Flexible Univariate Continuous Distributions

Fernando A. Quintana, Mark F.J. Steel and José T.A.S. Ferreira*

Abstract

Based on a constructive representation, which distinguishes between a skewing mechanism P and an underlying symmetric distribution F, we introduce two flexible classes of distributions. They are generated by nonparametric modelling of either P or F. We examine properties of these distributions and consider how they can help us to identify which aspects of the data are badly captured by simple symmetric distributions. Within a Bayesian framework, we investigate useful prior settings and conduct inference through MCMC methods. On the basis of simulated and real data examples, we make recommendations for the use of our models in practice.

Keywords: density estimation; location-scale; modal regression; moment existence; skewness; unimodality

1 Introduction

In many applications, modelling continuous variables using simple default distributions, such as the normal, may not be appropriate and we need to consider more flexible alternatives. In adopting such general models, however, we typically leave one question unanswered: which aspect of normality was inappropriate in the context of the specific application? Was it symmetry, unimodality, or the light tails? Or some combination of these? In other words, we wish to assess which features of the default distribution are not well supported by the data. Therefore, we will consider flexible modelling within a constructive framework,

*Fernando Quintana is Profesor Adjunto, Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile (email quintana@mat.puc.cl), Mark Steel is Professor, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (email M.F.Steel@stats.warwick.ac.uk) and J. Tomé Ferreira is with Endeavour Capital Management, London, U.K. (email tome.ferreira@endeavour-capital.com). *The work of the third author was conducted when he was at the University of Warwick. This research was partially funded by grants Fondecyt 1060729 and 7060193.*

provided by Ferreira and Steel (2006), which applies a skewing mechanism (a probability density function p on the unit interval) to an underlying symmetric distribution, say, F with density f. The form of the multiplicative skewing mechanism can give us a more precise idea of what specific aspects of F are not correctly capturing the data. This could include any of the characteristics already mentioned plus others such as the existence of moments or different left and right tail behavior. By separating out p and f, we maintain a certain amount of control over properties of the distributions, as well as interpretability.

We adopt flexible representations for either p or f. In a first model, we use a flexible skewing mechanism, based on Bernstein densities, which allows for a large variety of distributional shapes, keeping a simple default form for f. In a second modelling approach, we use a parametric construct for the skewing mechanism as in Ferreira and Steel (2006), but allow for a flexible f through a nonparametric mixture of uniforms. The latter framework will always generate unimodal distributions, which can be desirable in regression modelling.

In this paper, inference will be conducted within a Bayesian framework, but the sampling models proposed here could, of course, also be analysed through other inferential schemes. A Bayesian setting, however, makes it easier to control the properties of the models through careful prior choices.

This article is structured as follows. Section 2 presents and discusses the two models, including our proposed prior definition. Section 3 applies these models to several simulated data sets. Section 4 discusses applications to two real data sets, also considering extensions to linear regression with flexible error distributions. We conclude in Section 5 with some final remarks and recommendations.

2 Model Definition

We consider the general modelling strategy introduced in Ferreira and Steel (2006) based on a multiplicative skewing mechanism, leading to a univariate probability density function (pdf) of the form

$$h(y|F,P) = f(y)p[F(y)], \tag{1}$$

where F is the cumulative distribution function (cdf) corresponding to a symmetric unimodal pdf f, and p is the pdf of a distribution P on the unit interval. Throughout the paper, we will consistently use the notation F and P for the distributions (and their cdf's) defined above, and use H for the distribution generated by (1). Generally, pdf's are denoted by the corresponding lower-case letters and π will indicate a prior pdf or probability mass function. We will focus on random variables Y defined on the real line $\Re$, but extensions to

subsets of $\Re$ are straightforward, by appropriately defining the domain of F. The skewing mechanism P does not depend on the underlying symmetric distribution F. Without this requirement, other modelling strategies, such as the generalized skew-elliptical class of Genton and Loperfido (2005) can also be generated, but we would lose the separate interpretability of P and F. Throughout, we assume that both p and f exist and are continuous. The choice of p determines the form of h. For instance, $p(x) = 1$ for all $x \in (0, 1)$ means $h(y)$ becomes just $f(y)$. Other interesting properties are related to the limiting values of p. For example, $p(0) = 0$ implies a downweighting of the left tail, while $p(1) = 0$ does the same for the right tail. By making the limits of p zero or infinity, we can make the order of moment existence of $h(y)$ different from that of $f(y)$.

For modelling purposes, consider the location-scale family corresponding to (1), defined as

$$h(y \mid F, P, \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right) p\left[F\left(\frac{y-\mu}{\sigma}\right)\right]. \tag{2}$$

A general model for density estimation problems based on a sample of size n assumes that

$$Y_1, \ldots, Y_n \mid F, P, \mu, \sigma \sim h(y \mid F, P, \mu, \sigma),$$

with h defined as in (2). Extensions to regression models are immediate, as discussed in Subsection 4.2.

A flexible model based on (2) will be constructed using either of the two procedures we describe next. In the first, which we will call the **Bernstein-skew model**, we adopt a parametric form for the pdf f and model $p(x)$ nonparametrically using Bernstein densities. The other model, named the **flexible unimodal construct** considers a parametric definition of $p(x)$ and a flexible definition of $f(y)$ through a nonparametric mixture of uniforms. The two approaches have different properties and motivations. The flexible model for $p(x)$ will support multimodality of the resulting pdf $h(y)$; the other approach, however, is constructed with the specific purpose of generating only unimodal outcomes.

Density estimation is formalized by means of the posterior predictive density $p(Y_{n+1} \mid y_1, \ldots, y_n)$ which can be obtained via posterior simulation using MCMC algorithms.

2.1 The Bernstein-Skew Model

We start by specifying a convenient symmetric density $f(y)$. We choose a Student t_ν distribution with unknown degrees of freedom ν. Next, we adopt a flexible representation for the skewing mechanism $p(x)$. Given the constructive representation of skewed distributions introduced above, a completely unrestricted nonparametric treatment of the skewing mechanism may look appealing. As P can be any distribution in $(0, 1)$, the possibility to model

it in an unrestricted fashion seems tempting. However, this would effectively generate the entire class of continuous distributions and, thus, we would lose control over the properties of the resulting skewed distributions.

Here we try to reach a compromise between an unrestricted nonparametric skewing mechanism and one for which some interesting results are still available. We make use of Bernstein densities (see *e.g.* Petrone, 1999 and Petrone and Wasserman, 2002) to model p. Given a positive integer m and a weight vector $\boldsymbol{w}^m = (w_1^m, \ldots, w_m^m)$, a Bernstein density is defined as

$$p(x \mid m, \boldsymbol{w}^m) = \sum_{j=1}^{m} w_j^m \mathrm{Be}(x \mid j, m-j+1), \tag{3}$$

where $m \in \{1, 2, \ldots\}$, $\boldsymbol{w}^m$ is constrained by $w_j^m \geq 0$ for all $1 \leq j \leq m$ and $\sum_{j=1}^m w_j^m = 1$, and we use the Beta pdf's

$$\mathrm{Be}(x \mid j, m-j+1) = \frac{m!}{(j-1)!(m-j)!} x^{j-1}(1-x)^{m-j}, \qquad 0 \leq x \leq 1.$$

Bernstein densities have characteristics that make their use as flexible skewing mechanisms quite attractive. Observe that $m = 1$ leads us back to the original model F. Also, for any choice of m, if we take $w_i^m = 1/m,\ i = 1, \ldots, m$, then P is Uniform which implies that we retain the original symmetric distribution F. Further, as long as there is a $j^* \in \{1, \ldots, m^*\}$, with $m^* = m/2$ if m is even and $m^* = (m-1)/2$ otherwise, such that $w_{j^*}^m \neq w_{m-j^*+1}^m$, then $p(x|m, \boldsymbol{w}^m)$ is asymmetric.

Members of the class of distributions generated by this skewing mechanism are often multimodal. Multimodality becomes more common as m increases.

For a given ν, our choice of $f(y)$ has finite moments up to order less than ν. We now present the following definitions to characterize tail behaviour.

