Technical Report

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 of the distributions of data as an infinite mixture of distributions from a specified collection. However, 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. After drawing samples from the posterior with the rescaled data, we transform the inference back to the original scale. The low information prior is selected to provide a wide variety of components for the DPM in order to generate flexible distributions for the data on the fixed scale. This scale-data-and-rescale-inference 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 favorable performance with 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 eventhough 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 randomly generated from a DP with baseline distribution G_0 and concentration parameter ν. The choice of kernel density $f(\cdot)$ determines the mixture components to use in a DPM; for example, if $f(\cdot)$ is a normal kernel, then this DPM is a mixture of Gaussians.

*8701 Watertown Plank Road Milwaukee, WI 53226 yushushi@mcw.edu mmartens@mcw.edu abanerjee@mcw.edu laud@mcw.edu

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.

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 results from the Gaussian and Weibull DPMs using the proposed priors with those from empirical methods. 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. One's first instinct might be to use a G_0 with 'high variance' to express ignorance about the prior. However, 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 recommended standardizing the data and using a weakly informative prior 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 median and 95th percentile for the mixture of Gaussians model) from the investigator knowledgeable in the subject matter. 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 sampled parameters representing posterior inference to obtain originally targeted inference.

Currently, our "black box" has two major applications: 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 for estimating distributions as it is straightforward to apply it 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 quantites $\mathbf{a} \in \mathbb{R}^p$ 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:

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

Here $No(\mathbf{m}, \mathbf{U})$ denotes a normal distribution with mean $\mathbf{m}$ and precision matrix $\mathbf{U}$, $Ga(a, b)$ 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 rate 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 $\mathbf{\Psi}$ and the second level having independent Gamma and Wishart distributions for λ and $\mathbf{\Psi}$, respectively. To be specific, $(\boldsymbol{\mu}, \mathbf{T}) \sim NoWi(\mathbf{m}, \lambda, k, \mathbf{\Psi})$ means $\boldsymbol{\mu}|\mathbf{T}, \lambda \sim No(\mathbf{m}, \lambda\mathbf{T})$ and $\mathbf{T}|\mathbf{\Psi}, k \sim Wi(k, \mathbf{\Psi})$. 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 is set to have a $Ga(a, b)$ prior with $a = 1$ and $b = 1$ (Escobar and West, 1995).

This model assumes that $\mathbf{z}_i$ arise from an infinite mixture of normal distributions. Then they have the cumulative distribution function (CDF)

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

where Φ_p is the CDF of a p-variate normal distribution $No(\mathbf{0}, \mathbf{I})$, $\mathbf{\Lambda}_i$ comes from the unique Cholesky decomposition $\mathbf{T}_i = \mathbf{\Lambda}_i\mathbf{\Lambda}_i'$, 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

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

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 in a way so that, regardless of the original data set, the transformed data tend to have a similar location and dispersion; this will justify applying a common 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 derive these in turn from 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 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.

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. The following theorem, which is a corollary of Theorem 3 in Ferguson (1973), is used repeatedly in the derivations that follow:

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 $E(\boldsymbol{\mu}_i^{\otimes k}) = E(\boldsymbol{\mu}_0^{\otimes k})$ for $k = 1, 2$ and $E(\mathbf{T}_i^k) = E(\mathbf{T}_0^k)$ for $k = \pm 1$, where $\mathbf{v}^{\otimes 1} = \mathbf{v}$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}'$ for $\mathbf{v} \in \mathbb{R}^p$.*

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

$$\begin{aligned} v^2 &= Var(\mu_i) = Var(\mu_0) = Var[E(\mu_0|T_0, \lambda)] + E[Var(\mu_0|T_0, \lambda)] \\ &= Var(m_\mu) + E(\lambda^{-1}T_0^{-1}) = 0 + E(\lambda^{-1})E(T_0^{-1}) = \frac{b_\lambda}{a_\lambda - 1} \end{aligned} \tag{3}$$

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 π. 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 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.

Having specified v, we have three equations and two constraints for the hyperparameters, requiring $k_T > 2$ and $a_\lambda > 1$. It is unclear how to choose k_T, k_ψ, and a_λ exactly; however, smaller values 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 v, a_λ, k_T, and k_ψ chosen above, equations (1)-(3) give values for m_μ, b_λ, and W_ψ, completing the prior specification.

3.3 Hyperparamter 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 $\mathbf{I}(n-1)/n$. Using 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 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 (3)-(5). 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 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 the prior predictive distribution, that is, 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 accomodate 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; 9 points fell outside the range [-10,10] and are not shown. 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 large portions of the

