Electronic Journal of Statistics
Vol. 10 (2016) 3355–3373
ISSN: 1935-7524
DOI: 10.1214/16-EJS1158

Transformations and Bayesian density estimation

Andrew Bean, Xinyi Xu and Steven MacEachern*

Department of Statistics
The Ohio State University
1958 Neil Avenue
Columbus, OH 43210
e-mail: bean.243@osu.edu; xinyi@stat.osu.edu; snm@stat.osu.edu

Abstract: Dirichlet-process mixture models, favored for their large support and for the relative ease of their implementation, are popular choices for Bayesian density estimation. However, despite the models' flexibility, the performance of density estimates suffers in certain situations, in particular when the true distribution is skewed or heavy tailed. We detail a method that improves performance in a variety of settings by initially transforming the sample, choosing the transformation to facilitate estimation of the density on the new scale. The effectiveness of the method is demonstrated under a variety of simulated scenarios, and in an application to body mass index (BMI) observations from a large survey of Ohio adults.

Received January 2016.

1. Background

We study the problem of estimating an unknown continuous density f on the real line. A popular Bayesian approach models the unknown density using Dirichlet Process Mixtures (DPM) (c.f. [3], [10]). The prior for f is constructed by convolving a continuous kernel, frequently Gaussian, with a Dirichlet Process (DP) (c.f. [2]) distributed mixing distribution G. Many variations on this basic structure have been proposed, one of the simplest being the location mixture of Gaussian kernels given by

$$\begin{aligned} G &\sim \mathrm{DP}\big(MG_0\big), \\ \mu_i \big| G &\overset{iid}{\sim} G \\ X_i \big| \mu_i, \sigma &\overset{indep}{\sim} \mathrm{N}\big(\mu_i, \sigma^2\big) \qquad\qquad i = 1, \ldots, n \end{aligned} \tag{1}$$

where the base distribution G_0 is often taken to be a normal distribution with fixed parameters, and an inverse-gamma prior for the kernel variance σ^2 completes the model. Although more complex structures have been proposed, even this basic construction is known to be extremely flexible, having KL support on a wide class of continuous distributions [5].

*The authors would like to thank the Associate Editor and Referee for comments that improved the paper. This work was supported in part by the NSF under award number DMS-1209194. The views expressed in this paper are not necessarily those of the NSF.

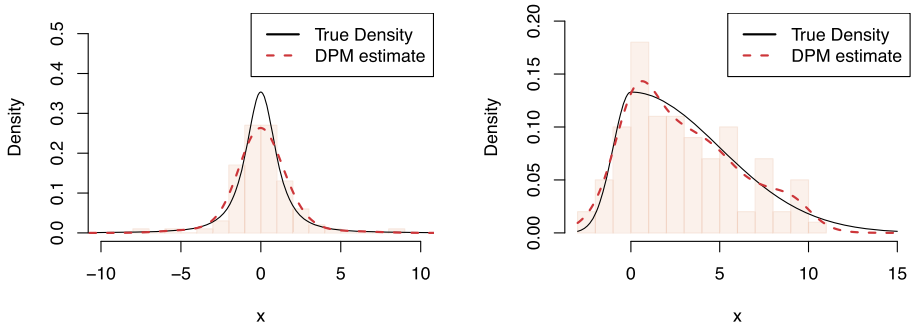

FIG 1. *DPM density estimates (dashed lines) based on samples of size 100 for two examples of the two-piece distributions (the true densities are shown in solid black). The leftmost density is symmetric, but has t_2 tails. The rightmost density has Gaussian tails, but is right-skewed.*

However, despite these models' flexibility, a fact long recognized in classical kernel density estimation is pertinent in the Bayesian context: some densities f are easier to estimate than others. In particular, densities f which possess heavy tails or severe skew can cause difficulties for estimation when the sample size is small, or even moderate. Applying (1) directly to estimate such skewed and heavy-tailed samples gives estimates that tend, qualitatively, to be overly bumpy or unsmooth in the tails. Quantitatively, the error for estimating such densities is inflated under a variety of metrics on the space of distributions, as will be shown in our simulation and data examples. For an understanding of these difficulties, consider the two examples shown in Figure 1. The plots show estimates of two so-called two-piece distributions, which were studied, for example, in [4] and [12]. The left plot is a (symmetric but heavy-tailed) Student t distribution with 2 degrees of freedom, while the right plot is a (light-tailed but skewed) split normal distribution where the scale parameter is larger above the median by a factor of 5. In both cases, the DPM estimates leave much to be desired. At left, in the heavy tailed plot, we can see that the DPM model underestimates kurtosis. At right, in the skewed case, the estimated density is bumpy, and appears to underestimate the degree of asymmetry seen in the true density.

To alleviate this problem, in this paper we propose a Transformation DPM (TDPM) method, which carefully selects a series of transformations from a parametric family $\{\varphi_\theta : \theta \in \Theta\}$ that is designed to symmetrize and shorten the tails of the density, so that the density of the transformed sample $Y_1, \ldots, Y_n$ is easier to estimate than the density of the original observations $X_1, \ldots, X_n$, in a sense made specific in Section 2. We fit the model (1) on this transformed scale, produce a DPM estimate $\hat{f}_Y$ of the density of the Y_i on the transformed scale, and then back-transform for an estimate $\hat{f}_X$ of the density of the original X_i, that is,

$$\hat{f}_X = \left(\hat{f}_Y \circ \varphi_{\hat{\theta}}\right) \cdot \varphi'_{\hat{\theta}}. \tag{2}$$

Figure 2 illustrates the application of the TDPM procedure to estimating the two-piece densities. It is clear that for both cases, the TDPM estimates are