Definition 1. Let G be the distribution of a random variable Y in $\Re$. We define:

(i) Largest left moment of G : $M_l^G = \sup\{q \in \Re_+ : \int_{-\infty}^0 |y|^q dG < \infty\}$.

(ii) Largest right moment of G : $M_r^G = \sup\{q \in \Re_+ : \int_0^\infty y^q dG < \infty\}$.

(iii) Largest moment of G : $M^G = \min\{M_l^G, M_r^G\}$.

If distribution G is symmetric with continuous pdf, as is the case for F, then $M_l^G = M_r^G = M^G$. As examples, the Normal and Logistic distributions have $M^G = \infty$ while the heavier tailed t_ν has $M^G = \nu$.

As the parameters of the Beta distributions in the mixture in (3) are never smaller than unity, $p(x|m, \boldsymbol{w}^m)$ is bounded and, therefore, one can easily deduce that $M_l^H, M_r^H \geq M^F$. Actually, if we exclude zero weights we obtain $M_l^H = M_r^H = M^F$. Now we discuss how to allow for more flexible tail behaviour.

The next Theorem presents a useful result regarding tail behaviour of $h(y)$. For the sake of brevity, we only analyse the effect on M_l^H. Treatment for M_r^H is entirely analogous.

Theorem 1. If $\lim_{y\to-\infty} \frac{p[F(y)]}{|y|^b}$ is positive and finite for some $b \in \Re$, then $M_l^H = M^F - b$.

The result in Theorem 1 is particularly interesting when $f(y)$ is a Student-t pdf, since we can derive the following:

Theorem 2. In the model (2) with skewing mechanism (3) and a Student distribution with ν degrees of freedom for F, the restriction $w_1^m = \cdots = w_k^m = 0$ implies that $M_l^H = (k+1)\nu$. Similarly, $w_{m-k+1}^m = \cdots = w_m^m = 0$ implies $M_r^H = (k+1)\nu$.

Therefore, we have a specific interest in the case where some of the weights at the extremes are zero. This is done by means of a prior distribution that assigns positive probability to such events. For simplicity, we consider here only the case $k = 2$, i.e., we assess whether the data supports existence of two or three times the number of original moments, in the left or the right-hand tail. The tail(s) that is (are) not affected by zero weights will have the same moment existence as F and will, in principle, allow us to learn about ν. In practice, when faced with very multimodal data, we will need to use fairly large values for m and we will then often fix ν to avoid introducing too much flexibility.

In summary, the Bernstein-skew model accommodates very general distributional shapes, with *e.g.* multimodality, but does allow us to control or monitor the relative tail behaviour (in terms of left and right moment existence) by manipulating the weights. By allowing for k zero weights on either end, we give the data the opportunity to inform us on differences in moment existence up to a factor of $k + 1$. Finally, both the skewing mechanism in (3) and the symmetric density f are smooth functions, which means that the Bernstein-skew model will generate smooth skewed distributions.

2.1.1 Prior specification

In the sequel, $\boldsymbol{w}^m_{-\boldsymbol{i}}$ represents the $\boldsymbol{w}^m$ vector with the coordinates in $\boldsymbol{i} \subset \{1, \ldots, m\}$ removed. Also a point-mass distribution at $\boldsymbol{a}$ is denoted by $\delta_{\boldsymbol{a}}(\cdot)$. In view of the discussion following

Theorem 2, we adopt the following prior for the weights in (3)

$$\begin{aligned}\pi(\boldsymbol{w}^m \mid m, \boldsymbol{c}^m, \theta_1^l, \theta_2^l, \theta_1^r, \theta_2^r) \;=\; & \theta_1^l(1-\theta_2^l)\left[\delta_0(w_1^m), \mathrm{Dir}_{m-1}(\boldsymbol{w}_{-1}^m \mid \boldsymbol{c}_{-1}^m)\right] \\ & +\theta_1^l\theta_2^l\left[\delta_{(0,0)}(w_1^m, w_2^m), \mathrm{Dir}_{m-2}(\boldsymbol{w}_{-(1,2)}^m \mid \boldsymbol{c}_{-(1,2)})\right] \\ & +\theta_1^r(1-\theta_2^r)\left[\delta_0(w_m^m), \mathrm{Dir}_{m-1}(\boldsymbol{w}_{-m}^m \mid \boldsymbol{c}_{-m}^m)\right] \\ & +\theta_1^r\theta_2^r\left[\delta_{(0,0)}(w_{m-1}^m, w_m^m), \mathrm{Dir}_{m-2}(\boldsymbol{w}_{-(m-1,m)}^m \mid \boldsymbol{c}_{-(m-1,m)})\right] \\ & +(1-\theta_1^l-\theta_1^r)\,\mathrm{Dir}_m(\boldsymbol{w}^m \mid \boldsymbol{c}^m), \end{aligned} \tag{4}$$

for known constants $\theta_i^l, \theta_i^r \in [0,1), i = 1,2$ such that $\theta_1^l + \theta_1^r < 1$. Here

$$\mathrm{Dir}_k(\boldsymbol{z}^k \mid \boldsymbol{c}^k) = \frac{\Gamma(c_1^k + \cdots + c_k^k)}{\Gamma(c_1^k)\cdots\Gamma(c_k^k)} z_1^{c_1^k - 1} \cdots z_k^{c_k^k - 1}$$

denotes the Dirichlet pdf, given a parameter vector of positive entries $\boldsymbol{c}^k = (c_1^k, \ldots, c_k^k)$, and where $\boldsymbol{z}^k = (z_1, \ldots, z_k)$ is constrained by $z_i > 0$ for all $1 \le i \le k$ and $\sum_{i=1}^k z_i = 1$. Also, θ_1^l and θ_1^r in (4) represent the marginal prior probabilities that $w_1^m = 0$ and $w_m^m = 0$, respectively. Likewise, θ_2^l and θ_2^r represent the conditional probabilities that $w_2^m = 0$ and $w_{m-1}^m = 0$, given that $w_1^m = 0$ and $w_{m-1}^m = 0$, respectively. If, for instance, $w_1^m = w_2^m = 0$, then $(w_3^m, \ldots, w_m^m)$ have a joint Dirichlet distribution in the appropriate space. By taking $\theta_1^l = \theta_1^r = 0$, we restrict left and right moments to be the same as that of the underlying symmetric distribution.

We treat the number of components m as unknown, and adopt a prior for it based on a Poisson distribution. In particular, we choose $0 < \psi < 1$ and $\lambda > 0$ and define $\pi(m)$ as

$$\pi(1) = \psi, \qquad \pi(m) = \frac{(1-\psi)\lambda^m \exp(-\lambda)}{m!(1 - \exp(-\lambda) - \lambda\exp(-\lambda))}, \quad m = 2, 3, \ldots. \tag{5}$$

The use of a point mass at one implies that the symmetric model F is assigned a positive prior probability equal to ψ.

The prior on $(m, \boldsymbol{w}^m, \mu, \sigma, \nu)$ is then chosen as

$$\pi(m, \boldsymbol{w}^m, \mu, \sigma, \nu) = \pi(m)\pi(\boldsymbol{w}^m \mid m)\pi(\mu)\pi(\sigma)\pi(\nu),$$

where $\pi(\nu) = \mathrm{U}(0, \nu_{max})$, $\pi(\sigma) = \mathrm{U}(0, S)$, $\pi(\mu) = \mathrm{N}(\mu_0, \tau)$, and $\pi(\boldsymbol{w}^m \mid m)$ and $\pi(m)$ are given by (4) and (5), respectively. Choices for the Dirichlet parameters $\boldsymbol{c}^m$ in (4) will be discussed in the next subsection.

