensitivity to median misspecification, Gaussian DPM: bias in top row, rmse in bottom row for CDF at 9 deciles; Section 5.1.

5.1 Sensitivity Analysis

To evaluate sensitivity to specification of the median (95th percentile) we varied this input, setting it to the true 30th, 40th, 50th, 60th, 70th (75th, 90th, 95th, 99th and 99.9th) percentiles of the underlying distribution. We then studied the performance of the posterior mean CDF at 9 deciles of the underlying distribution. Thus the true value of the estimation targets are 0.1 to 0.9 by increments of 0.1. We randomly generated 200 datasets of 100 observations each from the following three distributions:

1. t_2: the standard t distribution with 2 degrees of freedom, representing a distribution with tails heavier than those of the Gaussian;
2. lnorm: the lognormal distribution, exp[Normal$(2, 1)$], representing a skewed distribution;
3. mixnorm: a mixture of two Gaussians, 0.5 Normal$(0, 1^2)$ + 0.5 Normal$(4, 1.5^2)$, representing a multimodal distribution.

Performance was measured via bias $= \frac{1}{200}\sum_{i=1}^{200}(\theta_i - \theta)$ and root mean squared error (rmse) computed as $\sqrt{\frac{1}{200}\sum_{i=1}^{200}(\theta_i - \theta)^2}$ where θ_i is the posterior mean from dataset i and θ the true value of the inference target. Figures 12 and 13 display the performance measures in panels of six plots: bias in the top row, rmse in the second row. Columns

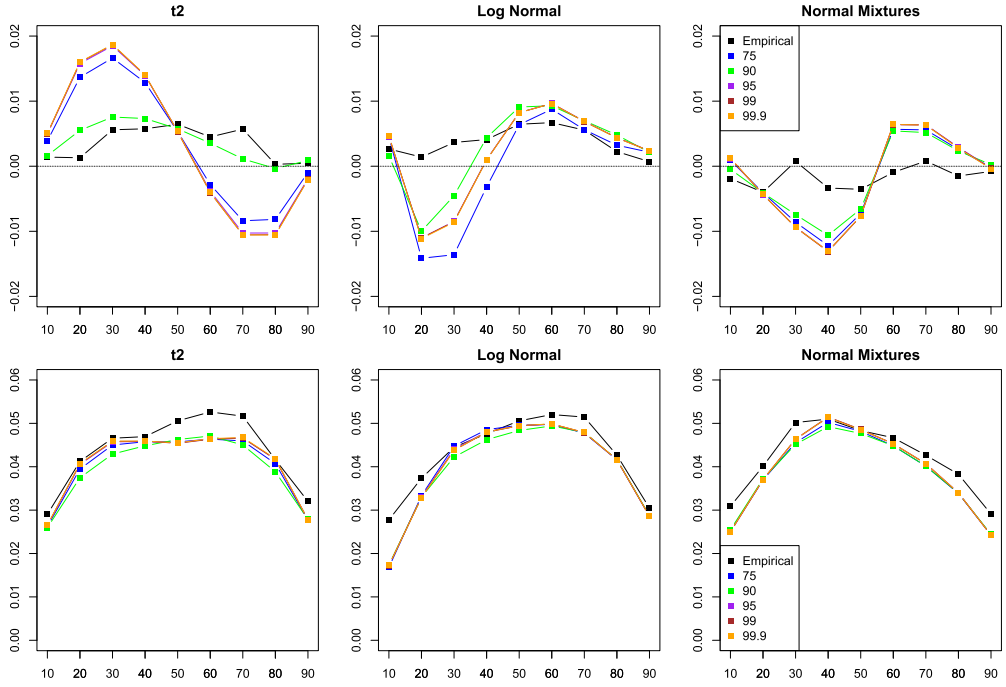

Figure 13: Sensitivity to 95th percentile misspecification, Gaussian DPM: bias in top row, rmse in bottom row for CDF at 9 deciles; Section 5.1.

correspond to the three data generating distributions. Horizontal axis markings in each plot indicate the percentile at which posterior CDF means were calculated. Different colors represent scaling input percentiles. An empirical estimate in black is also included as a benchmark. Bias is slightly worse and rmse slightly higher with misspecification of the median in the normal mixture case. Misspecification of the 95th percentile appears to be even less consequential. Overall, agreement with empirical estimates is reasonable.

