The Scaled Beta2 Distribution as a robust prior for scales, and a Explicit Horseshoe Prior for locations

Pérez, María Eglée and Pericchi, Luis Raúl
Department of Mathematics and Center for Biostatistics and Bioinformatics,
University of Puerto Rico, Rio Piedras, San Juan, PR

Abstract

In the tradition of "Conjugate Priors" analysis (and extensions to Conditional Conjugate Priors), priors for locations are assumed Normal and for precisions are assumed Gamma or equivalently for variances Inverted Gamma. Following an idea of deFinnetti (1961), by regarding each observation (or set of observations) as having its own precision taken from a super-population of scales distributed as a Gamma, the Student distribution, as a scale-mixture of Normals, replaces the Normal Gaussian Model. This procedure adds various forms of robustness to priors and models. In a parallel manner, it is natural to think of the sum of squares as having a *conditional* Gamma distribution, and mixing scales again by a gamma distribution we obtain what we call the Scaled Beta2 Distribution (SBeta2). We propose here this distribution as an alternative to the ubiquitous Inverted-Gamma IG prior for variances and standard deviations, or more generally for scales. Vague IG priors as a quasi-non informative prior for reciprocal of scales with very low hyper-parameters, say $IG(0.001, 0.001)$ percolates most of Bayesian Statistics. It is known however, as pointed out by Gelman (2006) (but see Berger (2006) to mathematically explain the phenomenon) that far from being quasi non-informative the "vague" IG, leads to very low variances of the effects and very strong shrinkages to the general mean. Gelman proposes Uniform and Half-Cauchy priors. In the same vein, we propose here the Scaled Beta2 family. Advantages of the SBeta2 are numerous, among them are: its flexibility, and for particular values of the hyper-parameters it can be as heavy or heavier tailed as the half Cauchy, and also different behavior at the origin can be modeled. Furthermore it is easy to simulate from, and it can be inside a Gibbs sampling scheme. Besides, both the scale or the square of the scale, (or their reciprocals) cam be modeled as a SBeta2, both for Normals or other Likelihoods like in the Student-t family. There is an additional bonus: when the conditional prior for the location is Cauchy for example, and the prior for the scale, is SBeta2,the marginal for the location can be obtained proportional of well known functions, and for integer values of the hyper-parameters of the SBeta2, the marginal densities for locations and cumulative distributions can be found explicitly. This adds insight in the analysis and make it easier to elicit the prior scale, which is very convenient in general and for Empirical Bayes modeling. When the scale is modeled as a

SBeta2, the marginal prior location has the shape of a horseshoe prior, Carvalho et. al (2010) and to the best of our knowledge, it is the first horseshoe prior that can be shown in closed form. We call these priors, the explicit Horseshoe priors. Horseshoe priors are also specifically well suited for full Bayes Hierarchical modeling, leading to strong shrinkage when there are is no conflict between data and priors, and discounting the prior when there is conflict. Alternatively, if instead of the scale, the square of the scale is assumed as a SBeta2, then Fúquene, Pérez and Pericchi (2011) obtained what is called a Student-Scaled Beta2 marginal prior, known in closed form, which is also very convenient as a heavy tailed prior for the location.

1 Introduction

The focus of this paper is to propose the Scaled Beta 2 Distribution SBeta2, as a convenient prior for standard deviations and variances (or more generally scales and squares of scales) and precisions, and for both informative and quasi-non-informative scenarios. The intention is to reduce the inordinate over use of the Inverted Gamma distribution.
We justify the suggestion of using the SBeta2 based on:

1. Flexibility of both tail thickness and behavior at the origin. This allows the user to add robustness and spread of the factors.

2. For particular values of hyper-parameters, is safer as a quasi-uninformative prior.

3. It is easy to simulate from SBeta2 in various ways, and it is possible in cases to include it into a Gibbs Sampler schema.

4. When coupled with a conditional Cauchy prior for locations, the marginal prior for locations can be found explicitly, as proportional to known transcendental functions, and for integer values of the hyper-parameters they are in analytical closed forms. Furthermore, for specific choices, the marginal is found to be an explicit Horseshoe prior (Carvalho et al, 2010) with a pole at the origin and heavy tail, leading to a sort of nearly optimal choice as a prior for sparse locations. this seems to be the first explicit Horseshoe prior in the literature.