2.1.2 Choices for Dirichlet parameters

In the prior for $\boldsymbol{w}^m$ in (4), the choice of the Dirichlet parameter vector $\boldsymbol{c}^m$ is critical. As we know, equal weights $w_i^m = 1/m$ result in maintaining the original symmetric distribution

F, and centring the prior over F might be a reasonable choice. Abstracting from the point masses at the extremes, this is obtained for equal values of c_i^m, say $c_i^m = b$. This will lead to

$$\mathrm{E}(w_i^m) = \frac{1}{m} \quad \text{and} \quad \mathrm{Var}(w_i^m) = \frac{m-1}{m^2(mb+1)},$$

implying a coefficient of variation equal to

$$\mathrm{CV}(w_i^m) = \sqrt{\frac{m-1}{mb+1}}.$$

As m increases, the sample information on each w_i^m will normally decrease, so it might be reasonable to impose more prior structure. This reasoning would have $\mathrm{CV}(w_i^m)$ decrease with m, which suggests choices for b that are increasing in m. For example $b = \sqrt{m}/10$ would lead to a coefficient of variation of 1.374 for $m = 5$ and 0.667 for $m = 500$.

Another consideration, however, is that accounting for outlying observations far out in the tails through the incorporation of extra (smaller) modes might be difficult if the prior on the weights is too tight. Such modes have to be accommodated by one or at best a few Bernstein densities with j close to 1 (in the left tail) or m (in the right tail). This implies that many of the weights out in the tails will have to be very close to zero. The densities building up the centre of the distribution (corresponding to j around $m/2$) do not have this problem, as there is rarely need to let the resulting density dip down all the way to zero in the central part of the distribution. This would suggest that the prior on the extreme weights should be less concentrated, and can be accommodated by taking *e.g.*

$$c_i^m = \frac{\sqrt{m}}{\left(10 + (\frac{m}{2} - i)^2\right)}, \tag{6}$$

which leads to coefficients of variation 50 times as large in the extremes than for the central weights (using $m = 300$). In addition, this makes the mean weights in the centre much larger than in the tails, which seems a priori reasonable. This also has the advantage that it concentrates the skewing mechanism to assign a relatively large amount of mass to the centre of the distribution. This ties the distribution down somewhat and avoids artifacts of too much flexibility, such as multimodal inference on β. Of course, now we lose the exact centring over F as the prior weights are no longer all the same. A less extreme version of the prior above is generated by

$$c_i^m = \frac{\sqrt{m}}{\left(50 + (\frac{m}{2} - i)^2\right)}, \tag{7}$$

which leads to a CV inflation by a factor 22 in the tails for $m = 300$.

In the sequel, we shall investigate a number of prior choices for $\boldsymbol{c}^m$, denoted as follows:

- Prior 1: $c_i^m = 1$
- Prior 2: $c_i^m = \sqrt{m}/10$
- Prior 3: $c_i^m = \sqrt{m}/50$
- Prior 4: the prior in (6)
- Prior 5: the prior in (7)

For prior 1 the variance coefficient slowly increases with m, while for all other priors it decreases. Prior weights are exchangeable for Priors 1-3, but not for Priors 4 and 5 as explained above.

2.2 The Flexible Unimodal Construct

The Bernstein-skew model is especially targeted to density estimation problems, where the goal is to provide as much flexibility as possible, without imposing restrictive conditions on the shape of the distribution. The flexible unimodal construct involves a different modelling strategy. Here we start with a general unimodal symmetric distribution, and incorporate a skewing mechanism p that is constrained to be unimodal with mode at $1/2$. From Theorem 2 in Ferreira and Steel (2006), this ensures that the resulting skewed distribution is unimodal with the same mode as f, a necessary property in the context of modal regression models. In particular, we follow Ferreira and Steel (2006) and choose their constructed skewing mechanism with proportional tails $p(x) = p(x \mid \delta, r)$. The latter accommodates skewness around the mode and in the tails of the distribution, without altering moment existence. Both tails are proportional to each other and $M_l^H = M_r^H = M^F$. The parameter $\delta \in \Re$ controls skewness around the mode, while $r > 0$ is the ratio of the right to the left tail, which induces skewness in the tails. For $\delta = 0$ and $r = 1$ we recover the original symmetric distribution F. Positive values of δ indicate positive (right) skewness in the central part of the distribution, and $r > 1$ leads to more mass in the right tail than in the left.

In order to aid the interpretation of P as a skewing mechanism, this approach formally restricts the class of P such that the only P leading to a symmetric h is the uniform. We choose the skewing mechanism such that p has two continuous derivatives. However, the smoothness of P will not be inherited by the resulting distribution H, as the symmetric distribution F (to be defined below) will not have the same smoothness properties. As a consequence, the flexible unimodal construct will not necessarily generate smooth predictives, but it will have a lot of flexibility to adapt to the data.

Thus, in this approach, we maintain unimodality and control skewness. An interesting skewness measure for unimodal distributions is the one proposed by Arnold and Groeneveld (1995), defined as 1 minus twice the mass to the left of the mode (denoted by AG skewness, henceforth). Ferreira and Steel (2006) give an expression for AG skewness as a function of δ and r, which applies here as we are using their proposal for P, and the choice of F does not affect this measure of skewness.

To model F in a flexible manner, we use the representation of unimodal symmetric densities as mixtures of uniform distributions:

$$f(y) = \int_0^\infty \frac{1}{2\theta} I\{y \in (-\theta, \theta)\}\, dG(\theta), \tag{8}$$

as discussed in Brunner and Lo (1989), where $I\{\cdot\}$ is the indicator function and where G is a distribution function on the positive real numbers. In particular, for G in (8) we adopt a stick-breaking prior distribution with a finite number of terms. Thus, G can be expressed as

$$G(\cdot) = \sum_{i=1}^{N} w_i \delta_{\theta_i}(\cdot), \tag{9}$$

where, for independent random variables $V_1, \ldots, V_{N-1}$ with $V_i \sim \text{Be}(a_i, b_i)$ and known a_i and b_i, we set $w_1 = V_1$ and $w_k = V_k \prod_{j=1}^{k-1}(1 - V_j)$ for $k = 2, \ldots, N$, and $V_N = 1$, which guarantees $P(\sum_{i=1}^{N} w_i = 1) = 1$. Furthermore, $\theta_1, \ldots, \theta_N$ are i.i.d. draws from some absolutely continuous centering probability measure G_0 on the positive real numbers, and independent of the weights w_i. Under this specification, the resulting G will be centered around G_0 for any given N, in the sense that for any Borel measurable subset of $\Re^+$, $\text{E}(G(B)) = G_0(B)$. A natural choice is an exponential G_0 with mean $1/\zeta$, which is thus the mean and the standard deviation of the atoms $\theta_1, \ldots, \theta_N$. We also experimented with different forms for G_0, such as a Gamma distribution with shape parameter 2 and unknown scale, but results were virtually identical.

The reason for using a finite representation in (9), rather than the more common infinite one (giving rise, for constant $a_i = 1$ and $b_i = M$ to Dirichlet process priors) is that it allows explicit expressions for f and F in (1) or (2). In the particular implementation used for the examples, we have taken N fixed, but we could, in principle, also allow for unknown N. However, we feel that taking, say, $N = 50$ allows for sufficient flexibility in practice.

The resulting skewed density function can be expressed as before in (2), where $F(y)$ now corresponds to the pdf in (8) and we use the constructed skewing mechanism $p(x \mid \delta, r)$. The resulting distribution will be unimodal with mode equal to μ. We adopt $a_i = 1$ and $b_i = M$ for $1 \leq i \leq N - 1$ so that a priori $\text{E}(w_k) = M^{k-1}/(M+1)^k$ for $1 \leq k \leq N - 1$

and $\mathrm{E}(w_N) = [M/(M+1)]^{N-1}$, which describes the expected weight given to each of the different atoms in (9). The prior on the remaining parameters is taken to be

$$\pi(\mu, \sigma, \delta, r, M, \zeta) = \pi(\mu)\pi(\sigma)\pi(\delta)\pi(r)\pi(M)\pi(\zeta),$$

where $\pi(\mu) = \mathrm{N}(\mu_0, \tau)$, $\pi(\sigma) = \mathrm{U}(0, S)$, $\pi(r)$ corresponds to a log-normal with scale τ_r, and $\pi(\delta)$ to a t distribution with ν_δ degrees of freedom and scale parameter τ_δ. Thus, the prior is centred over symmetry. For $\pi(M)$ we take a Gamma distribution with shape parameter α_M and mean α_M/β_M, and the prior for ζ is a Gamma with shape parameter α_ζ and unitary mean. Making ζ and M stochastic allow us to learn about the spread of the N atoms and the different weights assigned to each of them.

