On a transform for modeling skewness

Li Kang[*], Paul Damien[†] and Stephen G Walker[‡]

Abstract

In many applications, data exhibit skewness. In this paper we present a transform providing skewness, along similar lines to which one would use transforms within a class of density to obtain location and scale. Hence, in order to model data to include location, scale, skewness and shape, we need only to find a family of densities exhibiting a variety of shapes, since we can obtain location, scale and skewness via the transformations. The chosen class of density with the variety of shape is the simplest available and presented in the paper. Illustrations including regression and time series models are given.

Keywords: Auxiliary variable; Bayesian inference, Markov chain Monte Carlo, Mixtures of uniforms, Transform.

1 Introduction

Several applications involve data which exhibit unimodality, skewness and kurtosis, typically heavier tails compared to the normal distribution; see [22] for a comprehensive discussion. The quality of statistical inference heavily depends on the assumed probability model or distributions. To this end, considerable effort has been expended towards the construction of flexible, parametric classes of probability distributions aimed at better modeling real data.

Many, if not most, modeling ideas start with the normal distribution and extend to include shape and scale parameters. For example, introducing

[*]Department of Statistics and Data Sciences, University of Texas at Austin
[†]McCombs Schools of Business, University of Texas at Austin
[‡]Department of Mathematics, University of Texas at Austin

skewness to the normal distribution was done by [3], [4], [5]. The skew normal density is given by

$$f(x) = 2\,\phi(x)\,\Phi(\lambda x), \tag{1}$$

where ϕ denotes the standard normal density function, Φ the corresponding standard normal distribution function, and $\lambda \in \mathbb{R}$ is the skewness parameter, i.e. $\lambda = 0$ results in the normal distribution. [8] extended this distribution to scale mixtures of skew–normal (henceforth SMSN) distributions to allow for heavier tails. One such member is the skewed–t distribution, proposed by [12]; however, the family fails to capture all shapes, such as sharp peaks at the mode. The asymmetric Laplace distribution, introduced by [26], overcomes this, but on the other hand it only accommodates modes which are spiked; i.e. non–differentiable. Our family of densities include both spiked and non-spiked modal behavior.

Another class of distributions which encapsulates smooth and sharp modes, skewness and heavy tails, is the Skewed Exponential Power distribution (SEPD), proposed and studied by [11], [2], [17], [22], [9], among others. The SEPD is derived from the Exponential Power distribution ([21]; [7]) and is known by many names—Generalized Power Distribution (GPD), Generalized Error Distribution (GED), or Generalized Laplace Distribution (GLD). [27] further extend the SEPD to separately model the two tails, and a recent Bayesian analysis of the SEPD was undertaken by [18].

We start with a new construction of skew family distributions. Suppose a random variable X has location equal to 0, scale equal to 1, and zero skewness. If another random variable X' has the following representation,

$$X' = \mu + X \qquad\qquad X' = \sigma\, X, \tag{2}$$

then we can say that μ is the location parameter and σ is the scale parameter. Mimicking the logic of defining a location–scale family, for a skew family, we will motivate the following transform; for some $\tau > 0$,

$$X' = \widetilde{X} \times \begin{cases} \tau & \widetilde{X} > 0 \\ 1/\tau & \widetilde{X} < 0, \end{cases} \tag{3}$$

where

$$\widetilde{X} =_d \begin{cases} X_+ & \text{with probability} \quad \tau^2/(1+\tau^2) \\ X_- & \text{with probability} \quad 1/(1+\tau)^2. \end{cases}$$

Here X_+ is the variable X constrained to be positive and X_- the variable X constrained to be negative. We write this transform succinctly as $X' = X(\tau)$.

The motivation behind this transform are as follows; the basic idea is to accelerate the density on one side and to deccelerate it on the other side; see Figure 1. Any skew transform must do such, and our claim is that (3) is the easiest way to achieve it. The density function for X' in terms of the densities of X_+ and X_- is given by

$$f_{X'}(x) = \frac{2\tau^2}{1+\tau^2}\,\frac{1}{\tau}\,f_{X_+}(x/\tau) + \frac{2}{1+\tau^2}\,\tau\,f_{X_-}(x\tau),$$

which implies the value of the density at 0 is $f(0)\,\tau/(1+\tau^2)$, where $f(0) = f_{X_+}(0) = f_{X_-}(0)$, and so $f_{X'}$ is continuous. In fact, if

$$f_X(x) = \tfrac{1}{2}\int_0^\infty \mathbf{1}(-z < x < z)\,k(z)/z\,dz$$

for some density k on $(0,\infty)$, a representation which exists due to [16], then

$$f_{X'}(x) = \int_0^\infty \frac{\mathbf{1}(-z/\tau < x < z\,\tau)}{\tau + 1/\tau}\,k(z)/z\,dz.$$

If X is unimodal; then $X = U/Z$ for U uniform on $(0,1)$ and Z, independent of U, defined on $\mathbb{R}$. This characterization is due to [16]. Elegantly, we have $X' = U/Z'$, where Z' is easy to characterize, and which is based on the transformed Z, and so X' is also unimodal. Therefore, we define the skewed-location-scale family based on the original X as,

$$X' = \mu + \sigma\,X(\tau). \tag{4}$$

Given how we move to location, scale, and skewness, we will also consider shape as well shortly, we need to find a suitable choice for the original random variable X which, while has zero location, scale 1 (though it is not so important to be 1) and zero skewness, will come with some existing shape parameter that we denote by α. We will denote the density of our starting X as $k_\alpha(x)$.

To motivate a particular choice of $k_\alpha(x)$, we argue that one should start thinking about scale mixture of uniforms due to the characterization of [16]. For obvious reasons, a simple mixture is needed. To this end, consider a representation of the standard normal density as a mixture of uniforms:

$$\phi(x|z) = U(x \mid -\sqrt{z}, \sqrt{z}) \qquad\qquad z \sim \Gamma(\tfrac{3}{2}, \tfrac{1}{2}), \tag{5}$$

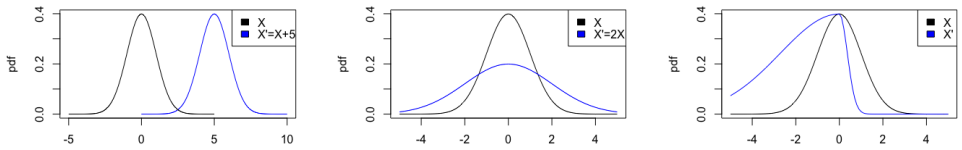

Figure 1: Location-Scale-Skew
X from Normal(0,1); The right most panel is showing the density for
$X' = X(\exp 1)$ when $X > 0$, and $X' = X(\exp(-1))$ when $X < 0$

where $\Gamma(a, b)$ denotes a gamma density with mean a/b and variance a/b^2, and $U(a, b)$ denotes the uniform density on (a, b). It is easy to check that $\phi(x) = \int_0^\infty \phi(x|z)\,\Gamma(dz|3/2, 1/2)$. Our proposed extension of (5) to include shape is simple; generalize the gamma density to have parameters $(\alpha, 1/2)$, namely to change its shape parameter. This shape property directly transfers to the induced density for x. Hence, take

$$k_\alpha(x) = \int \phi(x|z)\,\Gamma(dz|\alpha, \tfrac{1}{2}). \tag{6}$$

This representation is examined in Section 2.