5. Closed form results, also facilitates the elicitation process of hyper-parameters.

6. A different behavior at the origin, without a pole, of the marginal density of the location, is obtained if instead of the scale the square of the scale is modeled as a SBeta2. This strategy leads also to a very useful prior distribution, called the Student-Sbeta2 distribution, Fúquene, Pérez and Pericchi (2011).

The Beta2 Distribution is rather well known and can be written as,

$$Beta2(\psi|p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \cdot \frac{\psi^{(p-1)}}{(\psi+1)^{(p+q)}}, \text{ for } \psi > 0$$

which can be easily simulated as the odds $\psi = \frac{\varpi}{1-\varpi}$, and $\varpi \sim Beta(\varpi|p,q)$.
We will be working instead with the scaled version of the Beta2:

$$SBeta2(\psi|p,q,b) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)\cdot b} \cdot \frac{(\frac{\psi}{b})^{(p-1)}}{((\frac{\psi}{b})+1)^{(p+q)}}, \text{ for } \psi > 0, b > 0,$$

which can easily be simulated as an scaled odds of Beta random variables. An alternative way to simulate from a SBeta2 will be developed in the sequel.

2 Motivation

Perhaps the most natural motivation in Bayesian statistical modeling for the SBeta2 is in a Meta Analysis. Suppose we have k centers.Define $\bar{x}_i = \bar{x}_{fi} - \bar{x}_{pi}$, $S_i^2 = (n_{fi}-1)s_{fi}^2 + (n_{pi}-1)s_{pi}^2$ are the available summary quantities for the data.

Model I

The usual likelihood (non-robust) assumption is:

- **Exponential Family Likelihood (Non-Robust)**

$$\begin{aligned} d_i|\delta_i, \sigma_{Wi}^2 &\sim N\left(\delta_i, \sigma_{Wi}^2\left(\frac{1}{n_{fi}} + \frac{1}{n_{pi}}\right)\right), i = 1, \ldots, k \\ SS_i|\sigma_{Wi}^2 &\sim \text{Gamma}\left(\frac{n_{fi}+n_{pi}-2}{2}, \frac{1}{2\sigma_{Wi}^2}\right), i = 1, \ldots, k \end{aligned}$$

This is usually related with "vague" assignments as follows:

- **Conjugate Prior and Hyper-Prior: Non Robust**

$$\begin{aligned} \delta_i|\Delta, \sigma_B^2 &\sim \text{Normal}(\Delta, \sigma_B^2), i = 1, \ldots, k \\ \sigma_{Wi}^2 &\sim \text{InvGamma}(0.01, 0.01), i = 1, \ldots, k \\ \Delta &\sim \text{Normal}(0, 10^5) \\ \sigma_B^2 &\sim \text{InvGamma}(3, 50) \end{aligned}$$

Partial attempt to Robustness

$$d_i|\sigma \sim St_{\nu_1}(\delta_i, \sigma^2(1/n_1 + 1/n_2)), i = 1, \ldots, k \quad (1)$$

$$\delta_i|\sigma \sim t_{\nu_2}(0, \sigma^2), \quad (2)$$

$$\sigma \sim IGamma(a, b), \tag{3}$$

The assessment allows the posterior to reject the prior of σ in favor of the data and the prior $\delta_i|\sigma$. (Andrade and O'Hagan 2006). Different combinations of hyper-parameters change the speed and direction of rejection of conflicting information.

Important Note: The Normal model with the "vague" Inverted Gamma prior (a and b very big), is the standard model in Objective Bayesian Statistics. This model has been strongly criticized by Gelman (2006) as producing unduly small posterior variances. On the other hand, an alternative, attempting to make the model more robust is to replace the Normal by a Student as above but leaving the same vague Inverted Gamma as a prior. This "robustification" is not enough and is subject to unsatisfactory posterior variances as the Normal Inverted Gamma.