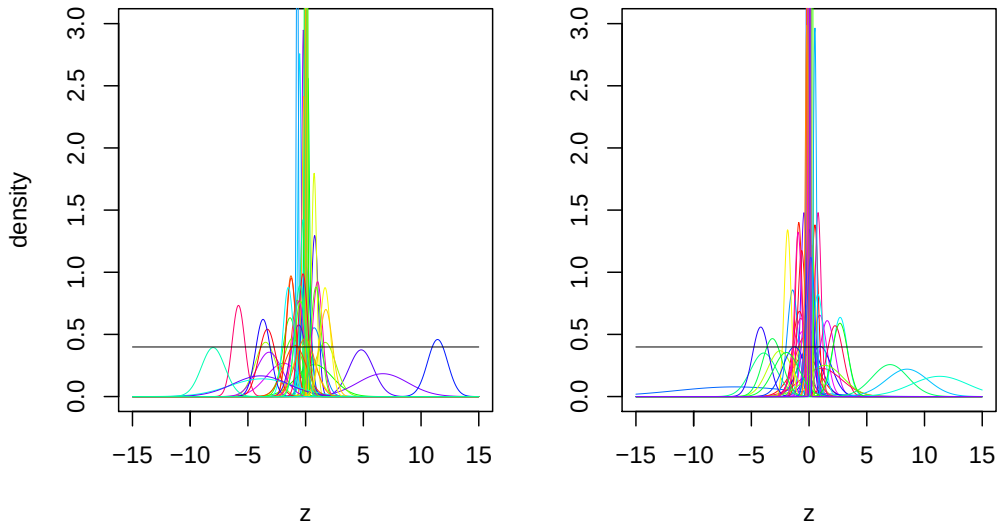

Figure 1: A Hundred Gaussian DPM Mixture Components from LIO Prior

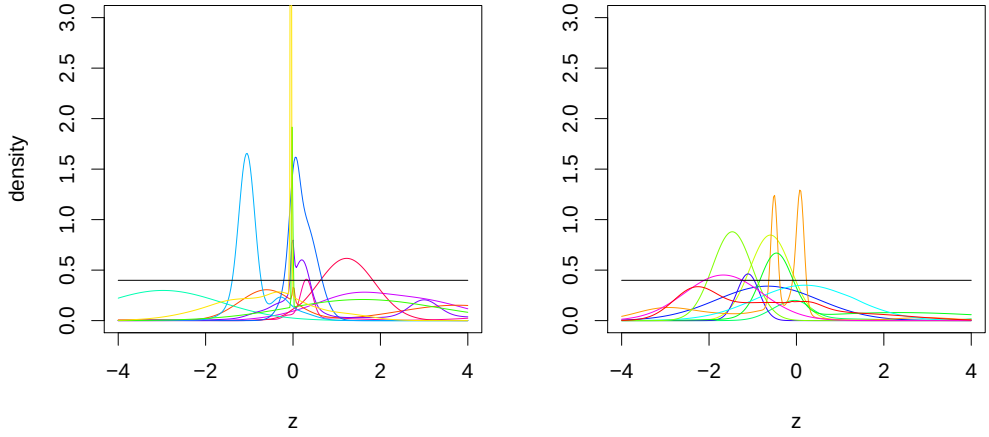

Figure 2: Twenty Prior Predictive Densities from Gaussian DPM with LIO Prior

graph, specifically the ranges $[-7, -4]$, $[-0.5, 0.5]$, and $[5, 10]$. This demonstrates that the Gaussian DPM with our LIO prior can adequately 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 known median and 95th percentile of the distribution. Sensitivity to such choices is considered in Section 5.

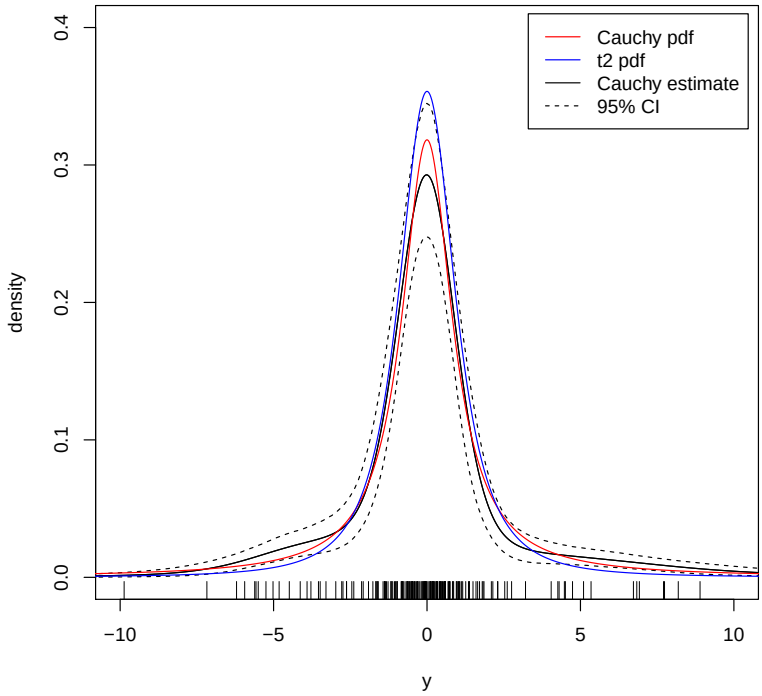

Figure 3: Density estimation of Cauchy distribution

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.

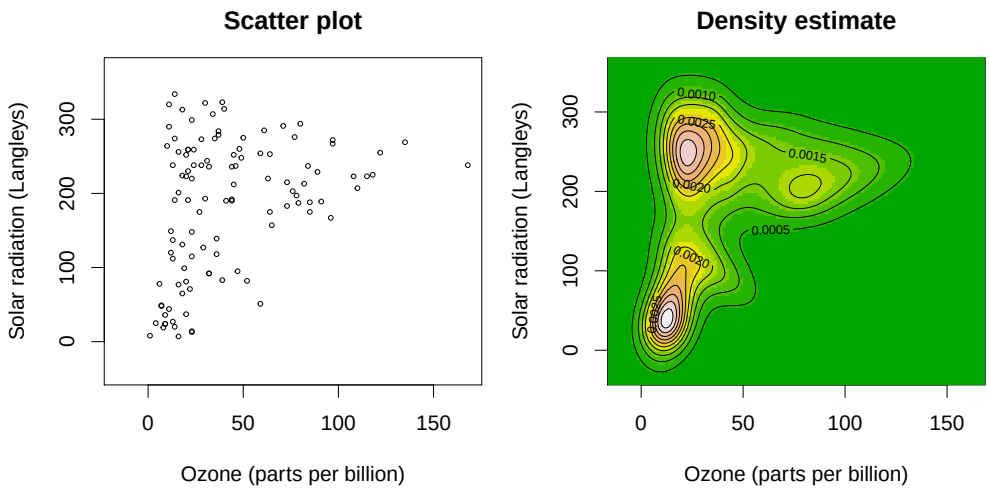

Figure 4: Plots from air quality data

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.