5.2 Comparison with Empirical Methods

Using the median and three specifications (90th, 95th, 99.9th) for the 95th percentile of the data generating distribution as input to the LIO prior, we compared the performance of the Gaussian DPM and the ECDF (empirical cumulative distribution function) for the three specified distributions. Here, we used 200 simulated datasets of 100 or 1000 observations each. Figure 14 shows bias and root mean squared error (rmse) at the deciles of the data's underlying distribution. For each distribution and 95th percentile specification, the plotted performance measures are averages of the corresponding quantities at the 9 deciles. On the horizontal axis we use "100D" and "100E" to denote respective results from the Gaussian DPM and ECDF on datasets of size 100; similarly, "1000D" and "1000E" show these for sets of size 1000. Unlike the previous figure, colors here represent data generating distributions. Plot symbol shapes indicate prior specifications.

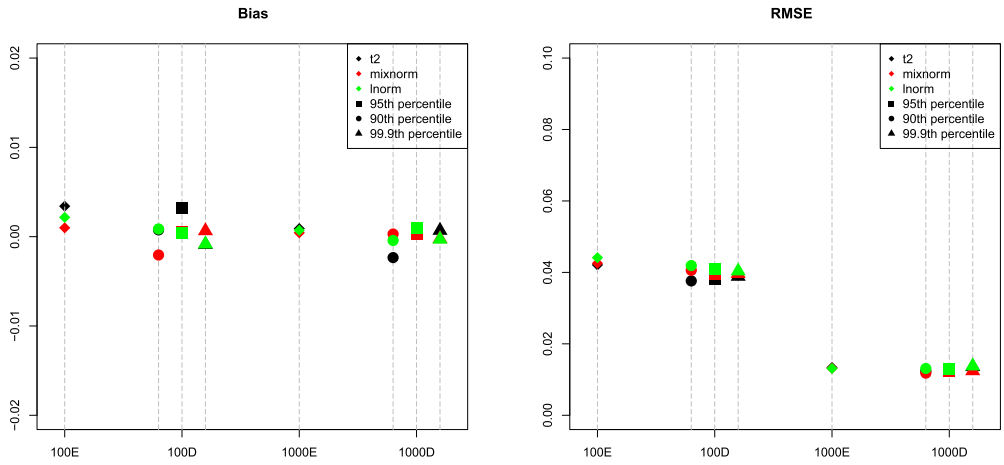

Figure 14: Gaussian DPM Comparison with Empirical DF at 9 deciles, combined; Section 5.2.

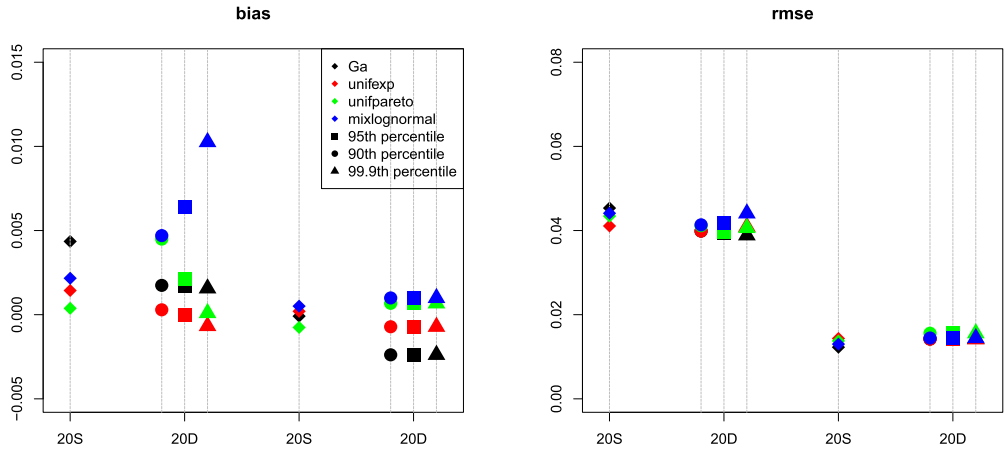

Figure 15: Comparison of Estimates of Survival function at 9 deciles from survival package (S) and LIO Weibull DPM (D); Section 5.2.

The DPM with the LIO prior and the ECDF perform very similarly with respect to bias and rmse.

For the mixture-of-Weibulls model, we used three specifications (90th, 95th, 99.9th) for the 95th percentile of the data generating distribution and compared performance with an empirical method, again using the same 4 data generating distributions as in the examples of the previous section. To see the impact of censoring rate and sample size, we added scenarios with 50% censoring (25% right censoring, 25% interval censoring) and 1000 observations. In Figure 15, the "S" on the x-axis represents the NPMLE estimates

from the R package "survival", while the "D" represents DPM of Weibulls model with LIO prior. The numerals preceding these letters indicate the censoring rate 20% or 50% (the latter figure in Supplementary Material). In each plot, the first 4 estimates are based on datasets with 100 observations while the rest are based on datasets with 1000 observations. Again, we see that the performance of the DPM is quite similar to the frequentist estimates in terms of bias and rmse.