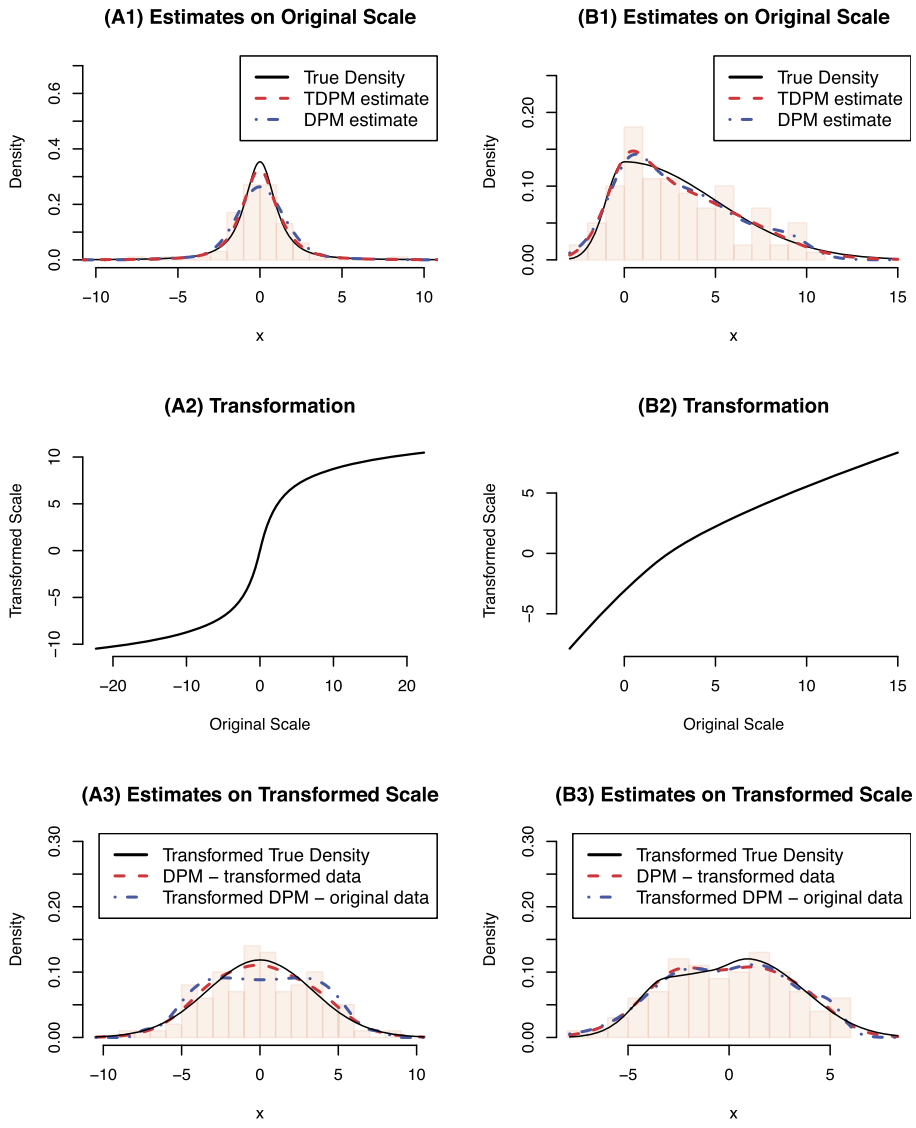

FIG 2. *Illustration of the Transformation-DPM technique. The heavy-tailed sample (left column, A1-A3) and skewed sample (right column, B1-B3) of figure 1 are transformed according to the symmetrizing and tail-shortening transformations of section 2. The DPM model is fit to the transformed samples in the bottom panels, then back-transformed to give the TDPM estimate on the original scale.*

closer to the true densities than the direct DPM estimates. This transformation-density-estimation technique has been investigated in the context of kernel density estimation and are shown to be effective for producing improved density estimates (c.f. [17] and [19]). Recently, [8] proposed to improve manifold learn-

ing and cluster detection through transformations with nonparametric warping functions. Our TDPM method, in contrast, utilizes a low dimensional parametric transformation, with the focus of minimizing a statistic that measures the ease with which a density can be estimated.

The rest of the paper is structured as follows. In Section 2, we describe a new family of transformations $\{\varphi_\theta : \theta \in \Theta\}$ that is rich enough to correct problems with both skew and heavy tails, and a method for selecting a series of transformations from the family. In Section 3, we discuss the use of such transformations in combination with the DP mixture model, and demonstrate the effectiveness of the method, under a variety of simulated scenarios, at reducing error for estimating the original f_X. In Section 4, the method is used to estimate and compare the distributions of body mass index across groups of individuals from the 2008 Ohio Family Health Survey. Finally, Section 5 concludes with a discussion of the contributions and of future work.

2. Transformations

Several authors [17, 13, 19] have considered the use of parametric transformations for reducing bias of kernel density estimates (KDEs), motivating their transformations through asymptotics of the kernel estimators. Because of the role of (1) as a Bayesian analogue of the Gaussian KDE [1], and the connections between KDE asymptotics and the asymptotic behavior of posteriors from DP mixture models [6], we hypothesize that these authors' transformation-density estimation methods are also relevant for DP mixture estimation, and that much of their work concerning transformation selection translates well to the setting of Bayesian density estimation.

In studying the transformation density estimation technique, we must address two central questions:

1. How to define an appropriate family of transformations $\{\varphi_\theta : \theta \in \Theta\}$, and
2. How to select a transformation $\varphi_{\hat{\theta}}$ from this family based on a sample?

The answer to the first question of course depends on the types of densities one wishes to estimate. A quick review of work on transformation kernel density estimation yields some ideas. [17] seeks to estimate skewed densities whose support is bounded below. They suggest a signed-power family of transformations which maps ranges $[c, \infty)$ to $\mathbb{R}$. [13] estimates kurtotic densities on $\mathbb{R}$ by first applying a kurtosis reducing transformation. Yang and Marron ([19]) suggest that for some densities, a second or third round of transformations can further improve the quality of kernel estimates. They consider an ensemble of three parametric Johnson transformations, and develop an iterative method for choosing between these families.

Bayesian methods for density estimation differ from classical methods in an essential way: they are driven by the likelihood. Consequently, natural evaluations of the success (or failure) of the methods in simulation settings is tied to likelihood, and, for real-data examples, to out-of-sample predictive performance.

Thus, to avoid a zero in the likelihood, it is essential to get the right support for the density, and we consider new families of transformations which preserve the support. For skew correction, we choose an alternative to the families of [17] and [19] that, unlike those transformations, maps $\mathbb{R}$ to $\mathbb{R}$. For kurtosis reduction, we forego the specific transformations in [13] and [19] in favor of a simple cdf-inverse-cdf transformation whose parameter can be directly related to the extreme tail index of the original density. All of the transformations we consider are strictly monotonic. The families we employ are detailed in Section 2.1.

To answer the second question, our approach is to develop a single statistic measuring the ease with which a density can be estimated. This statistic should be quickly and easily computable for a large collection of candidate transformations. In their work on transformation kernel density estimation, [13], [17], and [19] take this approach, and those authors agree on the form of the statistic. For each candidate transformation $Y_{\theta,i} = \varphi_\theta(X_i)$, one can compute an estimate of the criterion

$$L(\theta) = \sigma_{Y,\theta} \left[\int \left(f''_{Y,\theta}(y) \right)^2 dy \right]^{1/5}, \tag{3}$$

where $f_{Y,\theta}$ is the transformed density. This criterion is motivated through study of the L_2 asymptotics of kernel density estimates; larger values of $L(\theta)$ indicate that $f_{Y,\theta}$ is more difficult to estimate. In practice, estimation of (3) requires estimation of the integrated curvature $R(\theta) = \int \left(f''_{Y,\theta}(y) \right)^2 dy$, a problem which is well studied because of its role in "plug-in" bandwidth selectors for kernel density estimates [14]. Like [13], [17], and [19], we select transformations according to a kernel estimate $\hat{L}(\theta)$ of (3). The criterion (3) is motivated, and the estimator described in detail, in Section 2.2.

2.1. Family of transformations

We consider a collection $\{\varphi_\theta : \theta \in \Theta\}$ of transformations consisting of two parametric families: a skew-correcting transformation due to Yeo and Johnson [20], and a kurtosis-reducing Student-t cdf-inverse-Gaussian-cdf transformation.

These parametric transformation families, which will be described shortly, are appropriate for samples centered near 0 and appropriately scaled. For this reason, the original $\mathbf{X} = (X_1, \ldots, X_n)$ are first centered and scaled according to the sample median $m(\mathbf{X})$ and interquartile range $I(\mathbf{X})$, giving

$$\tilde{X}_i = \tilde{h}(X_i) = \frac{X_i - m(\mathbf{X})}{I(\mathbf{X})/5}.$$

This rescaling sets $I(\tilde{\mathbf{X}}) = 5$ as the desired interquartile range of $\tilde{X}_1, \ldots, \tilde{X}_n$. This constant affects the achievable shapes of Yeo-Johnson transformations (4); simulations suggest samples with an IQR of 5 allow effective estimation of the Yeo-Johnson parameter.