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. 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 $y_1, \ldots, y_n$ 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

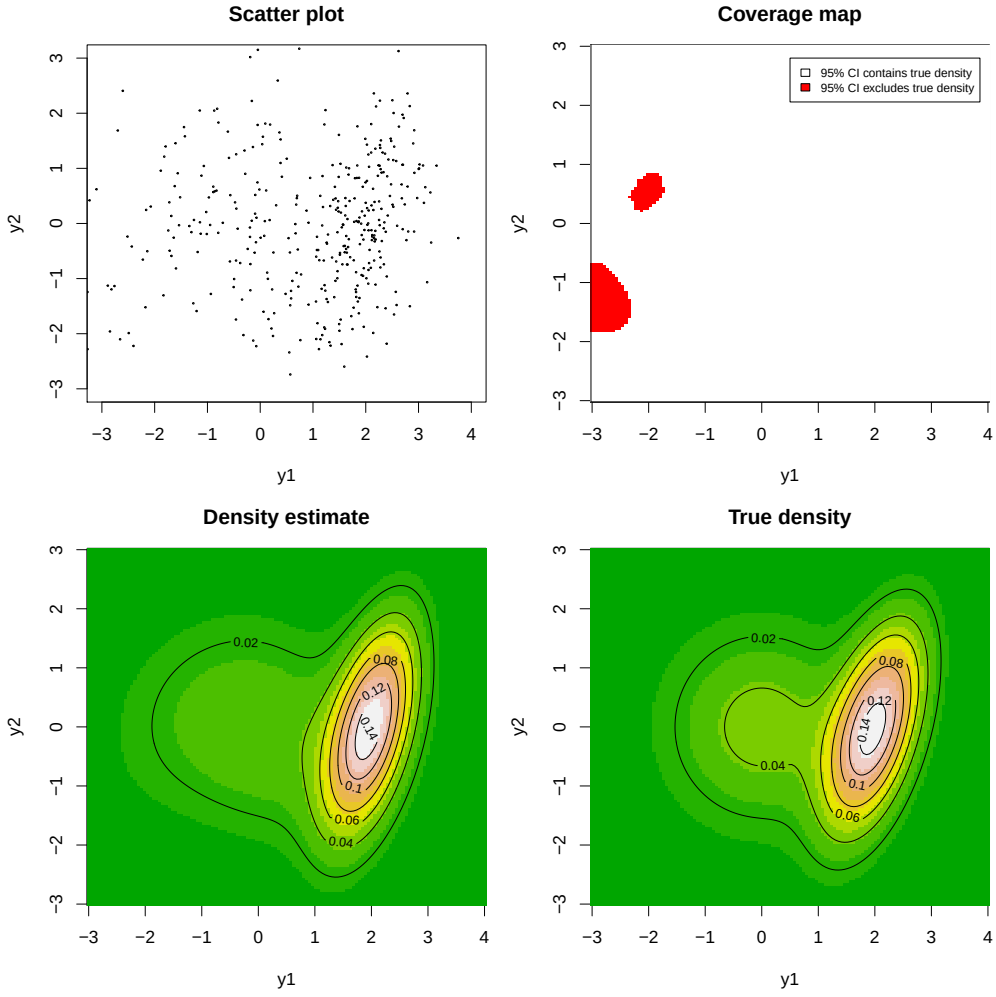

Figure 5: Plots from t_5 / normal mixture

(2006), we fit this model:

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

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 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 orginal model of Kottas (2006) uses a product of a gamma and a uniform-Pareto, the latter defined by $x \sim U - Par(a, b) \Rightarrow x|\phi \sim U(0, \phi),\ \phi \sim Pareto(a, b)$ with density of ϕ given by $ba^b\phi^{-(b+1)} I_{(a,\infty)}(\phi),\ a > 0, b > 0$. We use instead 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 $[ay] = \sum_{k=1}^{\infty} q_k Weib(\alpha_k, \lambda_k) \Rightarrow [y] = \sum_{k=1}^{\infty} q_k Weib(\alpha_k, \lambda_k a^{\alpha_k})$, 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 panel of Figure 7.

Working first with the marginal of λ (Figure 6, left panel), 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:

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

$$\alpha_0 \sim U - Par(1, 1), \quad \alpha_{00} \sim U - Par(1, 1), \quad \lambda_{00} \sim Ga(0.001, 0.001)$$

we used medians of posterior MCMC 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 panel of Figure 6.

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 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 panel of Figure 7 with $\alpha_\alpha = 0.2$ and $\lambda_\alpha = 0.1$. This completes the hyperparameter selection we recommend for the LIO prior.