A bivariate example is included in the Supplementary Material.

6 Convergence Considerations

Consistency of a Bayesian procedure is in a sense a frequentist validation of the procedure: For a nonparametric or semi-parametric Bayesian procedure, consistency implies convergence to a true unknown density as the number of data observations goes to ∞. We use Ghosal et al. (1999), Tokdar (2006), Wu and Ghosal (2008) and Wu and Ghosal (2010) to show convergence properties with the two LIO priors in previous sections.

Measuring convergence for a density estimation procedure is done in terms of concentration of the posterior probability around a neighborhood of the true unknown density. Let $X_1, \ldots, X_n$ be observed data $\in \mathbb{R}^p$ for some integer $p \geq 1$. Let $\mathcal{F}$ denote the space of all densities on $\mathbb{R}^p$. Let f_0 be some density on $\mathbb{R}^p$. Also for any $\epsilon > 0$ let us denote by $N_\epsilon^w(f_0)$ and $N_\epsilon^s(f_0)$ the neighborhoods of f_0 under weak and strong topology respectively. Let P_f denote the probability measure corresponding to a density f. Also let P_f^∞ denote the probability measure on the infinite dimensional random vector $\{X_i\}_{i=1}^\infty$, where the X_i are iid and $\sim f$. We begin with the formal definitions of posterior consistency.

Definition 1. *A prior* Π *is said to be weakly consistent at a density* f_0 *if for any* $\epsilon > 0$*, the random variable,*

$$\mathcal{X}_n(\epsilon) = \frac{\int_{f \in N_\epsilon^w(f_0)} \prod_{i=1}^n f(X_i) d\Pi(f)}{\int_{f \in \mathcal{F}} \prod_{i=1}^n f(X_i) d\Pi(f)} \to 1$$

as $n \to \infty$ *almost surely with respect to the measure* $P_{f_0}^\infty$*.*

Replacing the neighborhood under weak topology with the neighborhood under strong topology (also called L_1 topology), the definition of strong consistency is given as,

Definition 2. *A prior* Π *is said to be strongly consistent at a density* f_0 *if for any* $\epsilon > 0$*, the random variable,*

$$\mathcal{X}_n(\epsilon) = \frac{\int_{f \in N_\epsilon^s(f_0)} \prod_{i=1}^n f(X_i) d\Pi(f)}{\int_{f \in \mathcal{F}} \prod_{i=1}^n f(X_i) d\Pi(f)} \to 1$$

as $n \to \infty$ *almost surely with respect to the measure* $P_{f_0}^\infty$*.*

6.1 Equivalence Results

To establish posterior consistency properties of the LIO prior, we first show that it suffices to study consistency on the scaled data.

Lemma 1. *Let $Z_i = AX_i + b$, for each $i \in 1, \ldots, n$ be a linear scaling of the data $\{X_i\}_{i=1}^n$ for some positive matrix A in $\mathbb{R}^{p\times p}$ and any vector b in $\mathbb{R}^p$. Then, a prior Π achieves weak (strong) consistency at a density f_0 on $\{X_i\}_{i=1}^n$ if the induced prior $\widetilde{\Pi}$ achieved weak (strong) posterior consistency at the induced density $\widetilde{f_0}$ on $\{Z_i\}_{i=1}^n$.*

Proof. In Supplementary Material. □

Next we consider the class of densities at which consistency is shown. In the next lemma, we show that in addition to equivalence for posterior consistency, the regularity conditions and the density classes are also equivalent between the observed data and the scaled data.

Lemma 2. *Let $\{Z_i\}_{i=1}^n$ be a linear rescaling of the observed data $\{X_i\}_{i=1}^n$ as previously stated, with induced densities and priors between them. The following conditions for the induced density on rescaled data,*

1. *$\widetilde{f_0}(z)$ is nowhere 0 and is bounded above by M, $\forall z \in \mathbb{R}^p$*
2. *$|\int \widetilde{f_0}(z) \log \widetilde{f_0}(z) dz| < \infty$*
3. *For some $\delta > 0$, $|\int \widetilde{f_0}(z) \log \frac{\widetilde{f_0}(z)}{\phi_\delta(z)} dz| < \infty$, where $\phi_\delta(z) = \inf_{\|t-z\|<\delta} \widetilde{f_0}(t)$*
4. *For some $\eta > 0$, $\int \|z\|^{2(1+\eta)} \widetilde{f_0}(z) dz < \infty$,*

imply equivalent conditions on the density $f_0(x)$ on the observed data.