3 Results with simulated data

For all applications (on simulated data in this section and on real data in the next), we use the following prior settings. The Bernstein-skew model is used with $\nu \sim U(0, 100)$, $\sigma \sim U(0, 100)$, $\mu \sim N(0, 100^2)$. Throughout, we take $\theta_1^l = \theta_1^r = 0.24$ and $\theta_2^l = \theta_2^r = 1/6$ and $\psi = 0.1$. Furthermore, we vary prior choices for $\boldsymbol{c}^m$ and λ in the priors (4) and (5), and sometimes we keep ν fixed.

In the flexible unimodal construct, we use the same priors for μ and σ as in the Bernstein case and choose the prior hyperparameters $\tau_r = 5$, $\tau_\delta = 25$, $\nu_\delta = 5$, $\alpha_M = 0.4$, $\beta_M = 0.2$ and $\alpha_\zeta = 0.1$. We base the model on a stick-breaking prior for G with $N = 50$.

For both models, the prior is centred over symmetry.

3.1 Student t

We simulate samples of $n = 100$ and $n = 2000$ observations from a Student-t distribution with 5 degrees of freedom.

For the Bernstein-skew model, we use fairly low values for λ in the prior for m (see (5)), as these are obviously very regular and unimodal data. Posterior predictive results are virtually unaffected by the choice of $\boldsymbol{c}^m$ (see Subsection 2.1.2) and by the selection of $\lambda = 10$ or $\lambda = 20$, especially when we use Priors 4 and 5 (in (6) and (7)). For the sample of 100 observations, the left-hand panel of Figure 1 shows the predictive densities for Prior 5 with $\lambda = 10$ and $\lambda = 20$, overplotted on the normalized data histogram. For both samples, the generated data display a slight right skew (as measured by the AG skewness) and, as a consequence, the posterior probability of $m = 1$ is actually smaller than its prior counterpart, taken to be

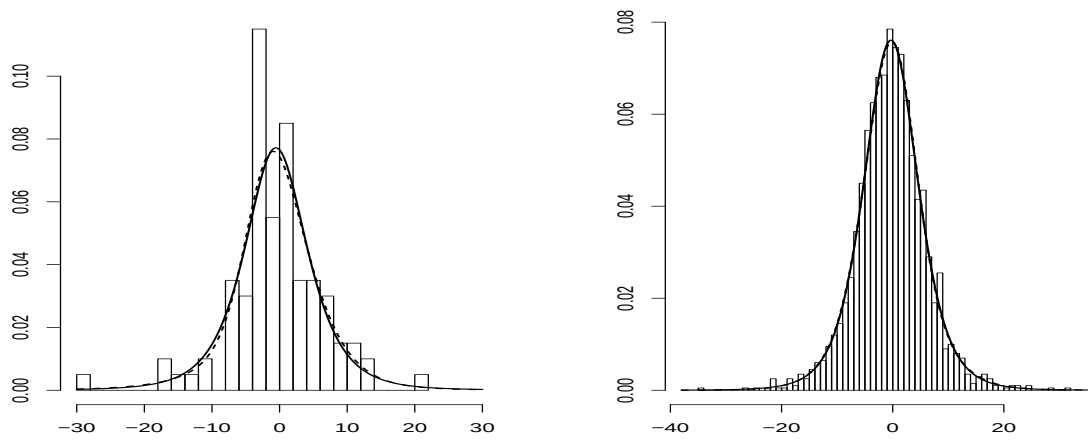

Figure 1: Simulated Student-t data, Bernstein-skew model: Posterior predictive densities, overplotted with the normalized data histogram. Left panel: $n = 100$, using Prior 5 for $\lambda = 10$ (solid line) and $\lambda = 20$ (dashed line). Right panel: $n = 2000$, with Prior 5 for $\lambda = 10$ (solid line) and $\lambda = 20$ (dashed line).

$\psi = 0.1$, for almost all combinations of $\boldsymbol{c}^m$ and λ. Nevertheless, the predictive results with the Bernstein-skew model are very smooth and similar to the actual data generating process.

For $n = 100$, the posterior probabilities of zero weights in the tails tend to be rather similar to the prior probabilities. For the large sample of 2000 observations, zero weights get very small posterior probabilities. For both samples, posterior inference on ν is quite concentrated around the value used to generate the data if we take $\lambda = 10$. For $\lambda = 20$ the increased amount of prior flexibility leads to slightly more dispersed posterior distributions for ν.

Using the flexible unimodal model on the same data, we obtain the posterior predictive results in Figure 2 for $n = 100$ and in Figure 3 for the large sample. The main difference with the Bernstein-skew results is in the smoothness: the flexible unimodal model leads to less smooth predictives, as expected, and very closely follows the shape of the data. By construction it can not reproduce the vaguely multimodal character of the small sample but closely fits most other aspects of the histogram. Interestingly, the Bernstein-skew model could accommodate multimodality, but does not do so in this case (see Figure 1) because of its inherent smoothness. Even with the large sample of 2000 observations, the flexible unimodal model tracks some minor irregularities of the data. The mode of the data is estimated at around -2 (μ has median -2.08, with first and third quartiles equal to -2.19 and -1.83).

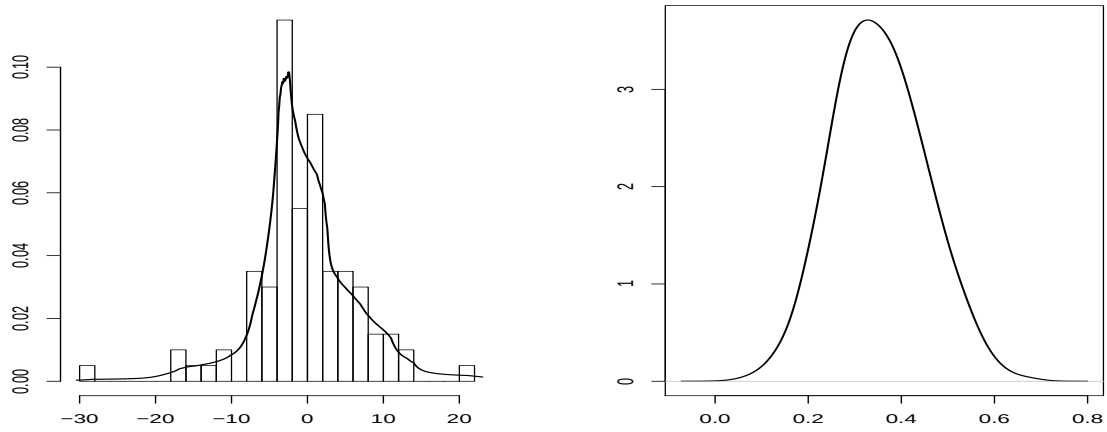

Figure 2: Simulated Student-t data, flexible unimodal construct, $n = 100$: Left panel: posterior predictive density, overplotted with the normalized data histogram. Right panel: posterior density of AG skewness measure.

In contrast, inference for μ in the Bernstein-skew model (where μ is not interpretable as the mode) is quite different with a posterior median equal to -0.32 and quartiles -0.43 and -0.23.

Figures 2 and 3 also display the posterior density of the AG skewness measure, which is in line with the slight positive skew in both samples. Taking as an example the case of $n = 2000$, inference on δ and r shows that some positive skewness is present both in the central mass of the distribution (the first and third quartiles of δ are 2.01 and 3.09) and in the tails (first and third quartiles of r are 2.13 and 2.68).