Selection of the appropriate transformation family proceeds with the rescaled data. After centering and scaling according to $\tilde{X}_i = h(X_i)$, a parametric trans-

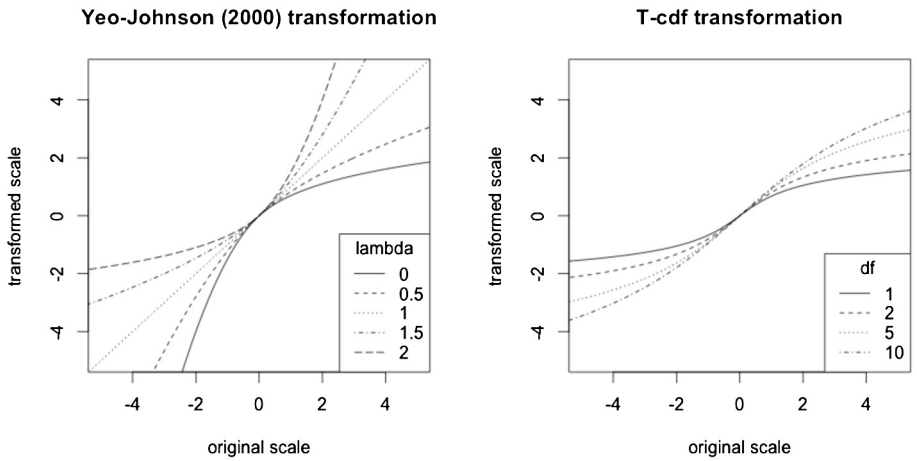

FIG 3. *The families $\tilde{\varphi}_{(1,\theta_1)}$ of Yeo-Johnson transformations and $\tilde{\varphi}_{(2,\theta_2)}$ of t-to-Normal cdf transformations.*

formation $\tilde{\varphi}_\theta$ is applied to the centered and scaled data to give the final transformation

$$Y_i = \varphi_\theta(X_i) = \big(\tilde{\varphi}_\theta \circ \tilde{h}\big)(X_i).$$

We search for an appropriate parametric transformation $\tilde{\varphi}_\theta$ over a parameter $\theta \in \Theta = \{(j, \theta_j) : j \in \{1,2\}, \theta_j \in \Theta_j\}$, where $j \in \{1,2\}$ indexes families of transformations with parameters θ_j.

The family $j = 1$ corrects excessively skewed distributions by employing the Yeo-Johnson power transformations ([20]) given by

$$\tilde{\varphi}_{(1,\theta_1)}(x) = \begin{cases} \big((x+1)^{\theta_1} - 1\big)/\theta_1 & x \geq 0,\ \theta_1 \neq 0 \\ \log(x+1) & x \geq 0,\ \theta_1 = 0 \\ -\big((-x+1)^{2-\theta_1} - 1\big)/(2-\theta_1) & x < 0,\ \theta_1 \neq 2 \\ -\log(-x+1) & x < 0,\ \theta_1 = 2, \end{cases} \tag{4}$$

where the real-valued parameter θ_1 is restricted to $\Theta_1 = [0, 2]$. See the left panel of Figure 3 for the shapes of the Yeo-Johnson transformations. These are closely related to the famous Box-Cox family and the signed power transformations of [17]. The Yeo-Johnson family is appropriate for correcting both left and right skew, with $\theta_1 > 1$ and $\theta_1 < 1$, respectively. Setting $\theta_1 = 1$ gives the identity transformation. The family possesses a symmetry property that $\tilde{\varphi}_{(1,\theta_1)}(x) = -\tilde{\varphi}_{(1,2-\theta_1)}(x)$.

For excessively kurtotic distributions, we propose a family of tail shortening cdf-inverse-cdf transformations,

$$\tilde{\varphi}_{(2,\theta_2)}(x) = \Phi^{-1}\big(T_{\theta_2}\big(x/b_{\theta_2}\big)\big), \tag{5}$$

where Φ and T_{θ_2} are, respectively, the cdfs of a standard normal and a t distribution. The degrees of freedom parameter $\theta_2 > 0$ of the t cdf controls the severity of

the kurtosis reduction. The rescaling constant $b_{\theta_2} = 5/\big(T_{\theta_2}^{-1}(0.75) - T_{\theta_2}^{-1}(0.25)\big)$ is required to again rescale the input sample $\tilde{X}_1, \ldots, \tilde{X}_n$ from their interquartile range of 5 to match the IQR of a t_{θ_2} distribution.

2.2. A criterion for estimating transformation parameters

The criterion $L(\theta)$ given in (3) can be motivated as a target function for guiding the transformation θ by studying of the asymptotics of the Gaussian kernel density estimator

$$\hat{f}_{Y,h_\theta}(y) = \frac{1}{n}\sum_{i=1}^{n} \phi_{h_\theta}(y - Y_{\theta,i}) \tag{6}$$

of the transformed density $f_{Y,\theta}$. The kernel in this classic estimator is $\phi_h(u) = (2\pi)^{-1/2}h^{-1}e^{-u^2/h}$. In this section, we review the asymptotic considerations of (6) that give rise to the criterion $L(\theta)$ in (3), a data-based estimator $\hat{L}(\theta)$ of L, and finally, a selection for $\hat{\theta} \in \Theta$.

Provided $f_{Y,\theta}$ possesses two bounded, continuous derivatives, the mean integrated squared error

$$\mathrm{MISE}(\theta) = \int \big(\hat{f}_{Y,h_\theta}(y) - f_{Y,\theta}(y)\big)^2\, dy \tag{7}$$

of the KDE $\hat{f}_{Y,h_\theta}$ admits an expansion that is quite standard in the kernel density estimation literature (c.f. [16]),

$$\mathrm{MISE}(\theta) = \mathrm{AMISE}(\theta) + o\Big(\frac{1}{nh_\theta} + h_\theta^4\Big), \quad n \to \infty,\ h_\theta \to 0,\ nh_\theta \to \infty. \tag{8}$$

These asymptotics allow different bandwidth selections $\hat{h}_\theta$ for each possible transformed sample $Y_{\theta,1}, \ldots, Y_{\theta,n}$, but require, for each θ, that the bandwidth shrinks to 0 more slowly than n^{-1}. For large n, the L_2 error of $\hat{f}_{Y,h_\theta}$ depends chiefly on the asymptotic mean integrated squared error

$$\mathrm{AMISE}(\theta, h_\theta) = \frac{R(\phi)}{nh_\theta} + \frac{1}{4}h_\theta^4 R(f''_{Y,\theta}), \tag{9}$$

where R is the curvature functional

$$R(f''_{Y,\theta}) = \int (f''_{Y,\theta}(y))^2\, dy \tag{10}$$

of the transformed density $f_{Y,\theta}$. This suggests a strategy of choosing both the transformation parameter θ and the bandwidth h_θ to reduce $\mathrm{AMISE}(\theta, h_\theta)$. For a given $R(f''_{Y,\theta})$, the bandwidth minimizing the AMISE is

$$h_\theta^* = C_1(\phi)\big(R(f''_{Y,\theta})\big)^{-1/5} n^{-1/5}.$$

Plugging h^*_θ into (9), we find that for each θ, the minimum AMISE over possible bandwidths $h_\theta > 0$ is

$$\text{AMISE}^*(\theta) = \text{AMISE}(\theta, h^*_\theta) = C_2(\phi)\left[\int \left(f''_{Y,\theta}(y)\right)^2 dy\right]^{1/5} n^{-4/5},$$

which depends on θ only through a positive power of the curvature $R(f''_{Y,\theta})$. Hence, the transformation θ which minimizes $R(f''_{Y,\theta})$ also minimizes the AMISE (9).