The proposed full family can also be derived from the nonparametric unimodal model described in [20] who constructed a family of nonparametric unimodal densities. The *parametric* family introduced in this paper—a Gamma Mixture of Uniform distributions (GMU)—is the mean density from the nonparametric model. According to [10], all unimodal densities can be represented as a scale mixture of uniform distributions. Hence, the previously mentioned families are members. The mixture of uniforms is useful in many contexts for a variety of reasons, as discussed, for example, in [14]. Hence, from this perspective, the most efficient mixture should be used: we show how such efficiency is achieved via the new GMU family introduced in this paper. Importantly, in our gamma mixture the ease with which the parameters can be interpreted is of practical value.

For now, note that the GMU family includes the Laplace distribution and the normal distribution as special cases. Furthermore, [23] showed that the Exponential Power family can be represented as a scale mixture of normals; of course, the normal distribution and the scale mixture of normals ([1]) are special cases of uniform scale mixtures.

To complete the notation, let $N(\mu, \sigma^2)$ and $tN(\mu, \sigma^2)$ the normal and left truncated normal distributions with mean μ and variance σ^2, respectively; likewise, $t\Gamma(\alpha, \beta)$ represents the left truncated gamma distributions with shape α and rate β.

Basic properties and a simulation study of the GMU are given in Section 2 where we compare GMU with the skew–normal, skew–t, asymmetric Laplace, and SEPD families. Simulation results show that the GMU family better captures features of the true underlying density when compared to the other families listed above. Bayesian inference is discussed in Section 3. Section 4 illustrates GMU in some broad classes of regression and time–series models, followed by illustrative analyses in section 5. A brief discussion is provided in Section 6.

2 Gamma Mixtures of Uniform densities

In this section, the model provided in the Introduction is further studied; i.e., we use X to have density (6) and to model with X' given by (4). In practice, we take $\tau = \exp(\lambda)$. Putting all the transformations together, our new class of asymmetric, unimodal densities, named as Gamma Mixture of Uniform distributions (GMU), can be written as follows:

Theorem. *If*

$$[X|Z = z] \sim U\Big(\mu - z\exp(-\lambda),\ \mu + z\exp(\lambda)\Big) \quad \textit{and} \quad Z \sim \Gamma(\alpha, \beta),$$

then $X \sim \text{GMU}(\mu, \alpha, \beta, \lambda)$ *with probability density function* $f(x)$ *is given by*

$$f(x) = \frac{\beta\,\text{sech}(\lambda)}{2(\alpha - 1)}\,\gamma\big\{r(x - \mu, \lambda)\big\}, \tag{7}$$

where

$$\gamma\big\{r(x-\mu, \lambda)\big\} = \int_{r(x-\mu,\lambda)}^{\infty} \Gamma(z|\alpha-1, \beta)\,dz \quad \textit{and} \quad r(x, \lambda) = \max\big\{\mu\exp(\lambda), x\exp(-\lambda)\big\}.$$

Proof. The joint density function of (x, z) can be written as

$$f(x, z) = \tfrac{1}{2}\,\text{sech}(\lambda)\ z^{-1}\,\frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1}\exp(-\beta z)\ \mathbb{1}\Big(z > \max\big\{(\mu-x)\ \exp(\lambda), (x-\mu)\ \exp(-\lambda)\big\}\Big),$$

and let $z(x) = \Big\{z : \ z > \max\big\{(\mu - x)\ \exp(\lambda), (x - \mu)\ \exp(-\lambda)\big\}\Big\}$. Integrating out z, we get

$$f(x) = \tfrac{1}{2}\,\mathrm{sech}(\lambda)\,\frac{\beta^\alpha}{\Gamma(\alpha)}\int_{\Omega(x)} z^{\alpha-2}\exp(-\beta z)\,dz = \frac{\beta\ \mathrm{sech}\,\lambda}{2(\alpha-1)}\int_{\Omega(x)} \Gamma(z|\alpha-1,\beta)\,dz$$

□

It is easy to see that the GMU family is unimodal. Its maximum (i.e. the mode) is achieved when $x = \mu$. In fact, the idea of introducing skewness using the τ and $1/\tau$ terms is not new, as it was used by [13], though our motivation via the transformation is new. It is the use of the "'simple" gamma distribution which distinguishes the two approaches leading to an important new result proved below. A well–known result ([10]) is that scale mixtures of uniform distributions coincide with the class of unimodal distributions; for example, [20] exploit this within the framework of Bayesian nonparametric inference. The parametric family of interest here turns out to be the mean of the nonparametric family.

Writing GMU as a scale mixture of uniform distributions shows how the shape parameter, α, and scale parameter, β, of the gamma density in the mixing kernel transfer directly to the shape and scale parameters for the data model. Also, μ and λ are clearly the location and skewness parameters, respectively. This simple, yet important, identification/interpretation of the parameters is an appealing aspect of the GMU representation.