In addition, the posterior distribution for ζ had median 0.427 and quartiles 0.288 and 0.633 for $n = 100$, and median 0.379 with quartiles 0.299 and 0.476 for $n = 2000$, revealing a fairly spread posterior distribution of the atoms. For M the posterior median for $n = 100$ was 2.779 (with quartiles 1.405 and 5.690 and posterior mean 4.282) and 3.935 for $n = 2000$ (with quartiles 2.819 and 5.269). This reveals that in both examples substantial posterior mass had to be assigned to quite a few different atoms in order to capture the data features. However, the difference in sample sizes is reflected in the more concentrated posterior distributions for $n = 2000$.

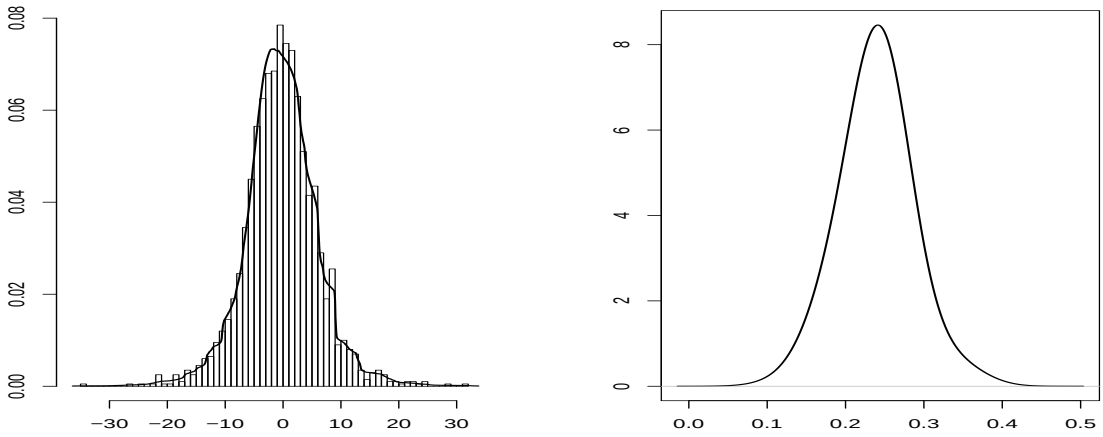

Figure 3: Simulated Student-t data, flexible unimodal construct, $n = 2000$: Left panel: posterior predictive density, overplotted with the normalized data histogram. Right panel: posterior density of AG skewness measure.

3.2 Mixture of two Normals

A sample of $n = 100$ data is generated from a mixture of two Normals, with density $0.3\mathrm{N}(-2, 4) + 0.7\mathrm{N}(4, 1)$, where $\mathrm{N}(\mu, \sigma^2)$ is the pdf of a Normal distribution with mean μ and variance σ^2. As the data are clearly bimodal, we use the Bernstein-skew model, but given the relatively regular nature of the data, we again adopt a fairly small value for the prior mean of m, choosing $\lambda = 10$ and 20. Figure 4 presents the posterior predictive densities for two choices of λ and all five priors on the weights mentioned in Subsection 2.1.2. Prior 1 is the only one which leads to perhaps too little separation between the two modes. Posterior means of the implied skewing mechanism are presented in Figure 5, which suggests that the behaviour induced by Prior 1 is indeed somewhat different from the others (it is somewhat shifted towards the right). Comparing the skewing mechanisms for $\lambda = 20$, it becomes clear that larger m allows for a lot of flexibility, which results in rather disparate skewing mechanisms. The predictives, however, are very similar to each-other, as the differences are largely counteracted by differences in the inference on μ, σ and ν. This suggests that, even in this bimodal case, priors with too much emphasis on large value of m are not required, and perhaps best avoided. The posterior means of the skewing mechanisms clearly indicate how the bimodality is induced and the relatively large values of $p(x)$ close to $x = 0$ generate the larger spread in the left tail. In line with this, there is strong posterior evidence for zero weights in the right tail, but not in the left tail.

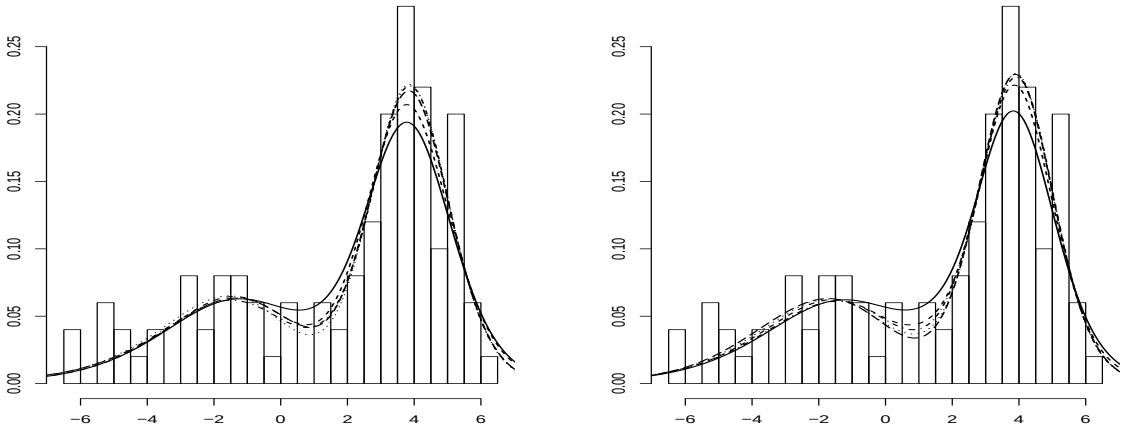

Figure 4: Simulated mixture of Normals data, Bernstein-skew model: Posterior predictive densities, overplotted with the normalized data histogram. Left panel: $\lambda = 10$. Right panel: $\lambda = 20$. Priors used are Prior 1 (solid line), Prior 2 (short dashes), Prior 3 (dots), Prior 4 (dot-dashed) and Prior 5 (long dashes).

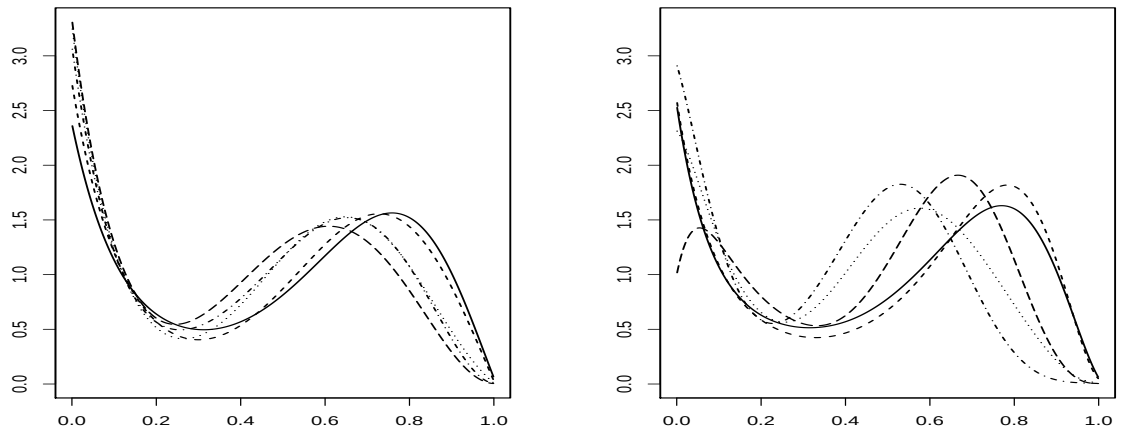

Figure 5: Simulated mixture of Normals data, Bernstein-skew model: Posterior means of the pdf of the skewing mechanism. Left panel: $\lambda = 10$. Right panel: $\lambda = 20$. Priors used are Prior 1 (solid line), Prior 2 (short dashes), Prior 3 (dots), Prior 4 (dot-dashed) and Prior 5 (long dashes).