Proof. In Supplementary Material. □

Earlier work in the literature (Walker, 2004; Choi and Schervish, 2007) contains other slightly different regularity conditions on the true density f_0, for all of which, equivalence can be shown. We omit a detailed description here for the sake of brevity.

6.2 Consistency Results on the Scaled Data

The LIO prior in this article is used for the following three scenarios:

1. Mixture of univariate normals for scalar responses
2. Mixture of Weibulls for scalar responses
3. Mixture of multivariate normals for vector responses

Items (1)&(2) have been dealt with in Ghosal et al. (1999) and Wu and Ghosal (2008). The work in Wu and Ghosal (2008) is restricted to showing consistency at true densities having a finite second moment, which excludes some commonly used densities, e.g., the Cauchy. Tokdar (2006) significantly weakens the second moment condition, while adding additional regularity conditions on the base measure. For our item (1), results of Tokdar (2006) Theorem 3.3 directly apply. This implies weak consistency for our procedure on a wide class of true densities, including those such as the Cauchy density.

We show here briefly that a similar weakening on conditions for our item (2) is also possible as our base measure satisfies similar regularity conditions in the next lemma.

Lemma 3. *Let* $\Pi = DP(G_0, \nu)$ *denote the prior specification for our mixture of Weibulls scenario, where the base measure* G_0 *is supported on* $\mathbb{R}^+ \times \mathbb{R}^+$*. The conditions (1)–(4) of Tokdar (2006)'s Theorem 3.3 implies weak consistency of our procedure.*

Proof. In Supplementary Material. □

The proof of weak consistency for the multivariate case - for our item (3) follows from Theorem 2 in Wu and Ghosal (2010). These results also do not permit true densities for which second moment is not finite. It is possible to further impose conditions on the base measure, implying conditions on the eigenvalues of covariance matrix, but this is fairly involved, not following from earlier results; a discussion of this is omitted here.

Strong consistency (also referred to as L_1 consistency) on a restricted class of densities as given by Theorem 3 in Wu and Ghosal (2010) applies directly to our scaled data procedure, and by virtue of our equivalence results, to the induced procedure on the observed data. Some weakening of the conditions of Theorem 3 is possible for admitting a broader class of true densities, once again by imposing strict decay conditions on the tails of the base measure, but further details are omitted here.

7 Discussion

We offer a technique for an omnibus low information prior specification that can handle data of various scales in a mixture-of-Gaussians model and a mixture-of-Weibulls model. Using data simulated from a variety of distributions we demonstrated the effectiveness of these prior specifications. To implement the Gaussian DPM model with our prior, we have developed a wrapper for the DPdensity function of the R package DPpackage (Jara et al., 2011) that provides density estimation for scalar and vector-valued random samples. This is included in the R package DPWeibull (https:cran.r-project/web/packages/DPWeibull) which also includes functions for DPM of Weibulls.

We illustrated this method of prior specification for DPMs of Gaussian and Weibull distributions. A similar approach can be used to obtain a LIO prior for a DPM of any location-scale family, such as t distributions. Additionally, a similar application could be used for mixtures of distributions from a family that, like the Weibulls, are closed under a change of scale; Gamma distributions are one such family.

The process of obtaining a low information prior only needs to be done once. It is designed to generate a vague but robust prior. The construction process could be based on moment arguments as in the mixture-of-Gaussians case, or may require a more constructive effort with trial and error as we described in the mixture-of-Webulls case.

Similar to De Iorio et al. (2004)'s Dependent Dirichlet process (DDP) of Gaussian mixtures model, we have extended DPM-of-Weibulls model to a DDP regression model for survival data that can directly model event time and address censoring as well as competing risks. This work is contained in the first author's recently completed dissertation at the Medical College of Wisconsin and will be published elsewhere.

We note that Bayesian nonparametric modeling has evolved much beyond the still-popular DPM model. For example, Gibbs-type partitions (Lijoi et al., 2008) and normalized random measures with independent increments (NRMIs; Regazzini et al. (2003); Lijoi et al. (2005, 2007)) are more general models with important special cases showing robustness with respect to prior specification for the total mass parameter. It is quite possible that the methods of this article could be extended to homogeneous NRMIs (James et al., 2009), perhaps straightforwardly. We hope to report on such developments in the future, noting that our contribution for now remains a starting point. In this regard, see also work in Argiento et al. (2010, 2016).

Supplementary Material

Supplementary Material for "Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models" (DOI: 10.1214/18-BA1119SUPP; .pdf).