There is a technical difficulty that, like the MISE itself, the curvature $R(f''_{Y,\theta})$ is not scale invariant. For example, standardizing $f_{Y,\theta}$ by its mean $\mu_{Y,\theta} = \text{E}_{f_X}(\varphi_\theta(X))$ and standard deviation $\sigma_{Y,\theta} = \text{Var}^{1/2}_{f_X}(\varphi_\theta(X))$, and considering the density $f_{Z,\theta}(z) = \sigma_{Y,\theta} f_{Y,\theta}(\sigma_{Y,\theta} z + \mu_{Y,\theta})$ of $Z = (Y - \mu_{Y,\theta})/\sigma_{Y,\theta}$, one finds that $R(f''_{Z,\theta}) = \sigma^5_{Y,\theta} R(f_{Y,\theta})$. This is an unacceptable property if we are to use R as a criterion for transformations.

Standardizing R for scale yields the criterion (3). $L(\theta)$ is used as a target function in the transformation density estimation schemes of [13], citeWand1991, and [19]. Transformations $\theta \in \Theta$ near the L-optimal $\theta^* = \text{argmin}_{\theta\in\Theta} L(\theta)$ lead to densities for which the global-bandwidth normal KDE $\hat{f}_{Y,h_\theta}$ in (6) will incur smaller L_2 errors for estimating the true $f_{Y,\theta}$.

In practice, of course, L cannot be evaluated, so θ^* is unknown. To concoct an estimate $\hat{\theta}$, we adopt the strategy of [13], [17], and [19] of first developing a kernel estimate $\hat{L}(\theta)$ of (3), and taking $\hat{\theta} = \text{argmin}\hat{L}(\theta)$. The estimates are of the form

$$\hat{L}(\theta) = \hat{\sigma}_{Y,\theta}\left[\hat{R}(f''_{Y,\theta})\right]^{1/5}, \tag{11}$$

where $\hat{\sigma}^2_{Y,\theta} = (n-1)^{-1}\sum(Y_{\theta,i} - \overline{Y}_\theta)^2$, and the chief difficulty is choosing an estimator $\hat{R}(f''_{Y,\theta})$ of the integrated curvature $R(f_{Y,\theta})$ in (10).

Several estimators of $R(f''_{Y,\theta})$ have been proposed. [9] suggests the popular "diagonals-in" choice

$$\hat{R}_2 = n^{-2}b^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{(4)}\left(b^{-1}(Y_{\theta,i} - Y_{\theta,j})\right), \tag{12}$$

derived from a further kernel estimate

$$\tilde{f}_{Y,b} = n^{-1}b^{-1}\sum_{i=1}^{n} K\left(b^{-1}(y - Y_i)\right), \tag{13}$$

where b is a bandwidth and K is a kernel for estimating $R(f''_{Y,\theta})$, not to be confused with h and ϕ in (6). Sheather and Jones ([14]) give a rule for selecting b of the form

$$b_{SJ} = C(f_{Y,\theta})D(K)n^{-1/7} \tag{14}$$

by setting the asymptotically dominant terms in the bias of (12) to zero and solving the resulting equation. The constant $C(f_{Y,\theta})$ involves higher order integrated squared derivatives of $f_{Y,\theta}$, which [14] suggests estimating in another

similar stage. An implementation of the bandwidth selector and curvature estimator (12) is given by the R package KernSmooth [15].

However, the strategy of [14] using (12) requires at least $2n^2$ operations, which is too costly for our setting, in which we would like to compute (12) for many candidate values of θ. We take a simple numerical approach to estimating the integrated curvature of the kernel approximation (13),

$$R(\tilde{f}''_{Y,b_{SJ}}) = \int \left(\tilde{f}''_{Y,b_{SJ}}(y)\right)^2 dy.$$

We use a bandwidth b_{SJ} selected according to the Sheather-Jones rule (14), and a normal kernel $K = \phi$. The integral is approximated over an equally spaced grid of 1000 points covering 20 sample standard deviations. The result is taken as $\hat{R}(f''_{Y,\theta})$ in (11). [18] gives some asymptotic results concerning properties of $\hat{L}(\theta)$ as an estimator of $L(\theta)$, and $\hat{\theta}$ of θ^*. The trials we conduct in the simulations and data analysis sections are concerned less with accuracy of $\hat{\theta}$ for θ^*, but more importantly with the accuracy of density estimation. To suit this purpose, the $\hat{L}$-optimal transformation performs quite well.

In Section 2.3, we describe rules for selecting, in an iterative fashion, a series of transformations from among the families described in Section 2.1.

2.3. Iterative transformation selection

As demonstrated by [19], we note that some difficult densities may benefit from multiple transformations applied in sequence. In particular, densities which are both heavy tailed and skewed may require application of both transformation families $\varphi_{(1,\theta_1)}$ and $\varphi_{(2,\theta_2)}$ described in Section 2.1.

We now describe a method for selecting, on the basis of the statistic $\hat{L}(\theta)$, a series of transformations

$$\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(K)} \in \Theta = \{(j, \theta_j) : j \in \{1, 2\},\ \theta_j \in \Theta_j\},$$

chosen from the families (j, θ_j) giving transformed sample values

$$Y_{\hat{\theta},i} = \left(\varphi_{\hat{\theta}^{(K)}} \cdots \circ \varphi_{\hat{\theta}^{(2)}} \circ \varphi_{\hat{\theta}^{(1)}}\right)(X_i). \tag{15}$$

In the simulations and data analyses of the next sections, we set Θ_1 and Θ_2 to be equally spaced grids over reasonable parameter values for the skew- and tail-transformations. For the Yeo-Johnson family (4), Θ_1 is set as a grid of 1000 equally spaced values between 0 and 2. For the cdf-inverse-cdf transformation, the set Θ_2 consists of 1000 possible values, equally spaced on the inverse scale between 1 and 20. In total, this yields 2000 candidate transformations plus the identity.

The total number K of transformations that should be applied for a given density is of great interest. The procedure should be sensitive enough to give a number K of transformation so that

$$f_{Y,\theta}(y) = \left(f_X \circ \varphi^{-1}_{\theta^{(1)}} \circ \cdots \circ \varphi^{-1}_{\theta^{(K)}}\right) \cdot \prod_{l=1}^{K} \left(\varphi^{-1}_{\theta^{(l)}}\right)'$$

has a small value of $L(\theta)$. However, it should not be too aggressive in suggesting transformations of already-easy densities such as the normal, nor should it apply transformations that do not give a substantial reduction in $\hat{L}$.