Compared to the above GMU representation, to rewrite the skewed normal distribution in (1) using scale mixture of uniforms, one needs to introduce two auxiliary variables u and z, with $f(x|u,z)$ given by

$$(8)\quad U\big(x \mid \max(-\sqrt{u}, \tfrac{z}{\lambda}), \sqrt{u}\big)\mathbb{1}(\lambda > 0) + U\big(x \mid -\sqrt{u}, \max(\sqrt{u}, \tfrac{-z}{\lambda})\big)\mathbb{1}(\lambda < 0)$$

with $u \sim \Gamma(3/2, 1/2)$ and $z \sim \phi$. Further, [13] construct their skew model as

$$(9)\qquad f(x) = \frac{2}{\gamma + 1/\gamma}\big[g(x\gamma)\mathbb{1}(x > 0) + g(x/\gamma)\mathbb{1}(x < 0)\big],$$

where g is a unimodal symmetric density on $\mathbb{R}$. We can write this as a scale mixture of uniforms,

$$(10)\qquad f(x|z) = U(x \mid -\tfrac{z}{2\gamma}, \tfrac{z\gamma}{2}) \quad \text{with} \quad z \sim \widetilde{g},$$

for some density $\widetilde{g}$ on $(0,\infty)$ where, for $x>0$, $g(x)=\int_{z>x} z^{-1}\,\widetilde{g}(z)\,dz$, and g is symmetric. In fact, [13] choose g to be a standard Student–t density with ν degrees of freedom. However, we argue that the above model is best understood as a scale mixture of uniforms. In order to recover the Student–t density for g, the $\widetilde{g}$ is of the form $\widetilde{g}(z)\propto z^2\,(1+z^2/\nu)^{-(\nu+3)/2}$, which is not standard. Given that location and skewness are a part of $f(x|z)$, we actually only need $\widetilde{g}$ to be a density with scale and shape parameters, the simplest of which is the gamma density.

Before examining theoretical features, it is instructive to visualize the density to better understand the role of the parameters collected together in $\theta=(\mu,\alpha,\beta,\lambda)$. Consider Figure 2. The curve's smoothness and differentiability at the mode and tails are controlled by $\alpha>1$; larger values result in heavier tails and smoothness at the mode. The scale is accounted for by $\beta>0$. Symmetry, left, and right skewness are attained when λ equals zero, less than zero, and greater than zero, respectively. Setting $\mu=0,\lambda=0,\alpha=2$ results in the Laplace $(0,1/\beta)$ distribution. Also, GMU$(0,2,3.5,0)$ is "very close" to the standard normal distribution but with slightly heavier tails. Consider this generalization of the GMU density given above in order to include the standard normal distribution in the GMU family. For some $\delta>0$, introduce the uniform component as

$$f(x|z)=\frac{1}{z^{\delta}(e^{-\lambda}+e^{\lambda})}\mathbf{1}\big(-z^{\delta}e^{-\lambda}<x<z^{\delta}e^{\lambda}\big)\quad\text{with}\quad z\sim\Gamma(\alpha,\beta).$$

Now $\alpha=\beta=\delta=\frac{1}{2}$ and $\lambda=0$ recover the standard normal density. This suggests we can enlarge the GMU family meaningfully by using a power z^{δ} in the uniform component, but we will not explore that possibility here.

3 Likelihood analysis

The likelihood function for n observations, $\mathbf{y}=(y_1,\ldots,y_n)$, from the GMU density in (7) is given by

(11)

$$l(\mu,\alpha,\beta,\lambda;\mathbf{y})=\frac{\beta^{n\alpha}}{\{\exp(\lambda)+\exp(-\lambda)\}^{n}\Gamma(\alpha)^{n}}\prod_{i=1}^{n}\left\{\int_{r(y_i-\mu,\lambda)}^{\infty}z^{\alpha-2}\exp(-\beta z)dz\right\},$$

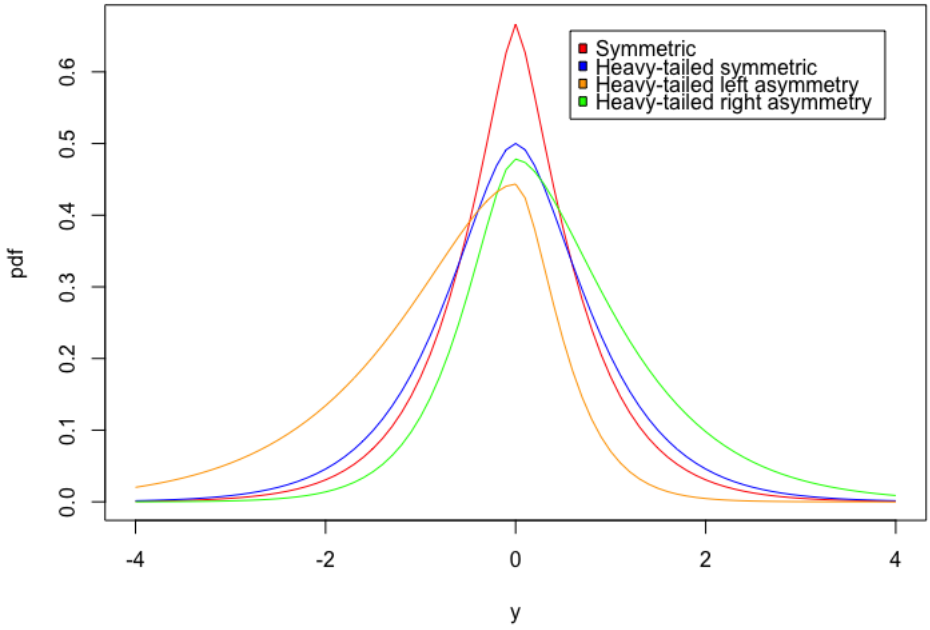

Figure 2: Symmetric and asymmetric GMU density
Parameter values for red curve: $\mu = 0, \alpha = 2.5, \beta = 2, \lambda = 0$; blue curve: $\mu = 0, \alpha = 3, \beta = 2, \lambda = 0$; green curve: $\mu = 0, \alpha = 3, \beta = 2, \lambda = 0.3$; orange curve: $\mu = 0, \alpha = 3, \beta = 2, \lambda = -0.5$

with log–likelihood,

(12)
$$\log l(\mu, \alpha, \beta, \lambda; \mathbf{y}) = n\alpha \log(\beta) - n \log\{\exp(\lambda) + \exp(-\lambda)\} - n \log \Gamma(\alpha) + \sum_{i=1}^{n} \log \left\{ \int_{r(y_i - \mu, \lambda)}^{\infty} z^{\alpha - 2} \exp(-\beta z) dz \right\}.$$

There is no analytical solution for the maximum likelihood estimators of the parameters of the GMU density. However, it is possible to obtain the MLE, $\widehat{\theta} = (\widehat{\mu}, \widehat{\alpha}, \widehat{\beta}, \widehat{\lambda})$, via Newton's method; see, for example, [25]. It is straightforward to obtain the gradient and Hessian matrix from (11) to implement Newton's method.