References

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The annals of statistics*, 1152–1174. MR0365969. 687

Argiento, R., Bianchini, I., and Guglielmi, A. (2016). "A blocked Gibbs sampler for NGG-mixture models via a priori truncation." *Statistics and Computing*, 26(3): 641–661. MR3489862. doi: https://doi.org/10.1007/s11222-015-9549-6. 700

Argiento, R., Guglielmi, A., and Pievatolo, A. (2010). "Bayesian density estimation and model selection using nonparametric hierarchical mixtures." *Computational Statistics & Data Analysis*, 54(4): 816–832. MR2580918. doi: https://doi.org/10.1016/j.csda.2009.11.002. 700

Canale, A. and Prünster, I. (2017). "Robustifying Bayesian nonparametric mixtures for count data." *Biometrics*, 73(1): 174–184. MR3632363. doi: https://doi.org/10.1111/biom.12538. 687

Chen, X. (2007). "A new generalization of Chebyshev inequality for random vectors." *arXiv preprint arXiv:0707.0805*. 683

Choi, T. and Schervish, M. J. (2007). "On posterior consistency in nonparametric regression problems." *Journal of Multivariate Analysis*, 98(10): 1969–1987. MR2396949. doi: https://doi.org/10.1016/j.jmva.2007.01.004. 698

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99(465): 205–215. MR2054299. doi: https://doi.org/10.1198/016214504000000205. 700

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(2): 635–644. MR2432438. doi: https://doi.org/10.1111/j.1541-0420.2007.00873.x. 687

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 678, 680, 687, 690

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. MR0350949. 677, 681

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis 3*. Chapman and Hall/CRC. MR2027492. 678

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics*, 2(4): 1360–1383. MR2655663. doi: https://doi.org/10.1214/08-AOAS191. 679

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). "Posterior consistency of Dirichlet mixtures in density estimation." *The Annals of Statistics*, 27(1): 143–158. MR1701105. doi: https://doi.org/10.1214/aos/1018031105. 697, 699

James, L. F., Lijoi, A., and Prünster, I. (2009). "Posterior analysis for normalized random measures with independent increments." *Scandinavian Journal of Statistics*, 36(1): 76–97. MR2508332. doi: https://doi.org/10.1111/j.1467-9469.2008.00609.x. 700

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011). "DPpackage: Bayesian semi-and nonparametric modeling in R." *Journal of Statistical Software*, 40(5): 1–30. MR3309338. doi: https://doi.org/10.1007/978-3-319-18968-0. 680, 699

Kottas, A. (2006). "Nonparametric Bayesian survival analysis using mixtures of Weibull distributions." *Journal of Statistical Planning and Inference*, 136(3): 578–596. MR2181970. doi: https://doi.org/10.1016/j.jspi.2004.08.009. 678, 689, 690, 692

Lijoi, A., Mena, R. H., and Prünster, I. (2005). "Hierarchical mixture modeling with normalized inverse-Gaussian priors." *Journal of the American Statistical Association*, 100(472): 1278–1291. MR2236441. doi: https://doi.org/10.1198/016214505000000132. 700

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(4): 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 700

Lijoi, A., Prünster, I., and Walker, S. G. (2008). "Investigating nonparametric priors with Gibbs structure." *Statistica Sinica*, 1653–1668. MR2469329. 700

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12(1): 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 677

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 679

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of normalized random measures with independent increments." *Annals of Statistics*, 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 700

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4(2): 639 – 650. MR1309433. 684, 691

Y. Shi, M. Martens, A. Banerjee, and P. Laud (2018). "Supplementary Material for "Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1119SUPP. 689

Tokdar, S. T. (2006). "Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression." *Sankhya: The Indian Journal of Statistics*, 68(1): 90–110. MR2301566. 697, 699

Turnbull, B. W. (1974). "Nonparametric estimation of a survivorship function with doubly censored data." *Journal of the American Statistical Association*, 69(345): 169–173. MR0381120. 692

Walker, S. (2004). "New approaches to Bayesian consistency." *Annals of Statistics*, 32(5): 2028–2043. MR2102501. doi: https://doi.org/10.1214/009053604000000409. 698

Wu, Y. and Ghosal, S. (2008). "Kullback Leibler property of kernel mixture priors in Bayesian density estimation." *Electronic Journal of Statistics*, 2: 298–331. MR2399197. doi: https://doi.org/10.1214/07-EJS130. 697, 699

Wu, Y. and Ghosal, S. (2010). "The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation." *Journal of Multivariate Analysis*, 101(10): 2411–2419. MR2719871. doi: https://doi.org/10.1016/j.jmva.2010.06.012. 697, 699