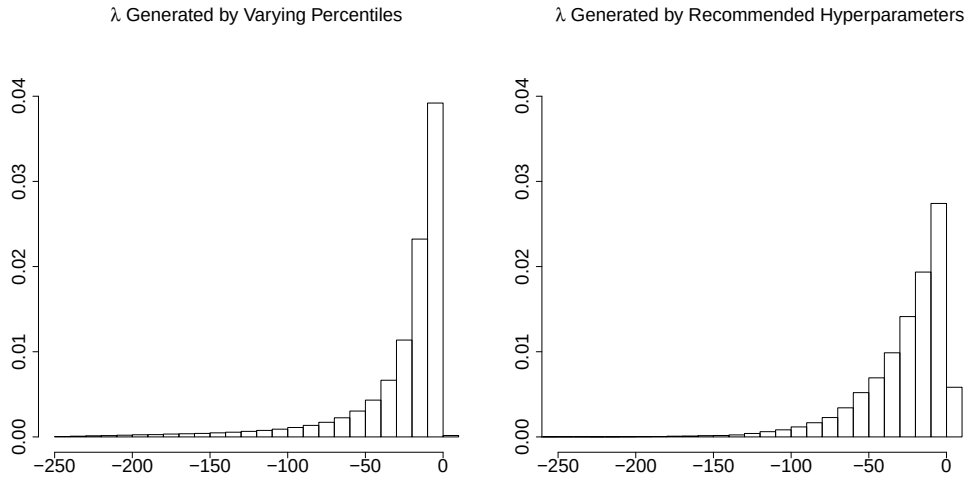

Figure 6: Histogram of the $\log(\lambda)$

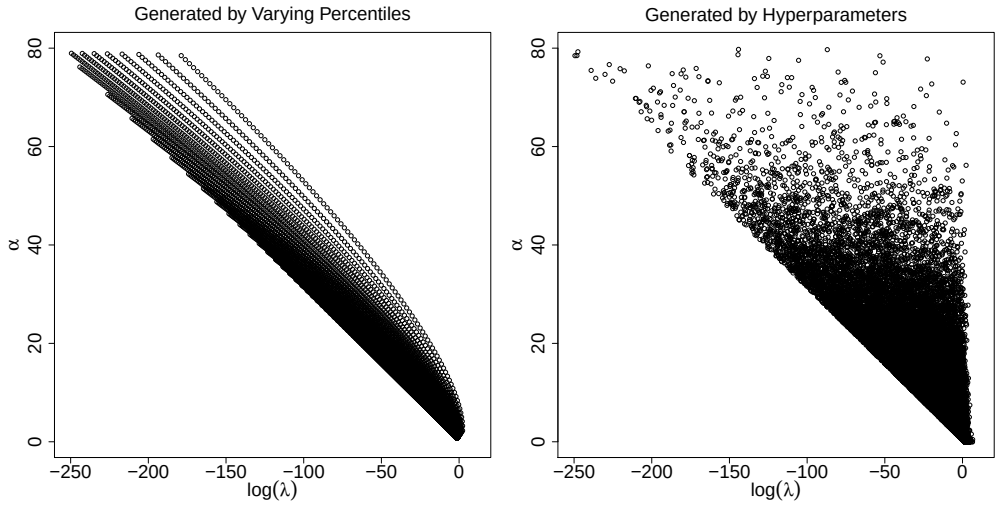

Figure 7: $\log(\lambda)$ and α

Figure 8 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, 1994). Colored lines show individual random survival functions. The black solid line

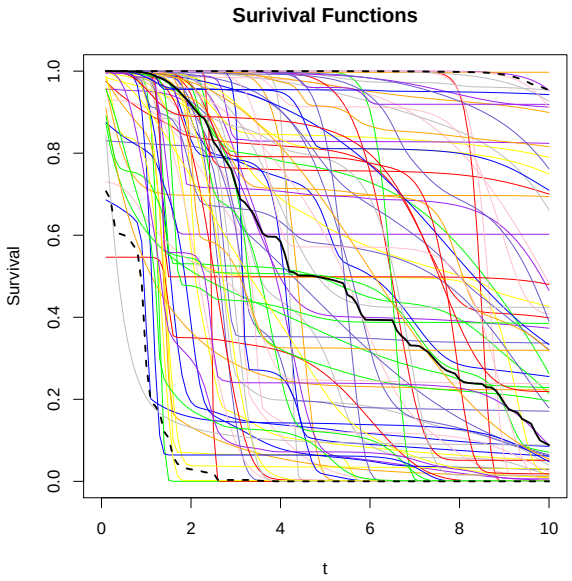

Figure 8: Survival Functions Generated from LIO prior

is the median of 10000 such realizations and the dashed black lines represent the 95% pointwise credible intervals. 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 four underlying distributions. Figures 9-12 show the results. 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 (Turnbull, 1974) estimates and 95% pointwise confidence intervals for them.

The data generating distribution for Figure 9 is Gamma(0.5,2). For Figure 10, observations were generated from a density, flat from 0 to 10/11 and exponentially decreasing after 10/11:

$$f(x) = \begin{cases} 1/11 & x < 10/11 \\ \exp(-(x - 10/11))/11 & x \geqslant 10/11 \end{cases}.$$

To consider a heavier tailed distribution, data in Figure 11 were generated from a Uniform-Pareto distribution, $U - Par(2, 2)$. Data for Figure 12 were generated from a mixture of lognormal distributions, $0.8LN(0, 0.25) + 0.2LN(1.2, 0.02)$, which was used in Kottas (2006). Finally, Figure 13 data generation was designed to mimic heavy right censoring, a situation not uncommon in practice. The data here consisted of 2000 obser-

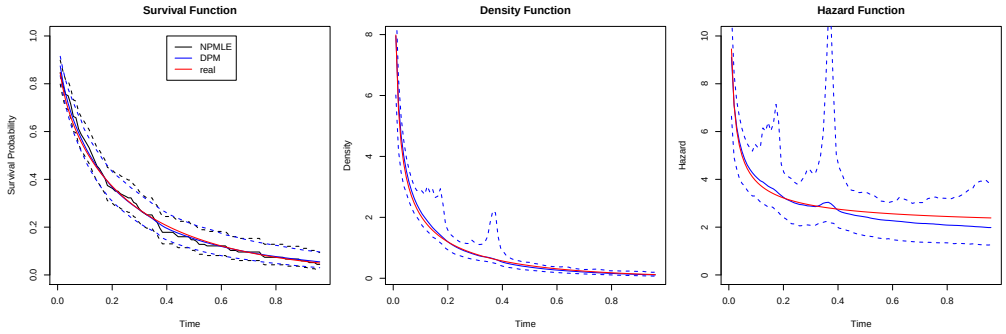

Figure 9: Survival Function, Density Function and Hazard Function Estimates of Gamma(0.5,2)

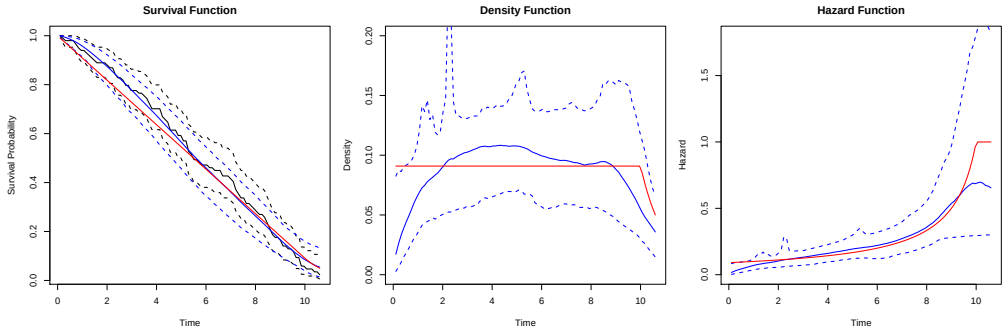

Figure 10: Survival Function, Density Function and Hazard Function Estimates of Unif-Exponential Distribution