Model II: A Comprehensive Robust Model: Student location as an scale mixture of Normals and Scale SBeta2 as a scale mixture of Gammas

We may "*robustify*" the normal by taking a scale mixture and integrating:

$$Student_\nu(\theta|\mu, \sigma^2) = \int Normal(\theta|\mu, \sigma^2/\rho) \cdot Gamma(\rho|\nu/2, \nu/2)d\rho.$$

We may proceed a similar path with the variances, as a scale mixture of Gammas.

$$SBeta2(\psi|p, q, b) = \int Gamma(\psi|p, b/\rho) \cdot Gamma(\rho|q, 1)d\rho,$$

see Section (3) for a proof. The procedure is justified, writing the likelihood on which each $\bar{x}_i$ and S_i^2 as with variance σ_B^2/ρ_i, and then the ρ_i are mixed, through a Gamma distribution: In a sense we mixed **also** the sum of squares and not only the sample means. This is a natural extension of the procedures that conduced to the student-t distribution. The (conditional) likelihood then reads: (see Appendix 7.1)

- Likelihood

$$\begin{aligned} d_i|\delta_i, \sigma_B^2, \rho_i &\sim \text{Normal}\left(\delta_i, \frac{\sigma_B^2}{\rho_i}\left(\frac{1}{n_{fi}} + \frac{1}{n_{pi}}\right)\right), i = 1, \ldots, k \\ SS_i|\sigma_B^2, \rho_i &\sim \text{Gamma}\left(\frac{n_{fi} + n_{pi} - 2}{2}, \frac{1}{2}\frac{\rho_i}{\sigma_B^2}\right), i = 1, \ldots, k \end{aligned}$$

- Prior

$$
\begin{aligned}
\rho_i &\sim \Gamma(p,q), i=1,\ldots,k \\
\delta_i|\Delta,\sigma_B^2 &\sim \mathrm{t}(\Delta,\sigma_B^2,9), i=1,\ldots,k \\
\Delta &\sim \mathrm{t}(0,10^5,5) \\
\sigma_B^2 &\sim \mathrm{SBeta2}(1,1,b), b>0, b\sim 3SS_B,
\end{aligned}
$$

where $b \sim 3SS_B$, is a rule of thumb discussed below.

3 Derivation of SBeta2 as a scale mixture of Gammas

In the sequel, we will use the following parametrization for the Gamma distribution: when $\psi \sim \Gamma(\alpha, b)$

$$
Ga(\psi|\alpha,b) = \frac{1}{\Gamma(\alpha)b^\alpha}\psi^{\alpha-1}\exp(-\frac{\psi}{b})
$$

With this parametrization, b is a scale parameter.

3.1 Scaled Beta2 distribution

We will take a scale mixture of Gammas as follows

$$
\begin{aligned}
\psi &\sim \mathrm{Gamma}(p, \frac{b}{\rho}) \\
\rho &\sim \mathrm{Gamma}(q,1)
\end{aligned} \tag{4}
$$

With this structure,

$$
\begin{aligned}
f(\psi,\rho) = Ga(\psi|\rho)Ga(\rho) &= \left\{\frac{\rho^p}{\Gamma(p)b^p}\psi^{p-1}\exp\left[-\frac{\psi}{b}\rho\right]\right\}\frac{1}{\Gamma(q)}\rho^{q-1}\exp[-\rho] \\
&\propto \psi^{p-1}\rho^{p+q-1}\exp\left[-\left(\frac{\psi}{b}+1\right)\rho\right]
\end{aligned}
$$

The marginal density of ψ is given by

$$\begin{aligned} f(\psi) &\propto \psi^{p-1} \int_0^\infty \rho^{p+q-1} \exp\left[-\left(\frac{\psi}{b}+1\right)\rho\right] d\rho \\ &\propto \psi^{p-1} \frac{\Gamma(p+q)}{\left(\frac{\psi}{b}+1\right)^{p+q}} \\ &\propto \frac{\psi^{p-1}}{\left(\frac{\psi}{b}+1\right)^{p+q}} \\ &= \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{1}{b} \frac{\left(\frac{\psi}{b}\right)^{p-1}}{\left(\frac{\psi}{b}+1\right)^{p+q}} \end{aligned} \tag{5}$$

If we transform $V = \psi/b$, it is easy to see that the distribution of V is a *Beta of the second type*, $V \sim Beta2(p,q)$. We will call the distribution given by (5) a *scaled Beta of the second type*, $\psi \sim SBeta2(p,q,b)$.