To control the sensitivity of the procedure, we simulate $10,000$ standard normal samples of size n, evaluate $\hat{L}$ for these, and continue applying transformations only so long as $\hat{L}(\theta)$ satisfies the two conditions: (1) $\hat{L}$ exceeds the sample $1-\alpha$ quantile $\hat{L}_{n,1-\alpha}$ of the 10,000 simulated normal $\hat{L}$ values for some small α, which we set to be $\alpha = 0.1$, and (2) the $\hat{L}$-optimal transformation reduces $\hat{L}$ by more than a minimum percentage, which we set to be 5%. Thus, at stage k, given the current transformed version of the data $X_1^{(k)}, \ldots, X_n^{(k)}$, the procedure finds the optimal transformation over the grids Θ_1 and Θ_2 of skew- and tail-transformation parameter values. Let $\hat{\theta}^{(k+1)}$ and $\hat{L}^{(k+1)}$ denote the minimizer and the minimum, respectively. If either

1. $\hat{L}^{(k+1)} \geq 0.95\hat{L}^{(k)}$, or
2. $\hat{L}^{(k+1)} \leq \hat{L}_{n,1-\alpha}$,

the transformation $\theta^{(k)}$ is not applied: we set $K = k$, and proceed with density estimation with the sample values (15). Otherwise, $\theta^{(k+1)}$ is applied, and another round of transformations is proposed.

3. Simulation study

3.1. Simulation design

We design a simulation study to investigate the utility of the transformation method in estimating densities of varying skewness and tail heaviness. We consider estimating two-piece densities of the form,

$$f_X(x) = \frac{2}{\sigma_1 + \sigma_2}\left[g\Big(\frac{x-\mu}{\sigma_1}\Big)I_{(-\infty,\mu)}(x) + g\Big(\frac{x-\mu}{\sigma_2}\Big)I_{(-\infty,\mu)}(x)\right], \tag{16}$$

where g is a symmetric density from a location-scale family, and $\sigma_1, \sigma_2 > 0$ are distinct scale parameters for the regions above and below the median μ of f_X. The parent density g controls the tail behavior of f_X, while the ratio of σ_1 to σ_2 controls the degree of skewness. Parametric estimation of these densities has been studied, for example, by [4] and [12]. In the simulations detailed here, we fix $\mu = 0$ and $\sigma_1 = 1$. We then study cases where $\sigma_2 = 1$, for symmetric distributions, and where $\sigma_2 = 5$, for right-skewed distributions. Additionally, we study two choices for the parent family g: a standard normal density and a heavy-tailed t density with 2 degrees of freedom. The resulting four densities are depicted in the leftmost column of Figure 4. From each of the four two-piece densities and each of three different sample sizes $n = 100, 200$ and 500, we perform 20 replicates of the simulation.

For each sample, we compare the direct DPM density estimate — found by fitting the basic DPM model (1) to the untransformed X_i so that $\hat{f}_X$ is the

posterior predictive density — to the Transformation DPM density estimate (2), with the transformation selected as described in Section 2, so that

$$\hat{f}_X = \left(\hat{f}_Y \circ \varphi_{\hat{\theta}}\right) \cdot \varphi'_{\hat{\theta}},$$

where $\hat{f}_Y$ is the posterior predictive density of the DPM (1) applied on the transformed scale to the Y_i's. In addition, we fit Griffin's modified DPM model, which can be represented by

$$\begin{array}{rcll} G & \sim & \mathrm{DP}(MG_0) & \\ \mu_i \big| G & \overset{iid}{\sim} & G & \\ X_i \big| \mu_i & \overset{indep}{\sim} & \mathrm{N}\Big(\mu_i,\, a\frac{\zeta_i}{\mu_\zeta}\sigma^2\Big) & \qquad i = 1, \ldots, n \end{array} \tag{17}$$

where $a \sim \mathrm{Beta}(\alpha, \beta)$, $G_0 = \mathrm{N}(\mu_0, (1-a)\sigma^2)$, ζ_i are iid from an inverse-gamma distribution with mean μ_ζ, and σ^2 is given an inverse-gamma prior. This model was proposed by [7] as an improved procedure over the basic DPM model (1) for density estimation, and is used as a benchmark in our simulation studies.

Computationally, the MCMC for the Bayesian methods is the most demanding component for CPU time. The basic location mixture (1) is a conjugate-style model, and is fit with a Gibbs sampler over a collapsed sample space, along the lines of Algorithm 3 of [11]. Griffin's model is non-conjugate; for this model, we implement the MCMC strategy suggested by the author in Appendix A of [7]. In comparison, estimation of the transformation is far less computationally demanding. Consider one example of a sample of size $n = 500$ from the skewed and heavy-tailed density. Using one core of a 2×Twelve Core Xeon E5-2690 v3 / 2.6GHz / 128GB machine, the sequence of transformations can be completed in 3.09 minutes, while 5000 iterations of MCMC for the basic DPM model requires 27.44 minutes.

3.2. Simulation results

For comparing point estimates to the truth, we employ the Hellinger distance, expressed here for a real-valued variate,

$$d_H(p, q) = \left(\int_{\mathbb{R}} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx \right)^{1/2}.$$

Hellinger distance is a useful metric for quantifying the distance between distributions p and q. Bayesian methods construct the point estimates using MCMC approximations to the posterior predictive density given the observations. For each estimate $\hat{f}_X$, we evaluate the Hellinger distance $d_H(\hat{f}_X, f_X)$ between the point estimate and the true two-piece density. Although the numerical results are presented only under the Hellinger distance, other metrics such as the total variation distance and the Kullback-Leibler distance have also been evaluated, and the results are similar under these alternative metrics.