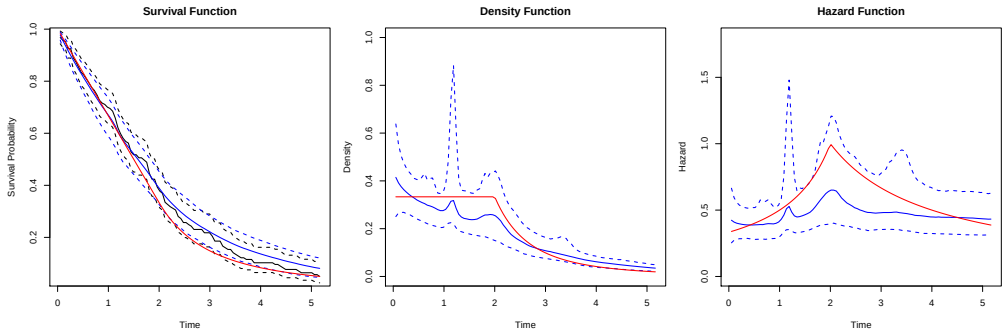

Figure 11: Survival Function, Density Function and Hazard Function Estimates of Unif Pareto(2,2)

vations, 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 appropriately reflecting lack of information there.

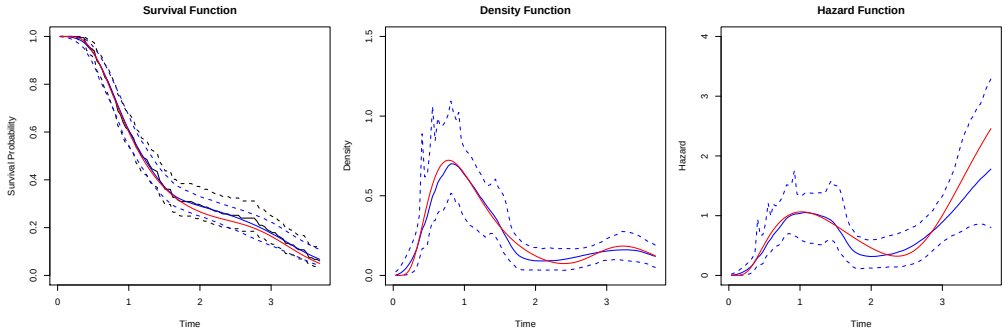

Figure 12: Survival Function, Density Function and Hazard Function Estimates of a mixture of Lognormal distributions

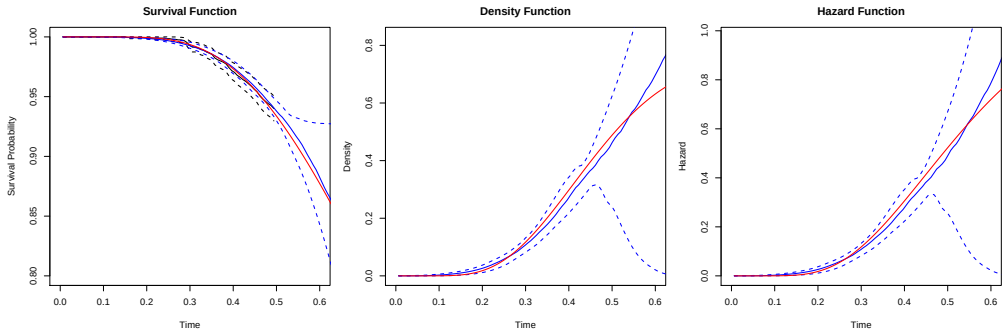

Figure 13: Survival Function Estimates of Heavily Right Censored Data

5 Sensitivity Analysis and Comparison with Empirical Methods

The only information that the LIO prior requires from the investigator is a guess 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.

5.1 Sensitivity Analysis

To evaluate sensitivity of the Gaussian DPM model to scale estimation, we use the 75%, 90%, 95%, 99%, and 99.9% percentiles of the data's underlying distribution as specifications of the upper percentile from the researcher; the median of the distribution is used as the specified median estimate. We consider three underlying 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.

We randomly generated 200 data sets of 100 observations each from each distribution. Figure 14 shows the performance, measured by bias and root MSE (rmse), of the Gaussian DPM model in estimating the cumulative distribution function under a range of scale specifications at the 9 deciles (from 10% to 90%) of the underlying distribution indicated on the horizontal axes. Symbol shapes denote data distributions; colors represent scaling choices. Black color is used for estimates from the empirical cumulative distribution function (ECDF). To make the plot easier to read, we added some horizontal jittering. Bias is slightly worse when using the 75th, 99th, and 99.9th percentiles while rmse is stable across all scaling choices and agrees closely with that of the empirical estimates.

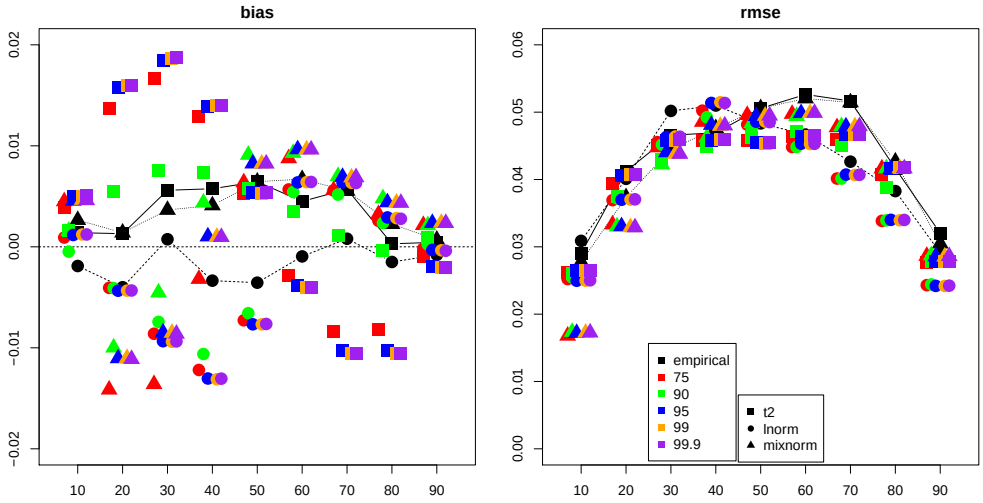

Figure 14: Sensitivity Analysis of Gaussian DPM

To conduct the analysis for DPM of Weibulls, we used 50%, 75%, 90%, 95%, 99% and 99.9% percentiles of the data generating distribution as the possible scale specifications from the user. The data were generated from the four distributions used in