3.3 Skew-t

The distribution used to generate the data in this case is the skew-t distribution proposed and analysed by Jones and Faddy (2003). This is an interesting distribution for our purpose as it displays both skewness and different moment existence in each tail. In particular, if X has a Be(a, b) distribution such a skew-t distributed random variable can be generated as

$$T = \frac{\sqrt{(a+b)}(2X-1)}{2\sqrt{X(1-X)}}.$$

In case that $a = b$ this results in a Student-t with $2a$ degrees of freedom. For $a < b$ we obtain left (negative) skewness and for $a > b$ the distribution is right skewed. The AG skewness equals $1 - 2B_z(a, b)$, where $B_z(a, b)$ is the regularised incomplete Beta function with $z = (a + 1/2)/(a + b + 1)$. In terms of moment existence, the right-hand tail behaves as $t^{-[2\min(a,b)+1]}$, whereas the left tail is always thinner in case of skewness and behaves like $t^{-[2\max(a,b)+1]}$. So tails will be proportional to those of a t-distribution with $2\min(a, b)$ degrees of freedom on the right and $2\max(a, b)$ degrees of freedom on the left. Finally, the distribution of T is always unimodal with mode given by

$$\frac{(a-b)\sqrt{a+b}}{\sqrt{(2a+1)(2b+1)}}.$$

We generate $n = 2000$ observations from this skew-t distribution with $a = 20$ and $b = 2$, which implies positive skewness with an AG skewness measure of 0.3645, tail behaviour characterised by $M_l^T = 40$, $M_r^T = 4$ and a mode equal to 5.897.

We find that predictive results with the Bernstein-skew model are virtually identical for each of the five prior choices on the weights, and closely match the data histogram, as illustrated in Figure 6 (left panel). In line with the tail behaviour of the skew-t distribution, as explained above, the posterior probabilities of the extreme weights in the right tail being zero is virtually zero, and inference on the degrees of freedom parameter ν is centered on fairly small values (most of the posterior mass is in the interval (2,4) for all priors). The left tail weights, however, do get appreciable zero probabilities, consistent with the much thinner left tails of the sampling distribution. Table 1 shows that, especially for Priors 1,2 and 4, these posterior probabilities are substantially higher than their prior counterparts (which are chosen to be 0.2). Thus, overall, the tail behaviour of the sampling distribution is relatively well captured by the Bernstein-skew model.

The flexible unimodal construct leads to the predictive presented in the right panel of Figure 6, which is, again, less smooth than the Bernstein-skew predictive and follows the data quite closely. The fit to the data is perhaps less good in the lower left tail, where the

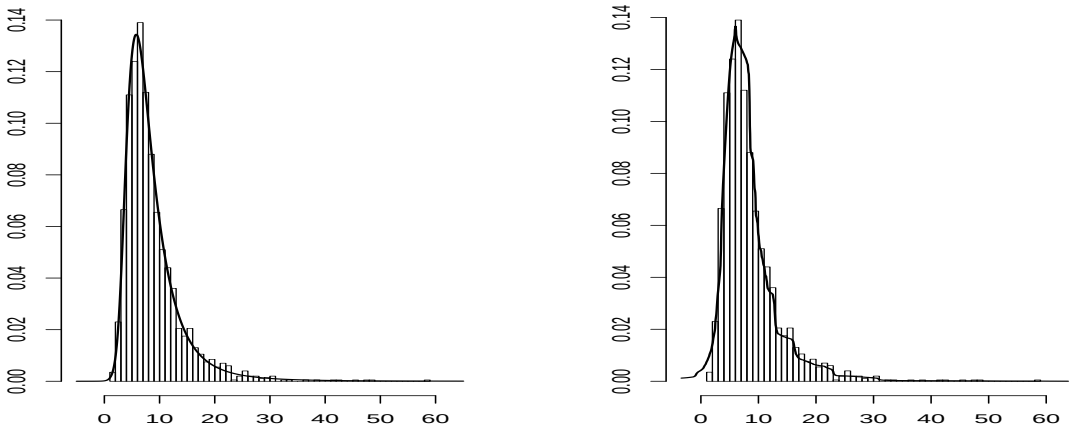

Figure 6: Simulated skew-t data: Posterior predictive densities, overplotted with the normalized data histogram. Left panel: Bernstein-skew model with Prior 5. Right panel: flexible unimodal construct.

Prior	$P(w_1^m = 0)$	$P(w_2^m = 0 \mid w_1^m = 0)$
Prior 1	0.99	0.95
Prior 2	0.91	0.61
Prior 3	0.49	0.27
Prior 4	0.70	0.47
Prior 5	0.42	0.22

Table 1: Simulated skew-t data, Bernstein-skew model: Posterior probabilities of zero weights in the left tail.

Bernstein-skew seems closer to the actual data. This seems a consequence of the fact that the skewing mechanism has a fairly simple parametric form (with only two parameters) and the rather extreme difference between the tails can not totally be accounted for (except for scaling by a factor r). Inference on skewness through the AG measure reveals a posterior distribution with median 0.332 and first and third quartiles 0.321 and 0.340, which is in line with the sample value of 0.329 (and a bit below the theoretical value of 0.3645). This right skewness originates to some extent from the central part of the distribution, but especially from the tails, as we can judge from the fact that r has posterior median 15.49 (with quartiles 15.07 and 16.05) while δ obtains posterior median 1.161 (quartiles 0.761 and 1.382). The posterior median of the mode μ is 5.78 (with quartiles 5.54 and 6.00), which fits with the theoretical value of 5.897. Also, the posterior median of M was 5.304 with quartiles 4.519

and 5.878, while for ζ we obtained a posterior median of 0.344 with quartiles 0.302 and 0.353. Thus, the form of the posterior predictive is a consequence of fairly spread atoms with relatively small weights each.

4 Real data applications

4.1 Galaxy data

These data are velocities of 82 distant galaxies, diverging from our own galaxy. They are described in some detail in Roeder (1990) and were used for both parametric and nonparametric mixture modelling (see Escobar and West, 1995, and Richardson and Green, 1997).

The normalized histogram of these data (see left panel of Figure 7) clearly shows the multimodal nature. This means we will need a fairly large amount of components, so we take $\lambda = 300$ in the prior for m in (5). The amount of flexibility that this entails means that we recommend fixing the degrees of freedom ν. We have used $\nu = 50$ in the plots presented, but $\nu = 5$ can equally well be used. This leads to differences in the inference on the weights w_i^m and μ and σ, but results in a virtually indistinguishable posterior predictive. The Bernstein-skew model with large m (the posterior mass for m is concentrated in between 210 and 260) is so flexible that the weights simply compensate for the tails of F and we can take pretty much any reasonable value for ν. Interestingly, the fact that we can learn only very little about ν is corroborated by the posterior probabilities of zero weights in the tails. The latter are not very different from the prior probabilities and they hardly change if we use $\nu = 5$ instead. However, we recommend against taking values much smaller than $\nu = 5$ as they tend to smooth out the posterior predictive too much. We have used the prior structure on the Dirichlet parameters in (6), but results with (7) are very similar. However, the use of exchangeable prior weights through $c_i^m = \sqrt{m}/50$ does not lead to a good fit of the data for $\nu = 50$, as weights remain too uniform and do not get close enough to zero (although this choice of the prior weights does fit well for $\nu = 5$).

4.2 Biomedical data

We now consider data on the body mass index of 202 Australian athletes. These data were used in *e.g.* Azzalini and Capitanio (2003) in a location-scale context and are included in the SN package available in R. Here we will use a linear regression model, with a constant term and three other covariates: red cell count, white cell count, and plasma ferritin concentration.