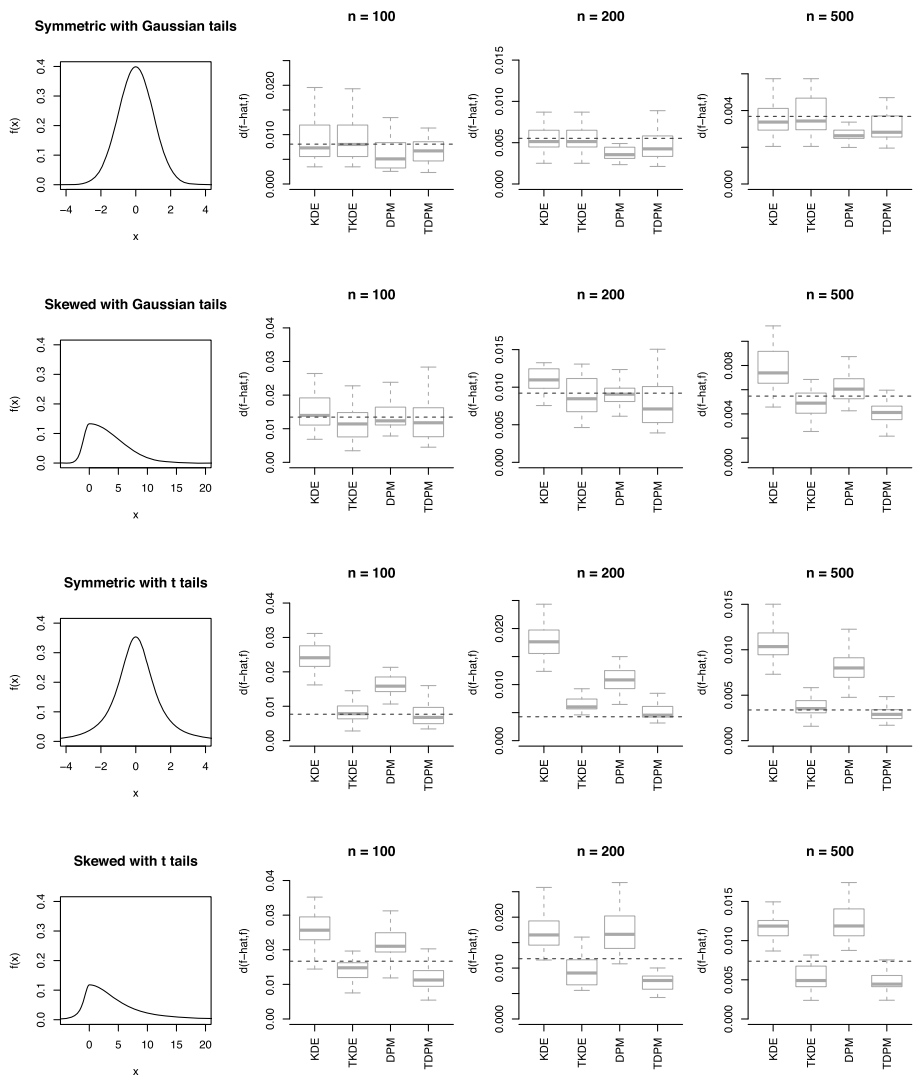

FIG 4. *Boxplots give a comparison of Hellinger error for four density estimates, KDE, TKDE, DPM, and TDPM. Each row represents a different two-piece scenario, and results are shown for 20 replicate samples from each scenario and sample size. The horizontal dotted line represents the median Hellinger distance obtained by fitting Griffin's (2010) model without transformation to those 20 samples.*

Figure 4 gives a comparison of the Hellinger error for the Transformation DPM (TDPM) density estimates, direct DPM density estimates, transformation KDE (TKDE), and direct KDE, across all simulation settings. The horizontal dotted lines represent the median of Hellinger distance for direct fits of Griffin's model (17). These numerical results again suggest that the TDPM approach gives improved estimates for skewed or heavy-tailed distributions. For

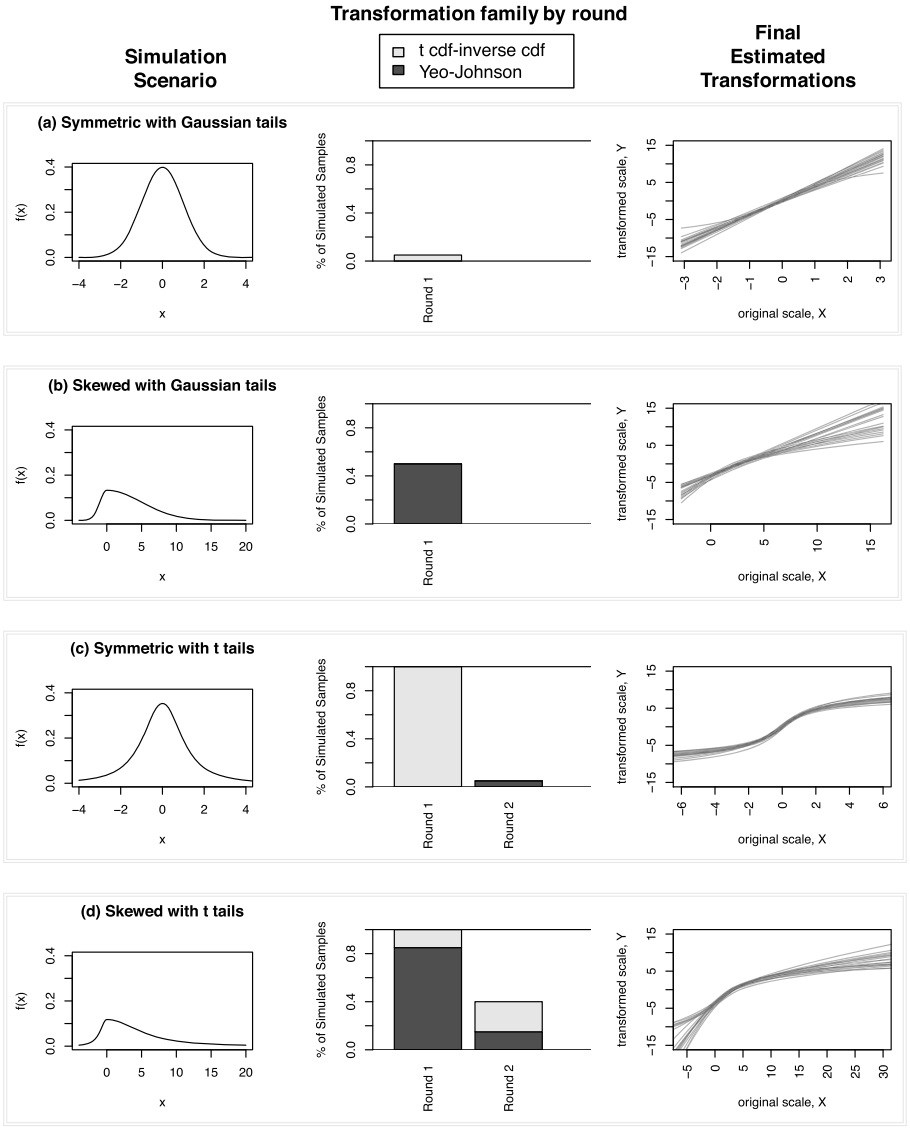

FIG 5. *Illustration of the estimated transformations based on 20 simulated samples of size* $n = 200$ *from each of the four two piece scenarios (a)-(d).*

the heavy-tailed scenarios in particular, transformations reduced the median Hellinger error by half in comparison to a direct fit of (1). Griffin's DPM (17) notably outperforms the basic DP mixture (1) for capturing skewness and heavy tails, but the TDPM method, combining (1) with a pre-transformation, gives still better results.

To investigate the transformations employed in the estimation procedures, we illustrate the selected transformations in Figure 5 for the samples with size

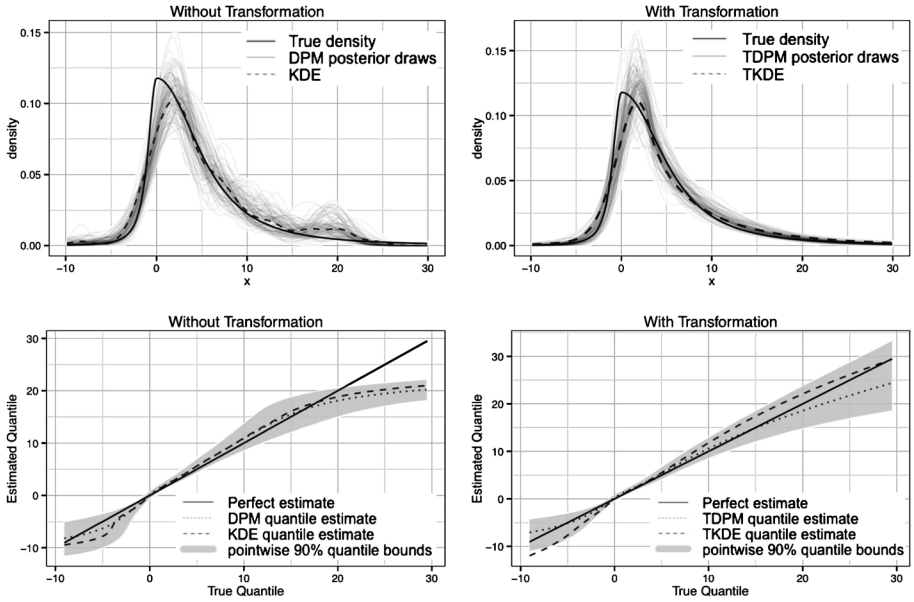

FIG 6. *Top: comparison of density estimates (via KDE and DPM) with and without transformation, based on a single sample of size* $n = 200$ *from the skewed and heavy tailed density with* $\sigma_2 = 5$ *and* $g(\cdot)$ *a* t *distribution with* 2 *degrees of freedom. Bottom: quantile-quantile plot comparing estimated to true quantiles for the same sample and model fits. Dashed and dotted lines represent point estimates of the quantiles, while the shaded regions represent pointwise 90% credible intervals for the true quantile based on the Bayesian fits.*