Underlying Density	Sample Size	Estimation Density	D
t_5	50	SN	0.0858
		SK-t	0.0760
		AEPD	0.0876
		GMU	0.0578
	500	SN	0.0486
		SK-t	0.0213
		AEPD	0.0272
		GMU	0.0211
SL(0,1,0.55)	50	SN	0.1009
		SK-t	0.1464
		AEPD	0.0876
		GMU	0.0578
	500	SN	0.0741
		SK-t	0.1413
		AEPD	0.0268
		GMU	0.0184
SK-t(5,1.5)	50	SN	0.0792
		SK-t	0.0678
		AEPD	0.0830
		GMU	0.0663
	500	SN	0.0529
		SK-t	0.0216
		AEPD	0.0302
		GMU	0.0247

Table 1: Simulation Study Comparison
SL: Skewed Laplace; SN: Skewed Normal; $SK-t$: Skewed t-Distribution; t: t-Distribution; AEPD: Asymmetric Exponential Power Distribution; GMU: Gamma Mixture of Uniforms; D: Average Maximum Distance between True and Estimated CDF

Here we compare the GMU density model with a number of alternative models using synthetic data. We generate 50 and 500 data points 100 times, respectively, from a Student–t distribution with five degrees of freedom; a skewed Laplace distribution ([26]) with skewness equal to 0.55; and a skewed–

t distribution (SK–t) with parameters 5 and 1.5. Density estimation for the generated samples is done using the GMU, SEPD, Skewed Normal (SN), and SK–t densities, and based on maximum likelihood estimation. We calculate the average maximum distance (D) between the true cumulative distribution function and the estimated cumulative distribution function.

Consider Table 1: In all the simulations, GMU outperforms the other models, barring only one instance, when the data are generated from a skewed-t, and when the sample size is 500. As expected, in this latter case, the skewed–t cumulative distribution function estimates are better. Importantly, when the data have heavy tails, under both symmetric and asymmetric cases, the GMU density outperforms SEPD. Results for data from the skewed Laplace distribution show that, unlike SK–t, the GMU density also does well when the data are spiked at the mode. This is because, as detailed earlier, the GMU density explicitly accounts for smoothness at the mode.

4 Bayesian analysis

To implement Bayesian analysis using Markov chain Monte Carlo (MCMC), the form of the density given in equation (7) is not useful due to the presence of a gamma integral. To obviate this difficulty, following [19], an auxiliary variable construct of the GMU density should prove helpful. To this end, introduce a latent variable z and write the joint density for y and z as

$$f(y,z) = \frac{\beta\,\mathrm{sech}(\lambda)}{2(\alpha-1)}\left[\frac{\beta^{\alpha-1}}{\Gamma(\alpha-1)}z^{\alpha-2}\exp(-\beta z)\mathbb{1}\{z \geq r(y-\mu,\lambda)\}\right]. \tag{13}$$

The density in (7) is recovered upon integrating out z. Let $\mathbf{z} = (z_1, \ldots, z_n)$ and write the complete likelihood function given in (11) using the above auxiliary variable representation:

(14)
$$l(\mu,\beta,\alpha,\lambda;\mathbf{y},\mathbf{z}) = \frac{\beta^{n\alpha}\exp(-\beta\sum_{i=1}^n z_i)}{\{\exp(\lambda)+\exp(-\lambda)\}^n\Gamma(\alpha)^n}\prod_{i=1}^n\left[z_i^{\alpha-2}\mathbb{1}\{z_i \geq r(y_i-\mu,\lambda)\}\right].$$

Next, we show how this form of the GMU$(\mu,\alpha,\beta,\lambda)$ density is amenable to full Bayesian inference. We set μ=0 to indicate a mode at 0, which will be our typical setting.

Since we plan to use a Gibbs sampler, conditional distributions, up to proportionality, are needed. With $p(\cdot)$ denoting a prior distribution, in the following we take gamma priors for β and α, and a uniform prior for λ. These prior choices were made to incorporate any potential prior information, and also so that the conditional distributions can be more readily sampled. From an inference perspective there is no loss in generality in making these selections, for one simply sets the prior parameter values to reflect diffuse or, if need be, informative beliefs. In the illustrative analyses, we work with diffuse prior choices. To obtain the conditional densities, we employ (14) along with the appropriate prior distributions detailed below. With $p(\beta) = \Gamma(c, d)$, the conditional distribution for β is the following gamma distribution,

(15)

$$p(\beta|\lambda, \alpha, \mathbf{y}, \mathbf{z}) \propto \beta^{n\alpha+c-1} \exp\left\{-\beta\left(d + \sum_{i=1}^{n} z_i\right)\right\} \equiv \Gamma\left(n\alpha + c, d + \sum_{i=1}^{n} z_i\right).$$

Since α has to be greater than 1, we use a gamma prior, $p(\alpha) = \Gamma(c', d')$, appropriately truncated, leading to,

$$p(\alpha|\lambda, \beta, \mathbf{y}, \mathbf{z}) \propto \frac{\beta^{n\alpha}}{\Gamma(\alpha)^n} \exp\left\{(\alpha - 2)\sum_{i=1}^{n} \log(z_i)\right\} p(\alpha). \tag{16}$$

The conditional density for α is not regular. Therefore, to draw posterior samples for α, we use a Metropolis–Hastings algorithm. For the proposal density, choose a truncated normal distribution with mean the current value of α, and with variance σ_α^2 that will be specified in the illustrations.

Finally, we take a flat prior for λ on $\mathbb{R}$. The conditional density for λ is given by

$$p(\lambda|\beta, \alpha, \mathbf{y}, \mathbf{z}) \propto (\operatorname{sech} \lambda)^n \, \mathbb{1}(\lambda \in A), \tag{17}$$

where

$$A = \left(\max_{y_i>0}\{-\log(z_i/y_i)\}, \min_{y_i<0}\{\log(-z_i/y_i)\}\right).$$

Given the above three conditional distributions, one now samples α, β, and λ iteratively via the following:

Algorithm 1 Bayesian inference for the parameters of the GMU density

1: Initialize α, β, λ. Then sequentially sample step 2-5 until convergence:
2: z_i from $\Gamma(\alpha - 1, \beta)$, left truncated on $r(y_i, \lambda) = \max\big\{ -y_i \exp(\lambda), y_i \exp(-\lambda)\big\}$.
3: β from $\Gamma(n\alpha + c, d + \sum_{i=1}^{n} z_i)$.
4: α via Metropolis–Hastings with the proposal from $\mathrm{N}(\alpha, \sigma_\alpha^2)$, restricted to $(1, \infty)$.
5: λ via Metropolis–Hastings with proposal coming uniformly from the set A.

5 Applications

The GMU density is now illustrated for some wide classes of models including linear regression with non–normal errors, the GARCH model, and Dynamic Linear Models.