For this distribution,

$$\begin{aligned} E[\psi] &= bE[V] = \frac{p}{q-1} b \text{ when } q > 1 \\ Var[\psi] &= b^2 Var[V] = \frac{p(p+q-1)}{(q-1)^2(q-2)} b^2 \text{ when } q > 2 \end{aligned}$$

It is easy to see that if $\psi \sim SBeta2(p,q,b)$, then $\phi = \frac{1}{\psi} \sim SBeta2(q,p,\frac{1}{b})$.

4 Using the scaled Beta2 as a prior for variances and precisions

Let σ^2 be the unknown variance of a given distribution. Assign a scaled Beta prior to σ^2 as follows

$$\begin{aligned} \sigma^2 &\sim SBeta_2(p,q,b) \\ \Rightarrow h = \frac{1}{\sigma^2} &\sim SBeta_2(q,p,\frac{1}{b}) \end{aligned}$$

For selecting values for the hyper-parameters p, q and b, the following properties should be taken into account

1. $f_{\sigma^2}(0) = \begin{cases} \infty & p < 1 \\ \frac{q}{b} & p = 1 \\ 0 & p > 1 \end{cases}$

2. $f_h(0) = \begin{cases} \infty & q < 1 \\ pb & q = 1 \\ 0 & q > 1 \end{cases}$

3. When $q \leq 1$, $E(\sigma^2) = \infty$ Similarly, when $p \leq 1$, $E(h) = \infty$.

4. $\text{mode}(\sigma^2) = \begin{cases} \frac{p-1}{q+1} b & p \geq 1 \\ 0 & \text{Otherwise} \end{cases}$, $\text{mode}(h) = \begin{cases} \frac{q-1}{p+1} \frac{1}{b} & q \geq 1 \\ 0 & \text{Otherwise} \end{cases}$

5. When $p = q = 1$, the median of the SBeta2 distribution is the scale parameter b.

In general, we want to have heavy tails for a robust inference, but we don't want to give high weight to $\sigma^2 = 0$. So, our suggestion for selecting the hyper-parameters p, q and b are taking $p = 1$ or slightly larger, $q = 1$ (or slightly larger) and b large.
The challenge here is then to check that the analysis yields sensible results as $b \to \infty$, i.e. as it becomes vague.
Rule of thumb: The assessment of b should not be too small. In fact it should never be smaller than $\hat{\sigma}_B$. The rule of $b \simeq 3\hat{\sigma}_B$, seems to be working well in practice.
Mathematical Restrictions: In Berger (2006) it is explained that it τ^2 is the higher level variance in a hierarchical model then, $\pi(\tau^2) \propto \tau^{-2}$, leads to an improper posterior. On the other hand $\pi(\tau^2) \propto \tau^{-1}$ is a good objective prior, leading to proper posteriors. This means for the scaled beta 2 that if $p \geq 1/2$ then the behaviour at $\tau^2 = 0$, will be as τ^{-1} or will place less mass at zero. So a sensible restriction is that $p \geq 1/2$ and also if $q \geq 1/2$ the same will result for the assessment of the prior for the precision.

5 Cauchy-Scaled Beta 2

Let θ be a Cauchy random variable with location parameter 0 and scale parameter τ, and assume $\tau \sim SBeta2(p, q, b)$. Then, the joint distribution of θ and τ is given by

$$\begin{aligned} \pi(\theta, \tau) &= \frac{1}{\pi\tau\left(1 + \left(\frac{\theta}{\tau}\right)^2\right)} \frac{1}{\text{Beta}(p, q)} \frac{1}{b} \frac{\left(\frac{\tau}{b}\right)^{p-1}}{\left(\frac{\tau}{b} + 1\right)^{p+q}} \\ &= \frac{(b)^q}{\pi\text{Beta}(p, q)} \frac{\tau^p}{(\tau^2 + \theta^2)(b + \tau)^{p+q}} \end{aligned}$$

An application of this distribution in a hierarchical modeling scenario is in Pericchi and Pérez (2010).