$n = 200$. For the skewed and heavy-tailed densities investigated here, the criterion (11) appears to be effective for identifying an appropriate remedial transformation. In each case, the transformation symmetrizes and shortens the tail of the original density, allowing more accurate density estimation on the transformed scale. When the true density is already "easy" to estimate[1], as is the case with the standard normal density shown in the top row of Figure 5, the selected transformations are close to linear. When the distribution is skewed, as in the second column, the skew-correcting transformations (4) of Yeo and Johnson are most commonly chosen. When the distribution is heavy tailed but symmetric, the tail-shortening copula transformations (5) are most common.

The Bayesian DPM model (1) allows a natural description of the uncertainty in the density estimation procedure. For a sample of size $n = 200$ from the skewed and heavy-tailed two-piece density with t_2 tails, in Figure 6 we show draws from the DPM and TDPM posterior distributions for the density. It can be seen that for this moderate sample size, the location-mixture (1) of Gaussians struggles to capture the polynomial tail decay of a t_2; while after a tail-correcting transformation, the TDPM model gives a more accurate depiction of

[1]As noted by [19], normal densities possess a curvature (3) that is already near the minimum among all densities with continuous second derivative.

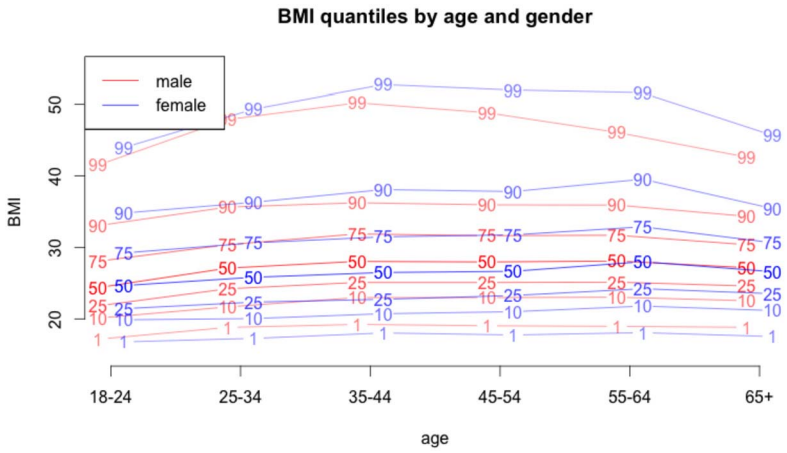

FIG 7. *Quantiles of the empirical distribution of BMI, separated by gender and age.*

the extreme tail behavior. Moreover, the DPM model yields unstable estimates of the extreme quantiles, with the empirical confidence levels of the credible intervals often falling much below the nominal values. On the other hand, the TDPM method provides much more stable estimates and more reliable credible interval coverage.

4. An application to BMI modeling

We demonstrate the use of transformations in modeling body mass index (BMI) measurements for Ohio adults grouped by age and gender. The data are from the 2008 Ohio Family Health Survey (OFHS), and are publicly available online (at http://grc.osu.edu/omas/datadownloads/ofhsoehspublicdatasets/). BMI is calculated as an individual's weight in kilograms divided by the square of height in meters. Due to the ease of its calculation, BMI is often used as a surrogate for more difficult measures of obesity, such as body fat percentage. Inference for the distribution of BMI, and for the extreme quantiles in particular, is of significant public-health interest. Individuals with extreme BMI have greater risk for a variety of health problems, including heart disease, stroke, and type-2 diabetes.

Figure 7 displays sample quantiles of the observed BMIs for each gender and age group in the study. The distributions of the BMI measurements are strongly-skewed, with heavy right tails, but the strength of these features varies with age and gender. We divide the survey respondents into subgroups by age and gender, with the aim of estimating BMI distributions for each age-by-gender group. For a given gender g and age group a, denote the set of BMI measurements by $\mathbf{x}_{ga} = \{x_{gai} : i \in \mathcal{I}_{ga}\}$, where $\mathcal{I}_{ga} = \{1, \ldots, n_{ga}\}$ are the indices for the respondents in the age-and-gender subgroup. The sizes of these subgroups are shown in Table 1.

TABLE 1
Ohio Family Health Survey (2008) sample sizes, divided into training and holdout samples.

		Age group						
		18-24	25-34	35-44	45-54	55-64	65+	Total
Female	OFHS sample size	1194	3226	4656	6268	6299	9191	30,834
	Training sample size	200	200	200	200	200	200	1200
	Holdout sample size	994	3026	4456	6068	6099	8991	29,634
Male	OFHS sample size	895	1838	2914	3892	3852	4660	18,051
	Training sample size	200	200	200	200	200	200	1200
	Holdout sample size	695	1638	2714	3692	3652	4460	16,851

To assess the effectiveness of the transformation DPM method for estimating the BMI densities, we draw subsamples of size 200 from each age-and-gender group, and measure the out-of-sample predictive likelihood. For gender g and age a, we partition the indices $\mathcal{I}_{ga} = \{1, \ldots, n_{ga}\}$ at random into training cases $\mathcal{I}_{ga}^{\text{train}}$ and holdout cases $\mathcal{I}_{ga}^{\text{hold}}$, with $|\mathcal{I}_{ga}^{\text{train}}| = 200$. Denote the training set of BMIs by $\mathbf{x}_{ga}^{\text{train}} = \{x_{gai} : i \in \mathcal{I}_{ga}^{\text{train}}\}$ and the holdout set by $\mathbf{x}_{ga}^{\text{hold}} = \{x_{gai} : i \in \mathcal{I}_{ga}^{\text{hold}}\}$. With the 200 training observations $\mathbf{x}_{ga}^{\text{train}}$, we form direct-DPM and transformation-DPM point estimates of the BMI density. As a measure of the quality of these density estimates, we consider the average log predictive likelihood for the holdout cases,

$$\overline{\mathcal{L}}_{ga} = \frac{1}{|\mathcal{I}_{ga}^{\text{hold}}|} \sum_{i \in \mathcal{I}_{ga}^{\text{hold}}} \log\bigl(\hat{f}_{ga}(x_{gai} | \mathbf{x}_{ga}^{\text{train}})\bigr). \tag{18}$$

Figure 8 gives a comparison of $\overline{\mathcal{L}}_{ga}$ for the DPM and TDPM fits.