So, in the model (2) for Y_i, we replace μ by $x_i'\beta$, where x_i groups the covariate values for

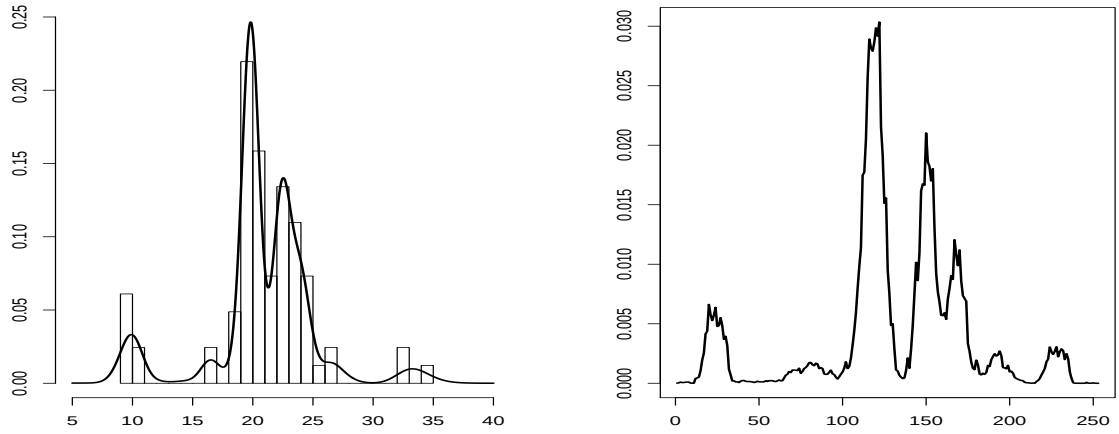

Figure 7: Galaxy data, Bernstein-skew model. Left panel: Posterior predictive overplotted with the histogram of the data. Right panel: posterior means of the weights. The degrees of freedom ν are fixed at 50, $\lambda = 300$, and the prior on the Dirichlet parameters is (6).

observation i and $\beta = (\beta_0, \ldots, \beta_3)$ is the four-dimensional vector of regression coefficients. The priors for the β_i's are independent normals with mean zero and variance 10,000.

We use the flexible unimodal construct, as this imposes unimodality and does not affect the mode, which is a desirable property in this regression context, which can then be interpreted as modal regression. Even though it was not designed for this case, we also apply the Bernstein-skew model to these data, with prior 5 as in (7) and two different priors on m, using $\lambda = 20$ and $\lambda = 100$. The resulting posterior densities for β and σ are presented in Figure 8, which also displays the posterior predictive density for the mean values of the covariates. Interestingly, inference on the regression coefficients corresponding to the three variable covariates is not affected much by the choice of model. The main differences are in the intercept β_0 and σ, but even these are not dramatically different. Of course, the latter parameters do not have the same interpretation for the Bernstein-skew and flexible unimodal construct models. Thus, despite its lack of unimodality constraints, the Bernstein-skew model leads to very similar inference on the parameters that can be compared across models $(\beta_1, \ldots, \beta_3)$ and, especially for the case with $\lambda = 100$, leads to very similar conditional predictive densities as the flexible unimodal construct.

Posterior inference on the degrees of freedom parameter ν in the Bernstein-skew model clearly indicates very heavy tails for the regression error term, with the first three quartiles equal to 1.16, 1.59, 2.21 for the case with $\lambda = 20$ and 0.90, 1.53, 2.38 for $\lambda = 100$. The

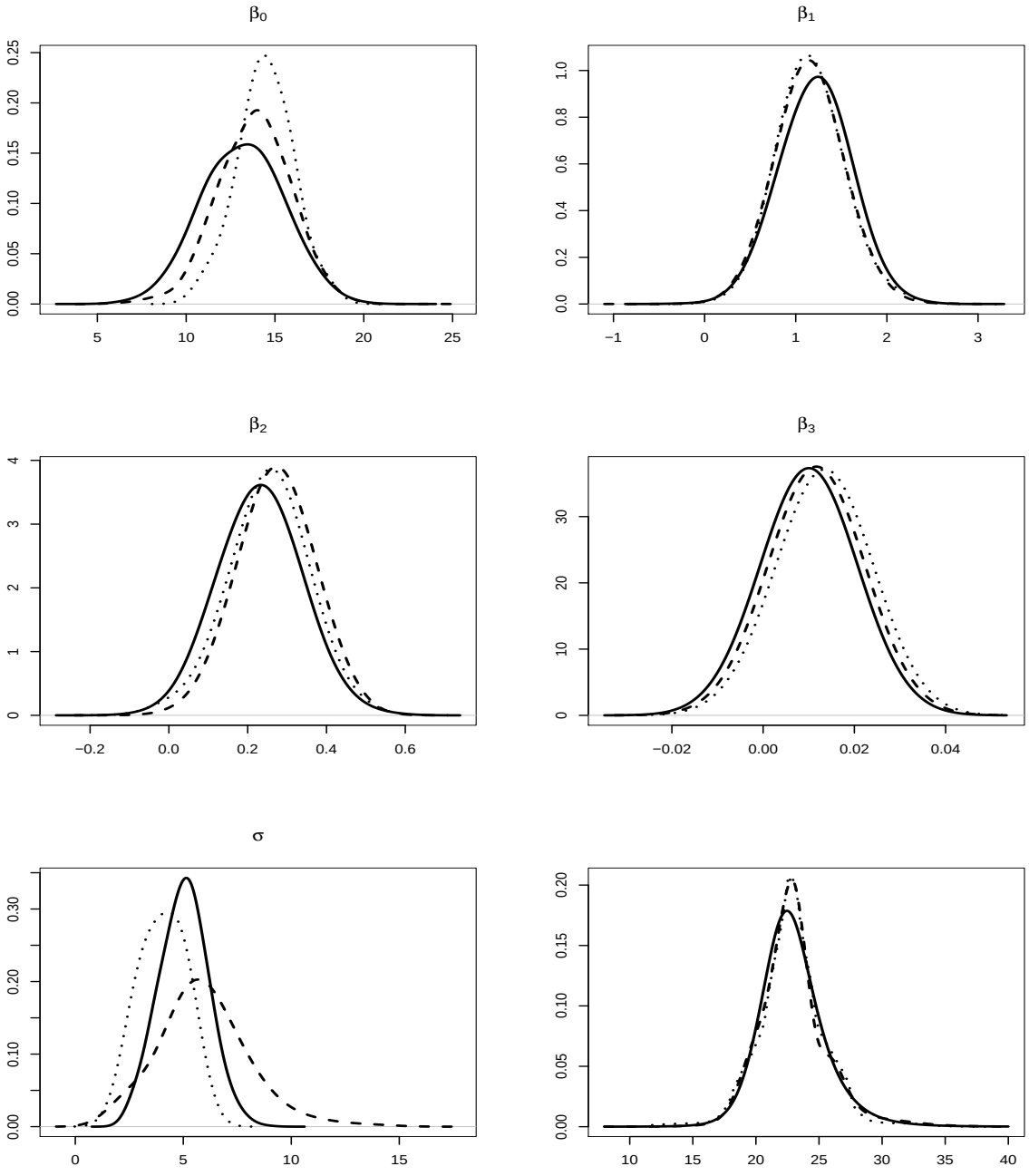

Figure 8: Biomedical data: Posterior densities of regression coefficients and σ. Lower right panel: posterior predictive density. Models used are the Bernstein-skew model with Prior 5 and $\lambda = 20$ (solid line), the Bernstein-skew model with Prior 5 and $\lambda = 100$ (dashed line) and the flexible unimodal construct (dotted line).

number of components used in the Bernstein-skew model does vary quite a bit with the prior assumptions, as it ranges from 8 to 30 for $\lambda = 20$ and from 93 to 133 if we adopt $\lambda = 100$. However, the resulting predictives and the inference on the other parameters are quite similar, with the model with many components generally closer to the flexible unimodal construct. So even with $\lambda = 100$ there is no suggestion that serious violations from unimodality have occurred. This is confirmed by Figure 9, which presents the posterior mean of the skewing mechanisms for all three models. Both Bernstein mean skewing mechanisms suggest unimodality, with a slight squeeze in the tails, especially the left tail. Yet, there is no strong evidence of zero weights, so this does not really counteract the very small values for ν we found. It also shows that the skew-Bernstein model with $\lambda = 100$ almost retains the mode of F (since p has a single mode at approximately 0.5), so that the regression function can approximately be interpreted as the mode, whereas there is a more noticeable shift for $\lambda = 20$ (which affects the posterior for β_0 in Figure 8). Note also the bump in the right hand shoulder of p for $\lambda = 100$, which is reflected in a similar bump in the predictive, mimicking that of the flexible unimodal construct. Clearly, the flexibility of F is what drives inference in the flexible unimodal construct, with very little added through the skewing mechanism. This is in line with a small amount of skewness. Indeed, this model identifies only a slight positive skew in the error distribution with median AG skewness equal to 0.09 and first and third quartiles 0.05 and 0.14, with small contributions of both the centre of the distribution and the tails. Interestingly, the posterior median of ζ was 0.923, with first and third quartiles 0.696 and 1.207 while for M the posterior median was 3.639 with quartiles 2.727 and 4.828. Again, the posterior weights are fairly small but the atoms are not so dispersed as in the previous examples using simulated data, which leads to a relatively smooth predictive pdf.

