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

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.

Theorem 1. *Let p 1, ($_i$; T_i) G, G DP(; G_0). Take ($_0$; T_0) G_0. Then E($_i$ ^{k}) = E($_0$ ^{k}) for $k = 1; 2$ and $E(\mathrm{T}_i^k) = E(\mathrm{T}_0^k)$ for $k =$ 1, where* v ^{1} = v *and* v ^{2} = vv^0 *for* v *2* R^p.

Given its parameters $_i$, the distribution of a data point z_i is normal with mean $_i$, 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 $_i$ and T

Having specified v, we have three equations and two constraints for the hyperparameters, requiring $k_T > 2$ and $a > 1$. It is unclear how to choose $k_T; k$, and a exactly; however, smaller values give less informative priors for the corresponding Gamma and Wishart distributed parameters. A choice of $a = 3{=}2$ implies has a scaled ^{2} distribution with 3 degrees of freedom, the minimal integer degrees that give $a > 1:$ Similarly, in the case $p = 1; Wi(k; W)$ is a scaled ^{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 $k = 1$. With $v; a ; k_T;$ and k chosen above, equations (1)-(3) give values for $m ; b ;$ and W , completing the prior specification.

3.3 Hyperparamter Selection for Vector Data

Here again, we need to specify 6 hyperparameters; the only changes are that m is a vector and W

Similar to the univariate case, we set a = 3=2 and k_T = p + 2 and k = p, the minimal integer degrees of freedom that satisfy k_T > p + 1 as required. Then we can obtain m ;b ; and W from equations (3)-(5). Using the fact that $^{2}_{1;}$ = z_1^2 $_{(1\quad)=2}$ for any , it is easy to see that the choice of hyperparameters for the vector data case reduces to the scalar case when p = 1.

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

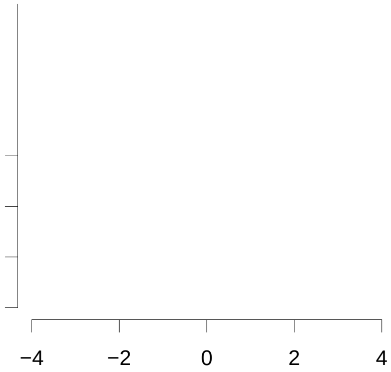

Figure 2: Twenty Prior Predictive Densities from Gaussian DPM with LIO Prior

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 $\begin{matrix}2\\0\end{matrix}$ and covariance matrix $\begin{matrix}1=3 & 1=3\\1=3 & 4=3\end{matrix}$. Figure 5 shows four plots: a scatter

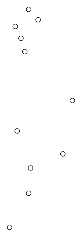

Figure 4: Plots from air quality data

The coverage plot shows whether the true density falls within the 95% pointwise cred-

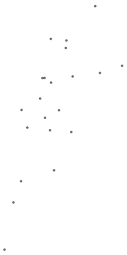

Figure 5: Plots from t_5 / normal mixture

(2006), we fit this model:

$$\begin{aligned}
z_i j \quad _i; \quad _i &\overset{ind:}{} \; Weib(y_i j \quad _i; \quad _i); \quad i = 1; :::; n \\
(\quad _i; \quad _i) jG &\overset{ind:}{} \; G; \quad i = 1; :::; n \\
G & \quad DP(G_0; \quad) \\
G_0 &= Ga(\quad j \quad _0; \quad _0) Ga(\quad j \quad ; \quad) I_{(f(\quad); 1)}(\quad) \\
_0 & \quad Ga(\quad _{00}; \quad _{00}) \\
& \quad Ga(a; b):
\end{aligned}$$

Here again, x Ga

various combinations of (;) resulted in the right panel of Figure 7 with = 0:2 and = 0:1. This completes the hyperparameter selection we recommend for the LIO prior.

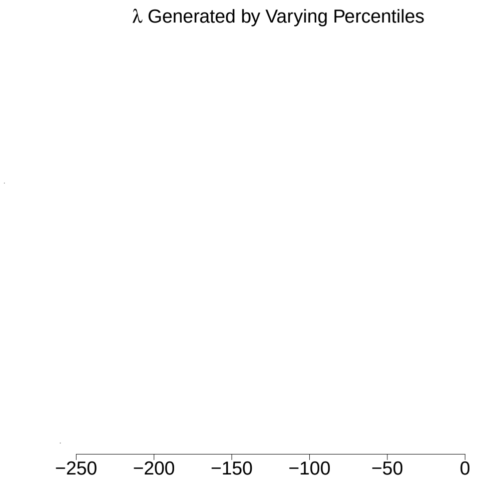

Figure 6: Histogram of the log( )