Some improvement can be seen for the middle-aged subgroups, which also tend to have the heaviest-tailed distributions of BMI (see Figure 7). The youngest

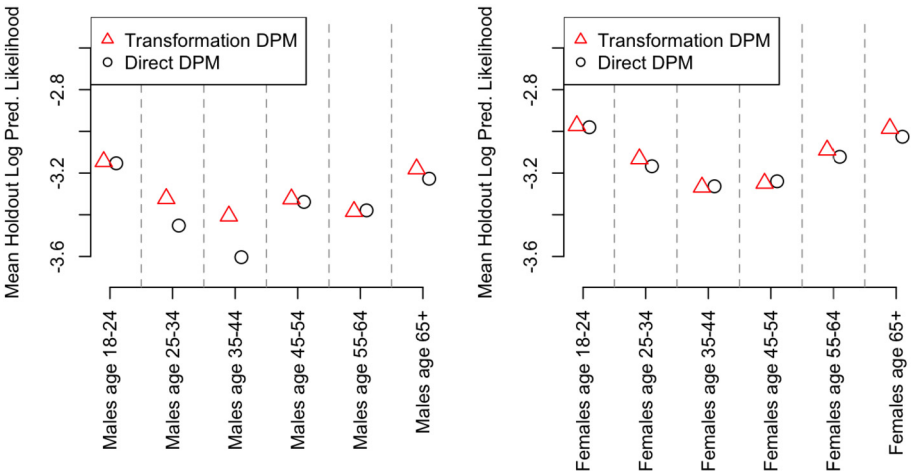

FIG 8. *Comparison of the average log predictive likelihood of the holdout cases* (18) *based on the DPM and TDPM fits.*

and oldest groups have relatively more symmetric and less heavy-tailed distributions. In such cases, there is less to be gained by transforming these samples prior to estimation, which was clear in the simulations of Section 3.

5. Discussion

In modern work on nonparametric Bayesian density estimation, many excellent and sophisticated models have been proposed, often by modifying the structure of the DP mixture (1). Here we follow a different path to improved performance, by deconstructing a difficult problem — estimation of an unknown density with extreme features — into two easier problems. First we choose a sequence of transformations to symmetrize and shorten the tails of the distribution, and second, we use basic DPM models to estimate the resulting "well-behaved" density. The evidence presented here suggests that devoting some attention to choosing a good transformation of the sample can yield substantial gains in performance for density estimation.

These two subproblems of the density estimation problem differ intrinsically in difficulty. The transformation part of the problem is low-dimensional and parametric, and so we expect estimation of the transformation parameters to follow conventional root-n asymptotics. The density estimation part of the problem is infinite dimensional, and so we expect both poorer large sample behavior and a slower asymptotic rate. Density estimators based on the DPM model are consistent under very mild conditions, and so we have little interest in finding the "optimal" transformation. Rather, we seek to find a decent one, and we then let a standard DPM model do its work. The difference in asymptotic rates for the two parts of the problem ensures that, for large samples, the transformation has little variation in comparison to the density estimate, motivating our choice to fix a single transformation rather than averaging over transformations. This strategy applies to a wide variety of statistical problems where different portions of the problem exhibit different rates. Among them, are problems where portions of a model differ greatly in dimension (as here) and problems where portions of the model are informed by different amounts of data, as in multiscale, local and treed regression models.

The advantages of the strategy we have pursued are twofold. First, conditioning on a single estimated transformation allows us to focus our computational resources on density estimation given the transformation. This task is simpler than averaging over transformations, and it allows us to rely on standard Markov chain Monte Carlo methods and code for fitting the model. Second, the single transformation approach provides a simpler conceptual framework, providing the user with a model which is easier to grasp.

There are many variations on the method we have presented. The families of transformations we have used could be replaced with other families. The transformations could be driven by likelihood rather than $\hat{L}(\theta)$. The rule for when to stop the iterative process of transformation selection need not be driven by the perspective of hypothesis testing (rule (1) of section 2.3). Information

criteria could be used to select transformation and density estimate, or a fully Bayesian approach could be used, averaging over transformations. These last two strategies are relatively expensive in terms of computation.

Nonparametric Bayesian density estimation is often used in multivariate settings. A natural question is how to extend this method to multivariate transformations to alleviate excess skewness and kurtosis. One promising route is to preprocess the data by first conducting a principal components analysis of the data. The transformation approach described herein could then be applied to each margin. Full development of such a method awaits further work.

The TDPM model can be represented in alternative forms. While we describe the model in terms of a transformation, followed by a DPM model, and finally completed with a back-transformation to the original scale, a mathematically equivalent version describes the model as a DPM with a non-standard base measure and kernel. Both presentations of the model have value; we believe the presentation here highlights our overall modelling strategy.

One of the great advantages of DP mixture models such as (1) and (17) is their flexibility, allowing them to capture unusual features in the target density. Only two such features, skew and heavy-tails, are investigated here with the transformation method. The usefulness of the transformation strategy has not been established for other situations, such as estimation of many-modal distributions. The basic idea, however, remains powerful. The area in which TDPM estimates show the most improvement over basic DPM is in the tails of the estimates. We would expect this advantage to persist, even if strange features are present in the body of the distribution.

References

[1] M. D. Escobar and M. West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995. MR1340510

[2] T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973. MR0350949

[3] T. S. Ferguson. Bayesian Density Estimation by Mixtures of Normal Distributions, 1983. MR0736538

[4] C. Fernandez and Mark F. J. Steel. On Bayesian Modeling of Fat Tails and Skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998. MR1614601

[5] S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143–158, 1999. MR1701105

[6] S. Ghosal and A. W. Van Der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29(5):1233–1263, 2001. MR1873329

[7] J. E. Griffin. Default priors for density estimation with mixture models. *Bayesian Analysis*, 5(1):45–64, 2010. MR2596435

[8] T. Iwata, D. Duvenaud, and Z. Ghahramani. Warped mixtures for nonparametric cluster shapes. *arXiv preprint arXiv:1206.1846*, 2012.

[9] M. C. Jones and S. J. Sheather. Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 11(6):511–514, Jun 1991. MR1116745

[10] A. Y. Lo. On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12(1):351–357, 1984. MR0733519

[11] S. N. MacEachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics – Simulation and Computation*, 23(3):727–741, jan 1994. MR1293996

[12] F. J. Rubio and M. F. J. Steel. Inference in two-piece location-scale models with Jeffreys priors. *Bayesian Analysis*, 9(1):1–22, 2014. MR3188293

[13] D. Ruppert and M. P. Wand. Correcting for Kurtosis in Density Estimation. *Australian Journal of Statistics*, 34(March 1991):19–29, 1992. MR1177270

[14] S. J. Sheather and M. C. Jones. A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690, 1991. MR1125725

[15] M. P. Wand. *KernSmooth: Functions for Kernel Smoothing Supporting Wand and Jones (1995)*, 2015. R package version 2.23-14.

[16] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994. MR1319818

[17] M. P. Wand, J. S. Marron, and D. Ruppert. Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343–353, 1991. MR1137118

[18] L. Yang. Root-n convergent transformation-kernel density estimation. *Journal of Nonparametric Statistics*, 12(4):447–474, 2000. MR1785394

[19] L. Yang and J. S. Marron. Iterated transformation-kernel density estimation. *Journal of the American Statistical Association*, 94(446):580–589, 1999. MR1702327

[20] I. K. Yeo and R. A. Johnson. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, 87(4):954–959, 2000. MR1813988