the examples of the previous section. For each distribution, we generated 200 datasets of 100 observations. Right censoring rate was set at 10% while interval censoring of 10% of the observations was accomplished for each dataset by ascribing obervations to fixed intervals. As in the previous plot, Figure 15 uses colors to represent scaling specifications, with black representing frequentist NPMLE results from the R package "survival". Again, each symbol represents a particular data generating distribution, with bias and rmse at the 9 deciles marked on the horizontal axes. The figure clearly indicates that 50^{th} and 75^{th} percentiles give poor results across all deciles. It appears safer to overestimate the 95^{th} percentile than underestimate it for the LIO prior in the Weibull DPM.

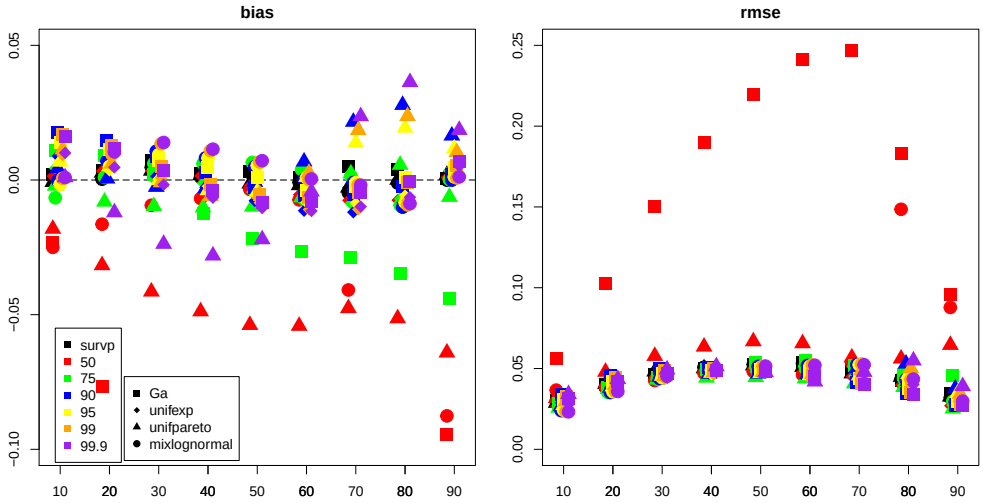

Figure 15: Sensitivity Analysis of Weibull DPM

5.2 Comparison with Empirical Methods

Using the median and 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 for the three specified distributions. Here, we used 200 simulated datasets of 100 or 1000 observations each. Figure 16 shows the results at the deciles of the data's underlying distribution. 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. 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 the 95th percentile of the data generating distribution and compared results 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%

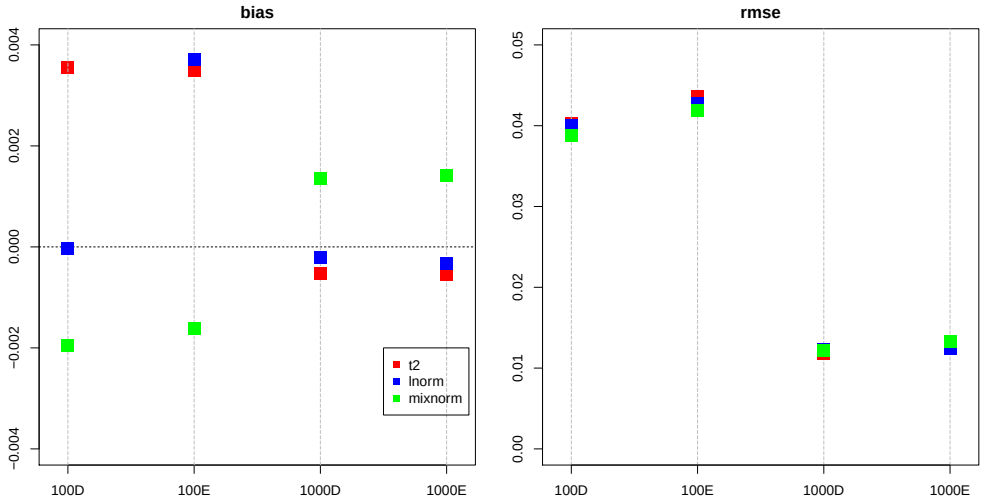

Figure 16: Gaussian DPM Comparison with Empirical CDF

right censoring, 25% interval censoring) and 1000 observations. In Figure 17, 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 preceeding these letters indicate the censoring rate 20 or 50 percent. 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.