5.1 Linear regression with non–normal errors

Consider the model

$$y = \mathbf{XB} + \epsilon \quad \text{with} \quad \epsilon \sim \mathrm{GMU}(0, \alpha, \beta, \lambda), \tag{18}$$

where $\mathbf{X}$ is an $n \times p$ covariate matrix and $\mathbf{B} = (b_1, \ldots, b_p)^T$ is a $p \times 1$ coefficient vector. The likelihood function for n observations with latent variable, z, is given by

$$l(\mathbf{B}, \alpha, \beta, \lambda; \mathbf{y}, \mathbf{z}) = \frac{\beta^{n\alpha} \exp(-\beta \sum_{i=1}^{n} z_i)}{\{\exp(\lambda) + \exp(-\lambda)\}^n \Gamma(\alpha)^n} \prod_{i=1}^{n} \Big[z_i^{\alpha-2} \mathbb{1}\{z_i \geq r(y_i - \mathbf{X}_i\mathbf{B}, \lambda)\} \Big] \tag{19}$$

with

$$r(y_i - \mathbf{X}_i\mathbf{B}, \lambda) = \max\Big\{ -(y_i - \mathbf{X}_i\mathbf{B}) \exp(\lambda), (y_i - \mathbf{X}_i\mathbf{B}) \exp(-\lambda) \Big\}.$$

Since jointly estimating the regression coefficient vector $\mathbf{B}$ and the GMU density's parameters is not possible, MCMC is used. Given $\mathbf{B}$, sampling the GMU density's parameters proceeds along the lines of Algorithm 1; then sample $\mathbf{B}$ conditional on the other parameters via a Metropolis–Hastings algorithm. Specifically, given the new latent $\mathbf{z}$, the complete conditionals for

β and α are exactly the same as in (15) and (16). The conditional density for λ stays the same but subject to a new constraint,

$$A' = \left(\max_{y_i > \mathbf{X}_i\mathbf{B}} \{-\log[z_i/(y_i - \mathbf{X}_i\mathbf{B})]\}, \min_{y_i < \mathbf{X}_i\mathbf{B}} \{\log[-z_i/(y_i - \mathbf{X}_i\mathbf{B})]\} \right).$$

Since the regression coefficient vector $\mathbf{B}$ is only involved in the constraint of $\mathbf{z}$, in order to get component-wise posterior samples for $\mathbf{B}$, write

$$y_i - \mathbf{X}_i\mathbf{B} = y_i - \sum_{l \neq j} X_{il} b_l - X_{ij} b_j.$$

It is clear that the complete conditional for b_j is its prior distribution subject to a certain constraint $\Omega = (ll, uu)$ with

$$ll = \max \left\{ \frac{y_i - \sum_{l \neq j} X_{il} b_l - z_i \exp(\lambda)}{X_{ij}}_{X_{ij} > 0, y_i - \sum_{l \neq j} X_{il} b_l > X_{ij} b_j}, \frac{y_i - \sum_{l \neq j} X_{il} b_l + z_i \exp(-\lambda)}{X_{ij}}_{X_{ij} < 0, y_i - \sum_{l \neq j} X_{il} b_l < X_{ij} b_j} \right\}$$

$$uu = \min \left\{ \frac{y_i - \sum_{l \neq j} X_{il} b_l - z_i \exp(\lambda)}{X_{ij}}_{X_{ij} < 0, y_i - \sum_{l \neq j} X_{il} b_l > X_{ij} b_j}, \frac{y_i - \sum_{l \neq j} X_{il} b_l + z_i \exp(-\lambda)}{X_{ij}}_{X_{ij} > 0, y_i - \sum_{l \neq j} X_{il} b_l < X_{ij} b_j} \right\}.$$

The entire algorithm is detailed in Appendix B.

5.2 GARCH

The GARCH model is widely used in applications, especially in modeling financial asset returns; for example, see [27] and the references therein. Here we use the GMU density to model a GARCH(1,1) process where, for $i = 1, \ldots, n$, the response variable, $y = (y_i)$, and volatilities, $\sigma^2 = (\sigma_i^2)$, are modeled as

$$y_i = \sigma_i \epsilon_i \quad \text{with} \quad \epsilon_i \sim \text{GMU}(0, \alpha, \beta, \lambda)$$

and

$$\sigma_i^2 = b_0 + b_1 y_{i-1}^2 + b_2 \sigma_{i-1}^2.$$

Using our auxiliary variable representation, the complete likelihood function for n observations is given by

(20)

$$l(b_0, b_1, b_2, \alpha, \lambda, \sigma; \mathbf{y}, \mathbf{z}) = \frac{\beta^{n\alpha} \exp(-\beta \sum_{i=1}^n z_i)}{\left(\prod_{i=1}^n |\sigma_i| \right) \{\exp(\lambda) + \exp(-\lambda)\}^n \Gamma(\alpha)^n} \prod_{i=1}^n \left[z_i^{\alpha - 2} \mathbb{1} \left\{ z_i \geq r\left(\frac{y_i}{\sigma_i}, \lambda \right) \right\} \right]$$

$$\text{with} \quad r\left(\frac{y_i}{\sigma_i}, \lambda\right) = \max\left\{ -\left(\frac{y_i}{\sigma_i}\right)\exp(\lambda), \left(\frac{y_i}{\sigma_i}\right)\exp(-\lambda)\right\}.$$

To ensure that the GARCH process is stationary, the following two constraints have to be imposed. First, the variance of the error density must equal one, which is obtained by setting:

$$\beta = \sqrt{\frac{\alpha(\alpha+1)\cosh(3\lambda)\,\mathrm{sech}(\lambda) - 3\sinh^2(\lambda)\alpha^2}{3}}. \tag{21}$$

Second, [6] noted that $b_1 + b_2 < 1$ is another necessary condition. The estimate of the parameter vector $\mathbf{\Phi}$, where $\mathbf{\Phi} = (b_0, b_1, b_2, \alpha, \lambda)$, is obtained by Bayesian inference. Given the new auxiliary variable $\mathbf{z}$, the complete conditional for α and λ is similar to (5.2) with β replaced by (21). The new sample constraint for λ is given by:

$$A'' = \left(\max_{\frac{y_i}{\sigma_i}>0}\{-\log[\sigma_i z_i/y_i]\}, \min_{\frac{y_i}{\sigma_i}<0}\{\log[-\sigma_i z_i/y_i]\} \right).$$

Since b_0, b_1 and b_2 are only involves in σ, the complete conditional for b_0, b_1 and b_2 can be reduced to,

$$p(b_0, b_1, b_2|\alpha, \lambda, \sigma; \mathbf{y}, \mathbf{z}) \propto \frac{1}{\prod_{i=1}^{n}|\sigma_i|} \prod_{i=1}^{n}\left\{ z_i \geq r\left(\frac{y_i}{\sigma_i}, \lambda\right)\right\}.$$

Since b_0 has constraint greater than 0, we use the normal distribution as both the prior and proposal distributions for b_0 truncated from 0. Due to the joint constraint, $b_1 + b_2 < 1$, we assign Dirichlet priors for b_1 and b_2, and again jointly propose $\mathrm{c}(b_{1,j+1}, b_{2,j+1})$ through Dirichlet $(\xi b_{1,j}, \xi b_{2,j}, \xi(1 - b_{1,j}, b_{2,j}))$, where j indicates the current state of the Markov chain, and ξ controls the spread of the proposal distribution—larger ξ returns a more concentrated density. Note that besides the above constraints, the new proposed b_0, b_1 and b_2 should also make each σ_i satisfy $z_i \geq r\,(y_i/\sigma_i, \lambda)$.