Figure 7: log( ) and

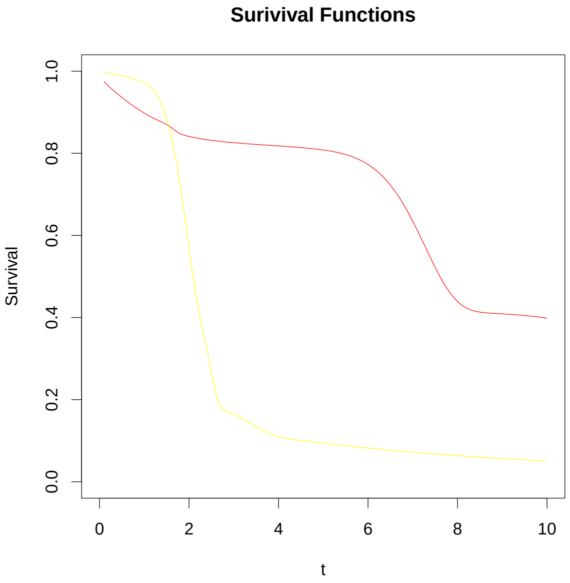

Figure 8: Survival Functions Generated from LIO prior

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 n

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

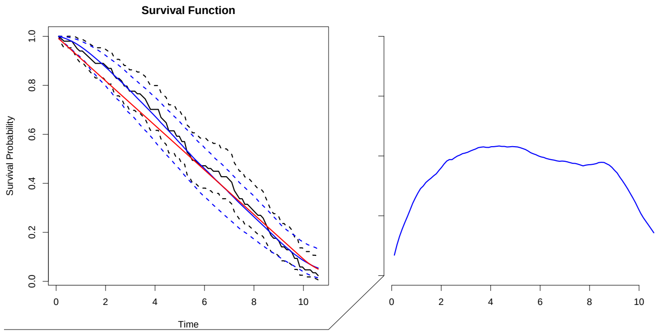

Figure 10: Survival Function, Density Function and Hazard Function Estimates of Unif-Exponential Distribution

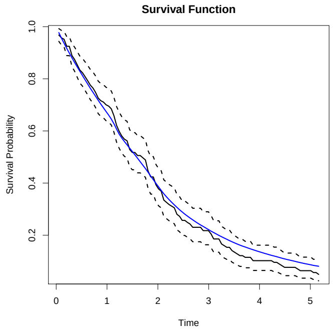

Figure 11: Survival Function, Density Function and Hazard Function Estimates of Unif Pareto(2,2)

vations, 95% right censored at 0:5, generated from the same mixture of log-normals as in

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, $\exp[\text{Normal}(2; 1)]$, representing a skewed distribution;
3. mixnorm : a mixture of two Gaussians, $0.5\ \text{Normal}(0; 1^2) + 0.5\ \text{Normal}(4; 1.5^2)$,

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.