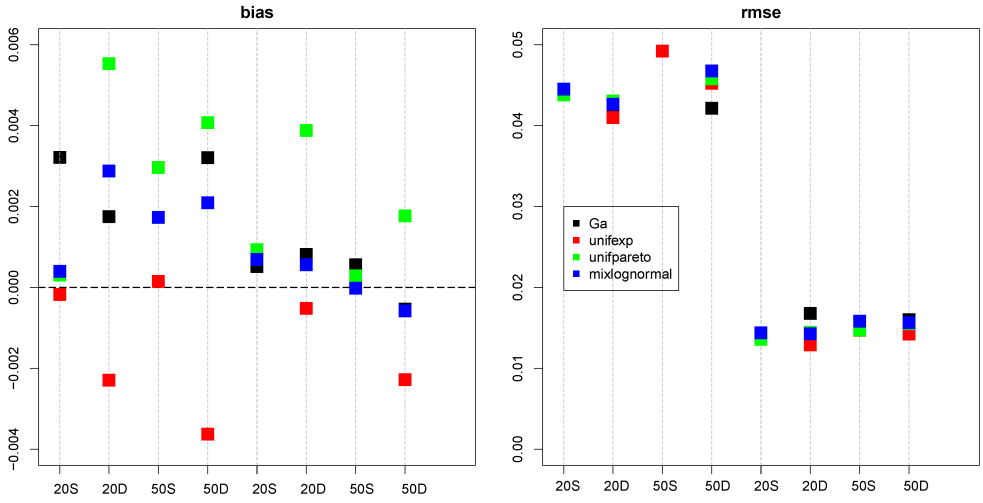

Figure 17: Comparison with Estimates from Survival package

6 Convergence Considerations

6.1 Summary

To summarize the results, here is what has been done in the rest of this write-up. There is Ghoshal style weak-consistency (call it GW, refer to Ghosal et al. (1999) and Wu and Ghosal (2008)) and there is Tokdar style weak consistency (call it TW, refer to Tokdar (2006)). Then there is Ghoshal style strong consistency (call it GS, Wu and Ghosal (2010)). At first we establish equivalence to show that it is sufficient to show strong/weak consistency on the transformed scale for the priors. Then TW is shown for our univariate priors: for the univariate normal, this follows directly from some Tokdar results, for the univariate Weibull, this requires a small amount of work. For multivariate normal, we appeal to GW and we do not show it for the larger TW class. Finally we appeal to GS to show strong consistency for all of our priors - we do not really prove anything for strong consistency but just remark in the end that this holds due to GS results.

6.2 Equivalence Results

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 ∞. 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 $\{X_i\}_{i=1}^\infty$, when each 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 referred to as the 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* $P_{f_0}^\infty$*.*

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

Lemma 1. *Let $Z_i = AX_i + b$, for each $i \in 1, \ldots, n$ be a linear rescaling 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. We begin with the proof for weak posterior consistency. Note that,

$$\widetilde{f_0}(z) = \frac{f_0(Ax+b)}{|A|},$$

where $|A| > 0$ since A is positive definite. For any $\epsilon > 0$, consider the $N_\epsilon^w(f_0)$ neighborhood. Then using the Portmanteau lemma,

$$\begin{aligned}
N_\epsilon^w(f_0) &= \left\{ f \in \mathcal{F} : \left| \int_{x\in\mathbb{R}^p} \phi(x) f(x) dx - \int_{\mathbb{R}^p} \phi(x) f_0(x) dx \right| < \epsilon, \forall \text{ bdd cont } \phi \right\} \\
&= \left\{ f \in \mathcal{F} : \left| \int_{x\in\mathbb{R}^p} \phi(x)|A|^{-1} f(x) dx - \int_{\mathbb{R}^p} \phi(x)|A|^{-1} f_0(x) dx \right| < |A|^{-1}\epsilon, \forall \text{ bdd cont } \phi \right\} \\
&= \{ f \in \mathcal{F} : \left| \int_{z\in\mathbb{R}^p} \phi(A^{-1}(z-b)) f(A^{-1}(z-b)) dz - \int_{\mathbb{R}^p} \phi(A^{-1}(z-b)) f_0(A^{-1}(z-b)) dz \right| \\
&< |A|^{-1}\epsilon, \forall \text{ bdd cont } \phi \} \\
&\equiv \left\{ \widetilde{f} \in \mathcal{F} : \left| \int_{z\in\mathbb{R}^p} \phi(A^{-1}(z-b)) \widetilde{f}(z) dz - \int_{\mathbb{R}^p} \phi(A^{-1}(z-b)) \widetilde{f_0}(z)) dz \right| < |A|^{-1}\epsilon, \forall \text{ bdd cont } \phi \right\}
\end{aligned}$$

where $\equiv$ represents an isomorphism between the sets. Note that,

$$\phi(A^{-1}(z-b)) = \phi(L(z)) = \psi(z) \text{ (say)} ,$$

where L is the linear operator, $L(z) = A^{-1}(z-b), \forall z \in \mathbb{R}^p$. Since L maps from $\mathbb{R}^p \to \mathbb{R}^p$, if ϕ is bounded and continuous, then so is ψ. Therefore,

$$\{\psi : \psi = \phi(L)\} \subseteq \{\phi : \phi \text{ is bounded continuous.}\}$$

However each ϕ is equal to $\phi(L^{-1}(L)) = \phi_1(L)$ (say). Since $L^{-1}(z) = A(z + A^{-1}b)$ is also a linear operator, mapping from $\mathbb{R}^p \to \mathbb{R}^p$, $\phi_1 \in \{\phi : \phi \text{ is bounded continuous }\}$. Therefore, $\phi = \phi_1(L) =$ some $\psi \in \{\psi : \psi = \phi(L)\}$ and,

$$\{\psi : \psi = \phi(L)\} \equiv \{\phi : \phi \text{ is bounded continuous}\}.$$

So the neighborhood,

$$N_\epsilon^w(f_0) = \left\{ \widetilde{f} \in \mathcal{F} : \left| \int_{z\in\mathbb{R}^p} \psi(z) \widetilde{f}(z) dz - \int_{\mathbb{R}^p} \psi(z) \widetilde{f_0}(z)) dz \right| < |A|^{-1}\epsilon, \forall \text{ bdd cont } \psi \right\} \equiv N_\delta^w(\widetilde{f_0}),$$

where $\delta = |A|^{-1}\epsilon$. Then the random variable,

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

$$= \frac{\int_{\widetilde{f} \in N_\delta^w(\widetilde{f}_0)} \prod_{i=1}^n \widetilde{f}(Z_i) d\widetilde{\Pi}(\widetilde{f})}{\int_{\widetilde{f} \in \mathcal{F}} \prod_{i=1}^n \widetilde{f}(Z_i) d\widetilde{\Pi}(\widetilde{f})}$$
$$= \widetilde{\mathcal{Z}}_n(\delta) \text{ (say)}.$$