The marginal distribution of θ can be calculated as

$$\pi(\theta) = \int_0^\infty \frac{(b)^q}{\pi\text{Beta}(p, q)} \frac{\tau^p}{(\tau^2 + \theta^2)(b + \tau)^{p+q}} d\tau$$

Note than when p and q are integers, the integrand is a rational function. For example, if $p = q = 1$,

$$\begin{aligned}\pi(\theta) &= \int_0^\infty \frac{b\tau}{\pi(b+\tau)^2\left(\theta^2+\tau^2\right)}\,d\tau \\ &= \frac{b}{\pi\left(b^2+\theta^2\right)^2}\left[-\left(b^2+\theta^2-\pi b|\theta|\right)+(b-\theta)(b+\theta)\left(\log(b)-\log(|\theta|)\right)\right]\end{aligned}$$

In this case, the cumulative distribution function also has a closed form, given by

$$\Pi(\theta) = \begin{cases} \frac{b\left(\pi b-\theta \mathrm{Log}\left[\theta^2\right]+2\theta\log[b]\right)}{2\pi(\theta^2+b^2)} & \theta<0 \\ \frac{1}{2} & \theta=0 \\ \frac{2\theta^2\pi+\pi b^2+2\theta b\log\left[\frac{b}{\theta}\right]}{2\pi(\theta^2+b^2)} & \theta>0 \end{cases}$$

For $p = 2$, $q = 1$,

$$\begin{aligned}\pi(\theta) &= \int_0^\infty \frac{b}{\pi\mathrm{Beta}(2,1)}\frac{\tau^2}{(\tau^2+\theta^2)(b+\tau)^3}d\tau \\ &= \frac{b\left(\left(b^2-3\theta^2\right)\left(\left(b^2+\theta^2-\pi b|\theta|\right)+\left(6b^2\theta^2-2\theta^4\right)\mathrm{Log}[b]+\left(-3b^2\theta^2+\theta^4\right)\mathrm{Log}\left[\theta^2\right]\right)}{\pi\left(b^2+\theta^2\right)^3}\end{aligned}$$

and the corresponding CDF is

$$\Pi(\theta) = \begin{cases} \frac{b\left(2\theta^3+3\theta^2\pi b+2\theta b^2+\pi b^3+\theta^3\log\left[\frac{1}{\theta^4}\right]+4\theta^3\mathrm{Log}[b]\right)}{2\pi\left(\theta^2+b^2\right)^2} & \theta<0 \\ \frac{1}{2} & \theta=0 \\ \frac{2\theta^4\pi+2\theta^3b+\theta^2\pi b^2+2\theta b^3+\pi b^4+4\theta^3 b\log\left[\frac{b}{\theta}\right]}{2\pi\left(\theta^2+b^2\right)^2} & \theta>0 \end{cases}$$

From the data, $b = 0.5283597$

6 Appendix

6.1 Derivations from the Likelihood

Assume we have $i = 1, \ldots, k$ sufficient statistics results from k centers:

$$\bar{x_i} \sim \mathtt{N}(\bar{x_i}|\mu_i, \sigma^2/(n_i \cdot \rho_i)),$$

$$S_i^2 = \frac{\sum(x_{ij}-\bar{x_i})^2}{n_i-1},\ \text{then}\ \frac{(n_i-1)S_i^2}{\sigma^2/\rho_i} \sim \chi^2_{n_i-1} = gamma((n_i-1)/2, 1/2)),$$

or

$$S_i^2 \sim gamma((n_i-1)/s, \frac{(n_i-1)\rho_i}{2\sigma^2}).$$

Assume $\rho_i \sim \Gamma(\alpha/2, \alpha/2)$ Results (integrating ρ_i):

1. $\frac{\bar{x_i}-\mu_i}{\sigma/\sqrt{n_i}} \sim t_\alpha$.