5 Recommendations and Conclusions

Based on a constructive representation, we have proposed two flexible modelling strategies for univariate continuous distributions. The general constructive representation we use separates the skewing mechanism P from the underlying symmetric distribution F. The Bernstein-skew model is defined by a nonparametric model for P through Bernstein densities, in combination with a Student t_ν specification for F. This can generate very general distributions and is ideal for density estimation. The flexible unimodal construct adopts a parametric specification for P and takes a very flexible form for F through a nonparametric mixture of uniforms. This imposes unimodality with control over the mode and was designed with modal regression modelling in mind.

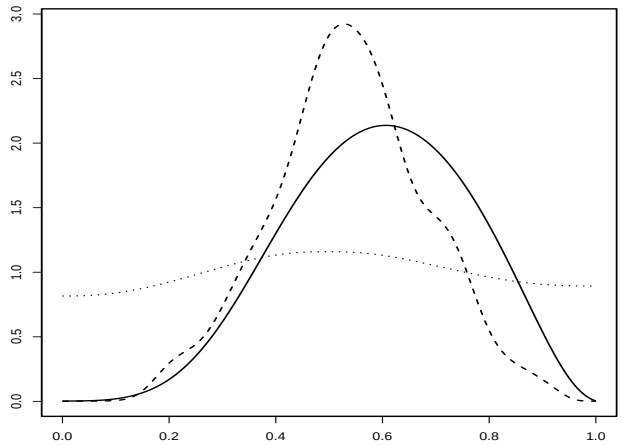

Figure 9: Biomedical data: Posterior means of p, the pdf of the skewing mechanism. Models used are the Bernstein-skew model with Prior 5 and $\lambda = 20$ (solid line), the Bernstein-skew model with Prior 5 and $\lambda = 100$ (dashed line) and the flexible unimodal construct (dotted line).

Our constructive approach helps us in more precisely identifying the key features of the data for which a simple model would be inappropriate. In particular, the Bernstein-skew model allows us to investigate moment existence and also leads to an easy calculation of the Bayes factor in favour of the symmetric model F. The flexible unimodal construct leads to straightforward inference on skewness, both emanating from the centre and the tails of the distribution. For both models, considering the skewing mechanism p is helpful in finding out how mass is shifted with respect to the underlying symmetric model F.

The question naturally arises which of the two proposed models to use in which situations. For multimodal data, we would naturally use the Bernstein-skew model, as it alone can accommodate multimodality. In cases of relatively "regular" (*e.g.* smooth, bimodal) data shapes, we would recommend to use relatively small λ and free ν. For situations with more irregular data, we think larger λ should be used (to induce sufficient flexibility) in combination with keeping ν fixed at an intermediate value. If the data are unimodal, the best model to use depends on the interests of the investigator: if the interest is primarily in density estimation or in the tail behaviour, it may still be best to use the Bernstein-skew model. If the main interest of the analysis is in inference on the mode or the skewness properties, the flexible unimodal construct would be the best option.

Note, finally, that the Bernstein-skew model generally leads to smooth predictives and the flexible unimodal construct tries to get as close to the data as it can, without any

clear preference for smoothness. However, the (sparse) parametric form of the skewing mechanism underlying the flexible unimodal construct does not always lead to sufficient flexibility. Overall, the Bernstein-skew model seems more useful for density estimation (it leads to smoother, more flexible shapes and can be used for multimodality) and should be recommended generally, except in those cases where we really wish to impose unimodality and where we are mainly interested in the location (as in modal regression) or in direct inference on skewness.

References

Arnold, B. C. and Groeneveld, R. A. (1995), "Measuring skewness with respect to the mode," *American Statistician*, 49, 34–38.

Azzalini, A. (1985), "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics*, 12, 171–178.

Azzalini, A. and Capitanio, A. (2003), "Distributions generated by permutations of symmetry with emphasis on a multivariate skew-t distribution," *Journal of the Royal Statistical Society, B*, 65, 367–389.

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

Fernández, C. and Steel, M. F. J. (1998), "On Bayesian modeling of fat tails and skewness," *Journal of the American Statistical Association*, 93, 359–371.

Ferreira, J. T. A. S. and Steel, M. F. J. (2006), "A constructive representation of univariate skewed distributions," *Journal of the American Statistical Association*, 101, 823–829.

Genton, M. G. and Loperfido, N. (2005), "Generalized skew-elliptical distributions and their quadratic forms," *Annals of the Institute of Statistical Mathematics*, 57, 389–401.

Jones, M. C. (2004), "Families of distributions arising from distributions of order statistics" (with discussion), *Test*, 13, 1–43.

Jones, M. C. and Faddy, M. J. (2003), "A skew extension of the t-distribution, with applications," *Journal of the Royal Statistical Society B*, 65, 159–174.

Petrone, S. (1999), "Bayesian density estimation using Bernstein polynomials," *The Canadian Journal of Statistics*, 27, 105–126.

Petrone, S. and Wasserman, L. (2002), "Consistency of Bernstein polynomial posteriors," *Journal of the Royal Statistical Society* B 64, 79–100.

Richardson, S. and Green, P.J. (1997), "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society* B 59, 731–792 (with discussion).

Roeder, K. (1990), "Density estimation with confidence sets exemplified by superclusters and voids in the galaxies," *Journal of the American Statistical Association*, 85, 617–624.

A Proofs

Proof of Theorem 1. If $\lim_{y\to-\infty}\frac{p[F(y)]}{|y|^b}$ is finite then, as P is a continuous distribution, it is possible to find $0 < K_1 \leq K_2 < \infty$ such that $K_1|y|^b \leq p[F(y)] \leq K_2|y|^b$ and thus

$$K_1\int_{\Re_-}|y|^{r+b}f(y)dy \leq \int_{\Re_-}|y|^r dH(y|F,P) \leq K_2\int_{\Re_-}|y|^{r+b}f(y)dy.$$

Now, as the largest moment of F is M^F, the left and right-hand side of equation above are finite if and only if $r+b < M^F$ and consequently $M_l^S = M^F - b$. □

Proof of Theorem 2. We shall apply Theorem 1 and, thus, consider the limiting behaviour of $p[F(y)]$. For a Student t distribution, $F(y)$ is

$$F(y) = 1 - \frac{1}{2}B_z(\nu/2,1/2), \text{ if } y>0, \text{ and } F(y) = \frac{1}{2}B_z(\nu/2,1/2), \text{ otherwise,}$$

where $1/z = [1+(y^2/\nu)]$ and $B_t(a,b)$ is the incomplete Beta function, regularized such that $B_1(a,b) = B(a,b)$. As we are considering limits for $y^2 \to \infty$ we need to consider small values of z. For these values $B_z(\nu/2,1/2)$ will behave like $z^{\nu/2}$ and thus like $|y|^{-\nu}$. As a consequence, in the extreme left tail we will obtain

$$p[F(y)] \approx \sum_{j=1}^{m} w_j^m \frac{1}{B(j,m-j+1)}|y|^{-\nu(j-1)},$$

so that if $w_1^m > 0$, we see that $p[F(y)]$ will behave like a constant and $M_l^H = M^F = \nu$. However, if the first k weights are zero, the relevant behaviour of $p[F(y)]$ will be as $|y|^{-k\nu}$ and using Theorem 1 we then obtain $M_l^H = M^F + k\nu = (k+1)\nu$. Results for the right-hand tail are derived in a totally analogous manner. □