Since $P_{\widetilde{f}_0}^\infty(\{Z_i\} \in S) = 1 \implies P_{f_0}^\infty(\{X_i\} \in AS + b) = 1$ and since $\widetilde{\mathcal{Z}}_n(\delta) \to 1$ a.s. by the conditions of the lemma, we have that, $\mathcal{X}_n(\epsilon) \to 1$ a.s., which completes the proof for equivalence of weak consistency. The proof of equivalence of strong consistency is similar with change in the type of neighborhood and is omitted. □

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 rescaled 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. We only show the proof for item (4). Others are similar and omitted.

$$\begin{aligned}
\int \|x\|^{2(1+\eta)} f_0(x) dx &= \int \||A|^{-1}(z-b)\|^{2(1+\eta)} f_0(|A|^{-1}(z-b)) dz \\
&= \int \||A|^{-1}(z-b)\|^{2(1+\eta)} \widetilde{f}_0(z) dz \\
&\leq (|A|^{-1})^{2(1+\eta)} \int \|z\|^{2(1+\eta)} \widetilde{f}_0(z) dz + (|A|^{-1} + \|b\|)^{2(1+\eta)} \\
&< \infty.
\end{aligned}$$

□

Earlier work in the literature (Walker, 2004; Choi and Schervish, 2007) contain other slightly different regularity conditions on the true density f_0, for all of which, equivalence can be shown - we avoid a detailed description here for the sake of brevity. In the rest of this exposition we consider results on the rescaled data only, based on the equivalence results derived.

6.3 Consistency results on the rescaled 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). However 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, such as the Cauchy density. 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, thus implying 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. The proof to show that conditions (1)-(4) imply weak consistency in general for a base measure supported on $\mathbb{R}^+ \times \mathbb{R}^+$ is similar to the proof of theorem 3.3 in Tokdar (2006) and is omitted. We show that the Weibull mixture formulation as specified in our article satisfies the base measure conditions (3)&(4) of Tokdar (2006)'s theorem 3.3.

As in remark 3.4 of Tokdar (2006), we have,

$$\begin{aligned} G_0((0,\infty) \times (\alpha^{1-\eta/2}, \infty)) &\propto Ga(\alpha \in (f(\lambda), u^{-(2-\eta)}) | \alpha_\alpha, \lambda_\alpha) \\ &\propto \int_{f(\lambda)}^{u^{-(2-\eta)}} \alpha^{\alpha_\alpha - 1} e^{-\lambda_\alpha \alpha} d\alpha \\ &\leq k_1 \alpha^{-\alpha_\alpha (2-\eta)}, \end{aligned}$$

for some positive scalar k_1. Also following similar arguments as in remark 3.4 of Tokdar (2006),

$$\begin{aligned} 1 - G_0((0,u) \times (0, e^{u^\eta} - \frac{1}{2})) &\leq Ga(\lambda \in (0,u) | \alpha_0, \lambda_0) + Ga(\alpha \in (f(\lambda), e^{u^\eta} - \frac{1}{2}) | \alpha_\alpha, \lambda_\alpha) \\ &\leq k_2 u^{-2\alpha_\alpha}, \end{aligned}$$

for some positive constant k_2 for large enough u. This completes the proof that our proposed Weibull mixture satisfies the additional conditions needed to show weak consistency on a large class of densities, including those without finite second moment. □

The proof of weak consistency for the multivariate case - for our item (3) follows from the results in theorem 2 in Wu and Ghosal (2010). Note that these results also do not permit 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 treatment is fairly involved and does not follow directly from earlier results - a discussion of this will be 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) applied directly to our rescaled data procedure, and by virtue of our equivalence results, to the induced procedure on the observed data. Some weaking 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 involved details omitted here.

7 Discussion

We offer a technique and low information prior specification that can handle data of various scales and demonstrated its value with the mixture of Gaussians model and the mixture of Weibulls model using data simulated from a variety of distributions. To implement the Gaussian DPM model with our prior, we have developed a wrapper for the DPdensity function of the R package DPpackage (**?**) that provides density estimation for scalar and vector-valued random samples.

We illustrated this method of prior specification for DPMs of Gaussian and Weibull distributions. However, a similar approach can be used to obtain a low information prior of mixtures of distributions from 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 for scaled data only needs to be done once and is selected to be vague but computationally reliable. While the LIO prior can be used as a default choice, sometimes substantive prior information is available in the context of the application. To incorporate prior information elicited from the investigator, we might consider this information when choosing what underlying mixture components should be likely to arise in the model. For example, if an expert suggests that most events will happen by the end of a clinical study, a statistician can give higher prior probability to components that have peaks close to 10 and, subsequently, get hyperparameters that are likely to generate those needed components by repeating the steps in the subsection 4.3.

Inspired by De Iorio et al. (2004)'s Dependent Dirichlet process (DDP) of Gaussian mixtures model, we plan to extend DPM of Weibulls model to a DDP regression model for survival data that could directly model event time and handle censoring data. For both De Iorio et al's model and our model, prior specification requires great subtlety. A rescaling approach might be useful for producing a scheme, similar to that offered for inference on a homogeneous population in this paper, for delivering robust inference

with these models.

References

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

Choi, T. and Schervish, M. J. (2007). "On posterior consistency in nonparametric regression problems." *Journal of Multivariate Analysis*, 98(10): 1969–1987.

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.

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

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

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

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.

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.

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.

Kottas, A. (2006). "Nonparametric Bayesian survival analysis using mixtures of Weibull distributions." *Journal of Statistical Planning and Inference*, 136(3): 578–596.

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12(1): 351–357.

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265.

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

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.

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

Walker, S. (2004). "New approaches to Bayesian consistency." *Annals of Statistics*, 32(5): 2028–2043.

Wu, Y. and Ghosal, S. (2008). "Kullback Leibler property of kernel mixture priors in Bayesian density estimation." *Electronic Journal of Statistics*, 2: 298–331.

— (2010). "The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation." *Journal of Multivariate Analysis*, 101(10): 2411–2419.