2. $\frac{(n_i-1)S_i^2}{\alpha\sigma^2} \sim Beta2((n_i-1)/2, \alpha/2)$. This result is equivalent to:

$$S_i^2 \sim SBeta2((n_i-1)/2, \alpha/2, \alpha\sigma^2/(n_i-1)).$$

6.2 A different marginal density

Suppose that Cauchy or more generally Student-t distributions are assumed for locations. We already showed that if the scale was assumed to be SBeta2, then the marginal for the location was a Horseshoe prior. Now we ask: What if instead of the scale, the square of the scale is assumed to ne SBeta2. That is:

$$\pi(\theta|\mu,\tau,b) = \frac{1}{\pi\sqrt{b}\tau} \cdot [1 + \frac{(\theta-\mu)^2}{b\tau^2}]^{-1},$$

and

$$\tau^2 \sim SBeta2(\tau^2|1,1,1/b).$$

What would be the marginal for θ? It is NOT a Horseshoe prior. Instead:

$$\pi(\theta) = \frac{1}{2\sqrt{b} \cdot (1 + \frac{|\theta-\mu|}{\sqrt{b}})^2}$$

This is an interesting marginal on itself, close to a Cauchy, and it does not have a pole at zero, so it is not a Horseshoe prior.
In Fúquene, Pérez and Pericchi (2011) this prior is studied and applied in detail. In fact a general result for the marginal of the location, for any p and q is obtained in terms of the Hypergeometric Function.
Result: Let $\theta \sim$ Student-t(μ,τ,υ) where υ are the degrees of freedom, μ the location and τ the scale of the Student-t density:

$$\pi(\theta|\tau^2) = \frac{k_1}{\tau}\left(1 + \frac{1}{\upsilon}\left(\frac{\theta-\mu}{\tau}\right)^2\right)^{-(\upsilon+1)/2}, \quad \upsilon > 0, -\infty < \mu < \infty, -\infty < \theta < \infty, \qquad (6)$$

where $k_1 = \dfrac{\Gamma((\upsilon+1)/2)}{\Gamma(\upsilon/2)\sqrt{\upsilon\pi}}$. Therefore

$$\pi(\theta) = \begin{cases} k\beta^q\nu/(\theta-\mu)^{q+1/2}2F1(p+q, q+1/2, (\upsilon+1)/2+p+q, 1-\beta\nu/(\theta-\mu)^2) & \text{if } \theta \neq \mu, \\ k_1\text{Be}(p-1/2, q+1/2)/(\beta^{1/2}\text{Be}(p,q)) & \text{if } \theta = \mu. \end{cases}$$

with $k = k_1\text{Be}(q+1/2, p+\upsilon/2)/\text{Be}(p,q)$. Where $Be(a,b)$ denotes the beta function and $2F1(a,b,c,z)$ denotes the hypergeometric function (see 15.1.1 of **Abramowitz and Stegun (1970)**). The $\pi(\theta)$ prior is called here the Student-t-Beta(υ,p,q,β) and it is a new wider class of hypergeometric heavy tailed priors.

7 References

Abramowitz, M and Stegun, I (1970) *Handbook of Mathematical Functions. National Bureau of Standards*, Vol. 46, Applied Mathematical Series.
Carvalho C.M, Polson N.G and Scott J.G (2010)The horseshoe estimator for sparse signals. **BIOMETRIKA**.
de Finetti, B. (1961). The Bayesian approach to the rejection of outliers. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, pp. 199-210. Univ. California Press: Berkeley, California.
Fúquene J.A., Pérez M.E. and Pericchi L.R. (2011) Modelling outliers and structural breaks in dynamic linear models with a novel use of a heavy tailed prior for the variances: An alternative to the Inverted Gamma. (Submitted).
Gelman, A. (2006)Prior distributions for variance parameters in hierarchical models(Comment on Article by Browne and Draper)*Bayesian Analysis*,Vol. 1, 3, pp. 515-534.
Pericchi, L.R. and Pérez, M.E. (2010) Limiting the shrinkage for the exceptional by Objective Robust Bayesian Analysis. Technical Report. Center for Biostatistics and Bioinformatics, UPR, Rio Piedras.