5.3 Dynamic linear models

The class of DLMs is very useful in both regression and time series analysis; see [24]. Comprising two equations, labeled observation and system, the error structure in both are assumed to follow different normal distributions. The advantage of the normality assumption is that it leads to the well-known Kalman filtering equations which provide an easy way to estimate parameters, as well as forecast values of the response variable. However, in a vast number of applications (such as asset pricing models), the observation equation is better modeled using a non-Gaussian error structure in order to account for skew and high levels of kurtosis. Such generalizations come at a price, for the Kalman filtering equations are no longer applicable. Here, we model the observation equation's error term as a GMU density and outline a solution to obtain parameter estimates and forecast values.

Let $\mathbf{T}$ denote the total number of time periods; $\mathbf{Y}$ is the vector of values of the dependent variable; $\mathbf{X}$ is the matrix of regressors; and $\mathbf{B}$ is a p-vector of regression parameters. To avoid over-fitting, a random walk model for the state equation is usually assumed in a DLM. At each time period t, we have:

$$y_t = \mathbf{X}_t\mathbf{B}_t + \epsilon_t \qquad \epsilon_t \sim \text{GMU}(0, \alpha, \beta, \lambda) \tag{22}$$

$$\mathbf{B}_t = \mathbf{B}_{t-1} + v_t \qquad v_t \sim N(0, W), \tag{23}$$

where W is a $p \times p$ diagonal covariance matrix. Bayesian inference for this model involves sampling iteratively from the following two steps.

1. Sample $\alpha, \beta, \lambda, W$ from $p(\alpha, \beta, \lambda, W|\mathbf{B}, \mathbf{X}, \mathbf{Y})$
2. Sample $\mathbf{B_t}$ from $p(\mathbf{B_t}|\alpha, \beta, \lambda, \mathbf{B_{-t}}, \mathbf{W}, \mathbf{X}, \mathbf{Y})$

Given $\mathbf{B}$, we use Algorithm 1 detailed earlier to sample all the parameters of the GMU density, if W is assumed known, as is typically done. If W is unknown, and if one assumes an Inverse–Wishart prior distribution, then the conditional posterior distribution is also an Inverse–Wishart. Predicting future values of y_t proceeds in a manner similar to the linear regression example discussed in the next section.

6 Illustrations

A simulated example is first detailed via a non-normal linear regression model, followed by a GARCH application using real data.

6.1 Linear regression

Consider the model

$$y_i = b_0 + b_1 x_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim \text{GMU}(0, \alpha, \beta, \lambda), \tag{24}$$

where $x_i \sim_{iid} \text{N}(0,1)$ for $i = 1, \ldots n$, with $n = 100$. The true values used to generate the errors, prior parameter values reflecting diffuse beliefs, posterior estimates, and the Geweke statistics used to assess MCMC convergence are presented in Table 2. Note that the prior distributions for α and β are gamma distributions (see earlier discussion) with diffuse hyper-parameter choices. The table also shows the Ordinary Least Squares (OLS) estimates for the regression parameters. The MCMC algorithm was carried out with $50,000$ iterations, a burn-in of $10,000$, and a thinning length of 500 samples for λ. The Geweke standard normal statistic, which tests the hypothesis that the means from the first 10% and last 50% of the samples from the posterior distributions of the parameters are the same, was used to assess convergence. If the MCMC chain has converged, the null is not rejected. It is evident from Table 2 that all parameters satisfy this test.

Parameters	α	β	λ	b_0	b_1
True values	3	2	0.5	0.3	0.5
Prior distribution values	$\Gamma(3,1)$	$\Gamma(3,1)$	–	–	–
Posterior means	2.66	1.66	0.48	0.35	0.61
95% intervals	(1.58,4.28)	(1.06,2.42)	(0.25,0.71)	(0.14,0.65)	(0.47,0.74)
Geweke statistics	-0.16	-0.04	0.55	-0.67	-0.62
OLS estimates	-	-	-	1.14	0.59

Table 2: Linear Regression Example with GMU Errors
We use $\Gamma(shape, rate)$ representation here for the prior distributions of α, β; non–informative prior for λ, b_0, b_1; OLS is the Ordinary Least Squares estimates.

To sample a one–step ahead predicted value, y_{n+1}, we simply use the posterior mean values of the model parameters. Hence, y_{n+1} has a $\text{GMU}(0, \widehat{\alpha}, \widehat{\beta}, \widehat{\lambda})$ distribution with

$$r(y_{n+1} - \widehat{b}_0 - \widehat{b}_1 x_{n+1}) = \max\left\{-(y_{n+1} - \widehat{b}_0 - \widehat{b}_1 x_{n+1}) \exp\widehat{\lambda}, (y_{n+1} - \widehat{b}_0 - \widehat{b}_1 x_{n+1}) \exp(-\widehat{\lambda})\right\}.$$

Figure 3 shows the trace plots for all posterior samples, Figure 4 plots the resulting posterior distributions of the model parameters, and Figure 5 shows the predictive density with true value shown in red and an approximate 95% prediction interval via dashed lines.