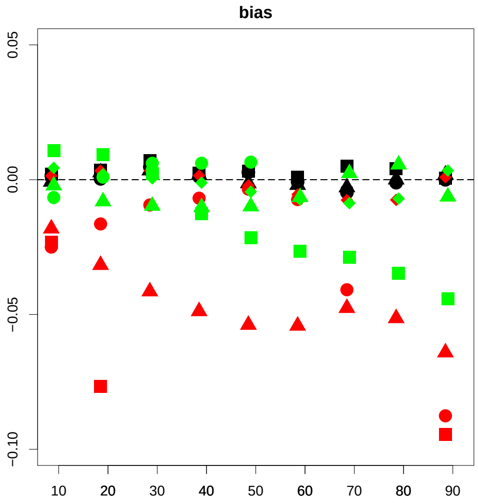

Figure 15: Sensitivity Analysis of Weibull DPM

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%

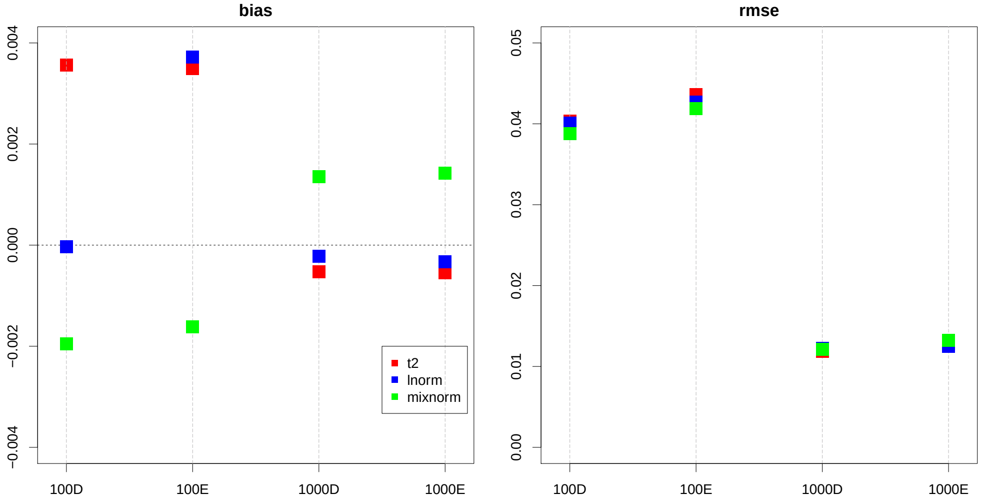

Figure 16: Gaussian DPM Comparison with Empirical CDF

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.

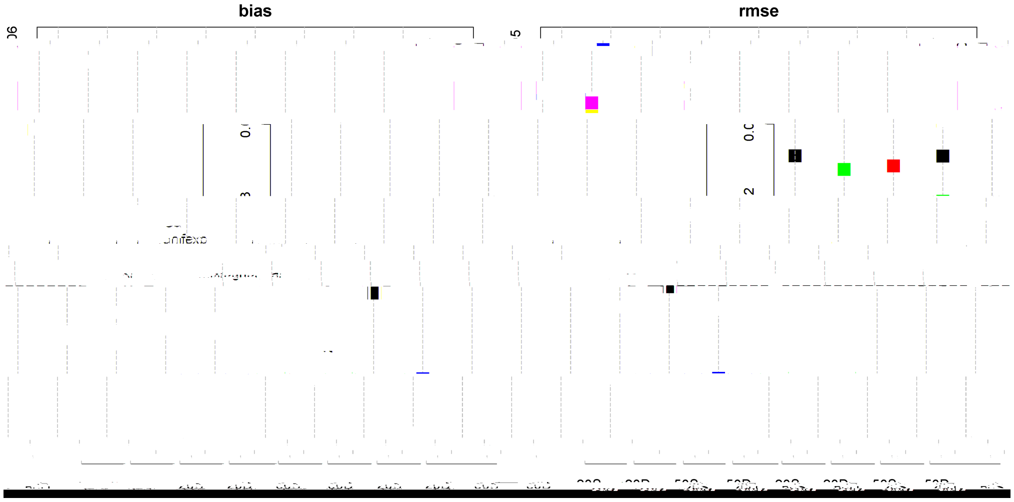

Figure 17: Comparison with Estimates from Survival package

6 Convergence Considerations

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 $\tilde{\Pi}$ achieved weak(strong) posterior consistency at the induced density $\tilde{f}_0$ on $\{Z_i\}_{i=1}^n$.*

Proof. We begin with the proof for weak posterior consistency. Note that,

$$\tilde{f}_0(z) = \frac{f_0(Ax + b)}{|A|};$$

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,

$$N_\epsilon^w(f_0) = \Big\{ f \in \mathcal{F} : \int$$

$$= \frac{\int_{\tilde f \in N^w(\tilde f_0)} \prod_{i=1}^n \tilde f(Z_i)\, d\tilde\Pi(\tilde f)}{\int_{\tilde f \in F} \prod_{i=1}^n \tilde f(Z_i)\, d\tilde\Pi(\tilde f)}$$
$$= \tilde X_n(\ )\ (\text{ say });$$

Since $P^\infty_{\tilde f_0}(\{Z_i\} \in S) = 1 \Rightarrow P^\infty_{f_0}(\{X_i\} \in AS + b) = 1$ and since $\tilde X_n(\ ) \to 1$ a.s. by the conditions of the lemma, we have that, $X_n(\ ) \to 1$ 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. *$\tilde f_0(z)$ is nowhere 0 and is bounded above by M, $\forall z \in \mathbb{R}^p$*
2. *$|\int \tilde f_0(z) \log \tilde f_0(z) dz| < \infty$*
3. *For some $\delta > 0$, $|\int \tilde f_0(z) \log \frac{\tilde f_0(z)}{\phi_\delta(z)} dz| < \infty$, where $\phi_\delta(z) = \inf_{\|t - z\| < \delta} \tilde f_0(t)$*
4. *For some $\eta > 0$, $\int \|z\|^{2(1+\eta)} \tilde f_0(z) dz < \infty$,*

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.

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

□

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

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 in-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 Americanthefop(20b0h0 g -333(9ake342.303 0 Td 577{8898(10):)-334(1969{*

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.