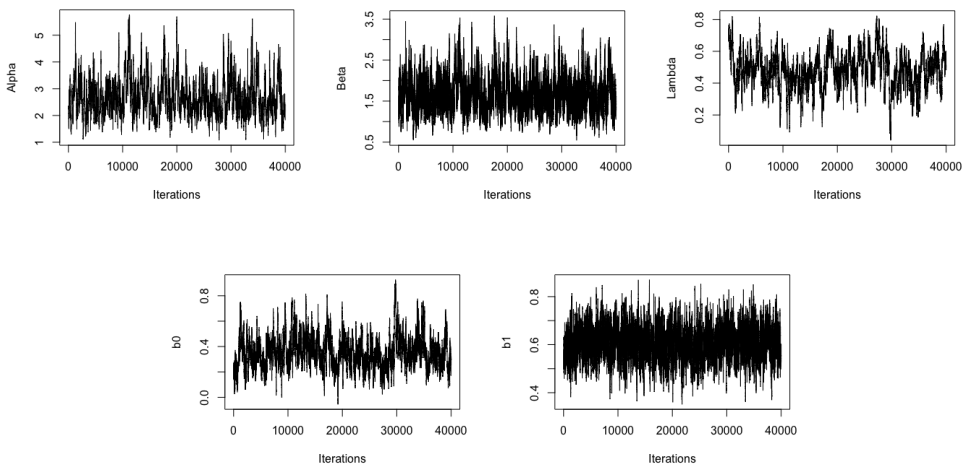

Figure 3: Linear Regression Example: posterior samples trace plots

6.2 GARCH

We applied the GARCH model to a stock returns dataset with $n = 735$ data points. The sample comprises the log daily return on the $S\&P$ 500 Index from Jan 2nd 2015 to Nov 30th 2017 obtained from the Wharton Research Services Database. The MCMC algorithm was carried out with 1,000,000 iterations, with a burn-in of 200,000, and thinning of 100. The convergence diagnostics are similar to the ones shown for the previous example and are therefore omitted. The posterior summaries are in Table 3.

Parameters	Prior distribution values	Posterior means	95%CI
α	$t\Gamma(3, 0.5)$	2.24	(1.86,2.69)
λ	-	0.028	(-0.02,0.08)
b_0	tN(0.006,0.0001)	0.0004	(0.0002,0.00058)
b_1	Dir(10,30,1)	0.24	(0.17,0.33)
b_2	Dir(10,30,1)	0.73	(0.65,0.81)

Table 3: GARCH Model with GMU Errors
We use $t\Gamma(shape, rate)$ representation for the truncated gamma prior distribution of α with hyper-parameter values depicting diffuse beliefs; non-informative priors for λ; and truncated normal prior, $\text{tN}(\mu, \sigma)$, for b_0; Dirichlet prior for b_1 and b_2.

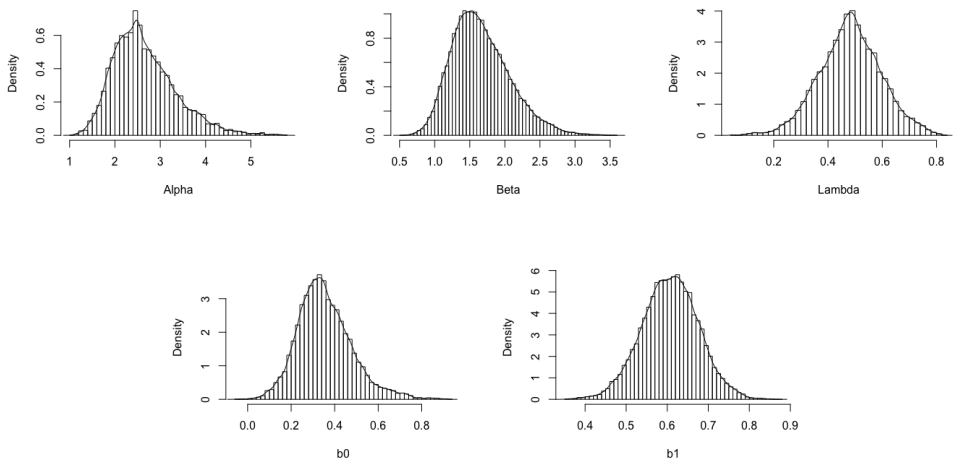

Figure 4: Linear Regression Example: Posterior distributions of model parameters

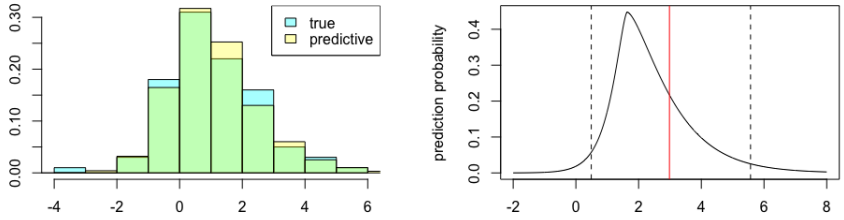

Figure 5: Linear Regression Example: Left panel: general predictive density, the overlapping of the two histograms turns to green; Right panel: One-step ahead prediction distribution based one left out X, with Red Line: actual value; Dashed Lines: 95% intervals

Figure 6 shows the one-step ahead predictive density and its response to varying bin sizes. Our prediction nicely recovers the underlying distribution implied by the true data. As the number of bins increases, the true and estimated distributions are virtually indistinguishable.

For comparison, the GARCH(1,1) model was also executed using the

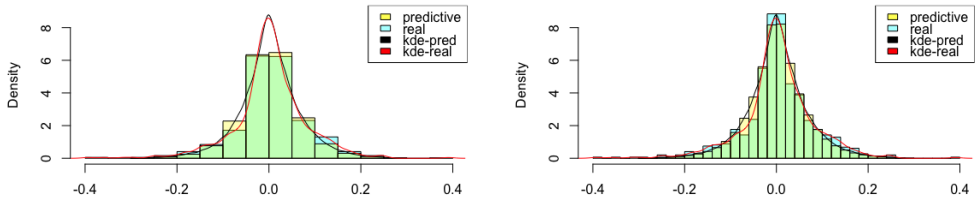

Figure 6: GARCH Model: predictive density
More bins in left panel and less bins in right panel; The overlapping of the two histograms turn green.

SEPD and skewed-t densities using the bayesDccGarch package in R. Table 4 shows the Mean Squared Errors (MSEs) under SEPD , skewed-t and GMU errors for the predictions using ten held-out samples. D is the statistic from a Kolmogorov-Smirnov test. This is a comparison of cumulative distribution functions where smaller D indicates smaller maximum difference between two CDFs. GMU returns the smallest MSE and D values (correspondingly largest p-value). Figure 7 shows the comparison of the predictive densities.

	GMU	SEPD	SK-t
Mean Squared Error	0.00087	0.00088	0.00089
D	0.034	0.056	0.11
p-value	0.3550	0.02656	$3.2e^{-08}$

Table 4: Estimated MSEs for 10 Out-of-Sample Predictions Using the GARCH Model;
$SK-t$: Skewed t-Distribution; SEPD: Skewed Exponential Power Distribution; GMU: Gamma Mixtures of Uniform. D is the statistic from a Kolmogorov-Smirnov test with its corresponding p-value.

7 Discussion

This paper develops a new parametric class of unimodal densities—a Gamma Mixture of Uniform Distributions—with applications to both regression and

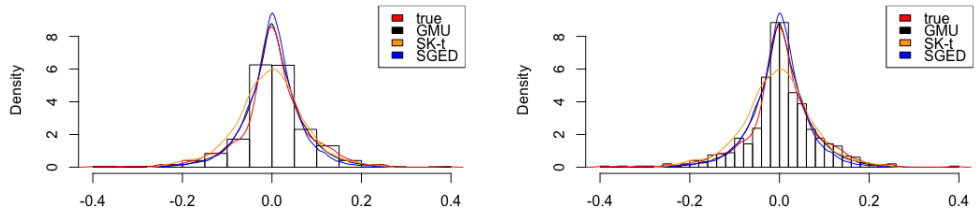

Figure 7: GARCH Model: Predictive density comparison

time-series models. The GMU family offers a practitioner the flexibility to model asymmetric/symmetric and heavy/light tailed data distributions: in this regard, a simulation example shows that when compared to other commonly used families, the proposed class performs better. A real data illustration further attests to this fact. One key reason is the new family, which exploits a result due to [10], has fewer restrictions stemming from its representation as a scale mixture of uniform distributions.

By employing a combination of data augmentation techniques and Gibbs sampling, full Bayesian inference in broad classes of popular models (Linear Regression, GARCH, Dynamic Linear Models, etc.) is possible. It is worth emphasizing that the full conditional distributions for the GMU density's parameters are largely shared by several broad classes of models, such as the ones detailed in this paper. In other words, MCMC implementation could be seamlessly adapted to obtain full Bayesian inference in a wide range of applied problems.

Extensions of the central idea in this paper would first involve developing a multivariate GMU representation. A starting point here would require a multivariate gamma distribution with

$$f(y|\omega) = \prod_{j=1}^{p} \frac{1}{\omega_j(e^{\lambda_j} + e^{-\lambda_j})} \mathbf{1}\big(-\omega_j e^{-\lambda_j} < y_j < \omega_j e^{\lambda_j}\big) \quad \text{with} \quad \omega \sim \Gamma_p,$$

where Γ_p denotes a p–dimensional gamma density; see for example [15]. Second, theoretical properties of Bayes estimates in multivariate formulations would be challenging. Finally, the asymmetric Laplace distribution is used in Bayesian quantile regressions to better model the distribution of a response

variable. Given that the GMU density was shown to outperform popular asymmetric families of densities, it would be worthwhile to engage the GMU family in quantile regressions.

References

[1] D.F. Andrews and C.L. Mallows, *Scale mixtures of normal distributions*, Journal of the Royal Statistical Society. Series B (Methodological) (1974), pp. 99–102.

[2] A. Ayebo and T.J. Kozubowski, *An asymmetric generalization of Gaussian and Laplace laws*, Journal of Probability and Statistical Science 1 (2003), pp. 187–210.

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

[4] A. Azzalini, *Further results on a class of distributions which includes the normal ones*, Statistica 46 (1986), pp. 199–208.

[5] A. Azzalini and A.D. Valle, *The multivariate skew-normal distribution*, Biometrika 83 (1996), pp. 715–726.

[6] T. Bollerslev, *Generalized autoregressive conditional heteroskedasticity*, Journal of Econometrics 31 (1986), pp. 307–327.

[7] G.E. Box and G.C. Tiao, *Bayesian Inference in Statistical Analysis*, Vol. 40, John Wiley & Sons, 2011.

[8] M.D. Branco and D.K. Dey, *A general class of multivariate skew-elliptical distributions*, Journal of Multivariate Analysis 79 (2001), pp. 99–113.

[9] P. Delicado and M. Goria, *A small sample comparison of maximum likelihood, moments and l-moments methods for the asymmetric exponential power distribution*, Computational Statistics & Data Analysis 52 (2008), pp. 1661–1673.

[10] W. Feller, *An Introduction to Probability Theory and Its Applications: Volume 2*, John Wiley & Sons, 1971.

[11] C. Fernandez, J. Osiewalski, and M.F. Steel, *Modeling and inference with υ-spherical distributions*, Journal of the American Statistical Association 90 (1995), pp. 1331–1340.

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

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

[14] T. Fung and E. Seneta, *A characterization of scale mixtures of the uniform distribution*, Statistics & Probability Letters 78 (2008), pp. 2883–2888.

[15] E. Furman, *On a multivariate gamma distribution*, Statistics & Probability Letters 78 (2008), pp. 2353–2360.

[16] A.Y. Khintchine, *On unimodal distributions*, Izvestiya Nauchno-Issledovatel?skogo Instituta Matematiki i Mekhaniki 2 (1938), pp. 1–7.

[17] I. Komunjer, *Asymmetric power distribution: Theory and applications to risk measurement*, Journal of Applied Econometrics 22 (2007), pp. 891–921.

[18] L. Naranjo, C.J. Pérez, and J. Martín, *Bayesian analysis of some models that use the asymmetric exponential power distribution*, Statistics and Computing 25 (2015), pp. 497–514.

[19] R.M. Neal, *Slice sampling*, Annals of Statistics 31 (2003), pp. 705–767.

[20] C.E. Rodríguez and S.G. Walker, *Univariate Bayesian nonparametric mixture modeling with unimodal kernels*, Statistics and Computing 24 (2014), pp. 35–49.

[21] M.T. Subbotin, *On the law of frequency of error*, Mat. Sb. 31 (1923), pp. 296–301.

[22] P. Theodossiou, *Skewed generalized error distribution of financial assets and option pricing*, Multinational Finance Journal 19 (2015), pp. 223–266.

[23] M. West, *On scale mixtures of normal distributions*, Biometrika 74 (1987), pp. 646–648.

[24] M. West and J. Harrison, *Bayesian Forecasting & Dynamic Models*, Springer, New York City, 1999.

[25] T.J. Ypma, *Historical development of the Newton–Raphson method*, SIAM review 37 (1995), pp. 531–551.

[26] K. Yu and J. Zhang, *A three-parameter asymmetric Laplace distribution and its extension*, Communications in Statistics-Theory and Methods 34 (2005), pp. 1867–1879.

[27] D. Zhu and V. Zinde-Walsh, *Properties and estimation of asymmetric exponential power distribution*, Journal of Econometrics 148 (2009), pp. 86–99.