Implementation of a Robust Bayesian Method

J. Portela

Dpto de Estadística e I. Operativa III. Escuela de Estadística.

Universidad Complutense de Madrid.

M. A. Gómez-Villegas

Dpto de Estadística e I. Operativa I. Facultad de Matemáticas.

Universidad Complutense de Madrid.

Abstract

In this work we study robustness in Bayesian models through a generalization of the Normal distribution. We show new appropriate techniques in order to deal with this distribution in Bayesian inference. Then we propose two approaches to decide, in some applications, if we should replace the usual Normal model by this generalization. First, we pose this dilemma as a model rejection problem, using diagnostic measures. In the second approach we evaluate model's predictive efficiency. We illustrate those perspectives with a simulation study, a non linear model and a longitudinal data model.

Keywords: Bayesian Inference, Bayesian robustness, Exponential Power distribution, Markov Chain Monte Carlo.

1 Introduction

We will use the form of the Exponential Power distribution , $EP(\theta,\sigma,\beta)$,

$$p(y \mid \theta,\sigma,\beta) = w(\beta)\sigma^{-1}\exp\left[-c(\beta)\left|\frac{y-\theta}{\sigma}\right|^{2/(1+\beta)}\right]$$

where $-\infty < y < +\infty,\ -\infty < \theta < +\infty,\ -1 < \beta \leq 1,\ \sigma > 0,$

$c(\beta)$ is a positive and bounded function , and

$$w(\beta) = \frac{c(\beta)^{\frac{1+\beta}{2}}}{2\Gamma[\frac{1}{2}(3+\beta)]}$$

It can be shown that the mean of the distribution is θ and its variance,

$$V(Y) = \frac{\Gamma[\frac{3}{2}(1+\beta)]}{\Gamma[\frac{1}{2}(1+\beta)]c(\beta)^{1+\beta}}\sigma^2$$

The Exponential Power distribution (Box, Tiao,1973) is commonly used in Bayesian robustness studies. It is a family of symmetric distributions that generalise Normal distribution, having more or less kurtosis than this distribution as parameter β varies. In previous Bayesian works, it is usual to take β as a fixed value, and observe, in a exploratory way, the consequences of changing β values. In this work we will always consider β as a random variable from the beginning.

Usually $c(\beta)$ is fixed in a way that leads to $V(Y) = \sigma^2$ for every β. However, an interesting reparametrization to simplify the Bayesian treatment of this distribution is to take $c(\beta) = \frac{1}{2}$. In this particular case, when $\beta = 0$, $V(Y) = \sigma^2$ and $PE(\theta,\sigma,\beta)$ is a $N(\theta,\sigma)$ distribution.

Posing $c(\beta) =$ constant make computations simpler, so we will use that reparametrization in this work. A detailed study of the Exponential Power distribution can be seen in Marin (1998). A multivariate generalization in the form $p(\mathbf{y} \mid \boldsymbol{\theta},\Sigma,\beta)$ is shown in Gómez, Gómez-Villegas and Marín (1998).

In the first part of this paper we develop tools to work with this distribution in a Bayesian frame. Then we study how to determine, through the use of discrepancy measures, whether we should continue with the usual Normal model or use the Exponential Power model. In the last part, we propose an alternative way to validate the use of this distribution, and we show its usefulness in applications like a non linear model and a longitudinal data model.

2 Monte Carlo treatment of EP distribution

Following the Bayesian paradigm, we consider θ,σ and β as random variables, with prior distributions $p(\theta)$, $p(\sigma)$ and $p(\beta)$, considered independents through this work. We will take $p(\theta,\sigma,\beta) \propto p(\theta)p(\beta)\frac{1}{\sigma}$, where $p(\beta)$ is the Uniform distribution over the $(-1,1)$ interval, and $p(\theta)$ any continuous and bounded distribution over $(-\infty,+\infty)$.

Having a sample $y_1,...,y_n$ from an Exponential Power distribution, the posterior distribution of parameters will take the form

$$p(\theta,\sigma,\beta \mid \mathbf{y}) \propto p(\theta)p(\beta)[w(\beta)]^n \sigma^{-(n+1)} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left|\frac{y_i-\theta}{\sigma}\right|^{2/(1+\beta)}\right] \tag{1}$$

We will need to use numerical methods to deal with this distribution, if we want some information about the conditional distributions $p(\theta \mid \mathbf{y})$,$p(\sigma \mid \mathbf{y})$ and $p(\beta \mid \mathbf{y})$. We will show a way to get estimations of these distributions via Monte Carlo, and next we propose simulation methods to get samples from the full posterior distribution.

2.1 Posterior densities estimations

Posing $p(\sigma) \propto 1/\sigma$, we can obtain an analytical expression proportional to $p(\theta,\beta \mid \mathbf{y})$ through direct integration in σ. It can be shown that this expression is integrable with the stated $p(\theta)$ and $p(\beta)$, hence it is a valid posterior density for θ and β , as is assumed in (Box, Tiao, (1973)). A proportional approximation of parameter posterior densities through Monte Carlo integration is then obtained. Taking samples of size m from prior distributions $p(\beta)$ and $p(\theta)$ gives:

$\hat{p}(\theta \mid \mathbf{y}) \propto \frac{1}{m}\sum_{i=1}^{m} p(\theta,\beta_i \mid \mathbf{y})$

$\hat{p}(\beta \mid \mathbf{y}) \propto \frac{1}{m}\sum_{i=1}^{m} p(\theta_i,\beta \mid \mathbf{y})$

$\hat{p}(\sigma \mid \mathbf{y}) \propto \frac{1}{m^2}\sum_{i=1}^{m^2} p(\sigma,\theta_i,\beta_i \mid \mathbf{y})$

These estimations allow for graphical representations of the densities and the estimation of posterior mode . If we want to calculate concrete probabilities we should use numerical integration to normalize the density. The precision of this Monte Carlo approach depends on the particular application. In the cases we have worked on, stabilization of Monte Carlo variance estimator is observed since $m = 300$.

2.2 Obtaining samples from posterior distributions

The aim is to get samples from the full posterior $p(\theta,\sigma,\beta \mid \mathbf{y})$. Assuming we know how to generate random samples from the Exponential Power distribution as well as from other known densities (Devroye (1984)).

2.2.1 Direct use of the Gibbs Sampler

In order to use the Gibbs sampler it is necessary to develop methods to generate samples from each posterior conditional density.

- Generation of samples from $p(\sigma \mid \theta,\beta,\mathbf{y})$

In this case, taking in (1) the transformation $z = \sigma^{2/(1+\beta)}$ then $z{\propto}IG(a,b)$, where $a = \dfrac{n(1+\beta)}{2}$ and

$b = \frac{1}{2}\sum_{i=1}^{n} |y_i - \boldsymbol{\theta}|^{2/(1+\beta)}$ so a direct method to get samples σ from $p(\sigma \mid \theta, \beta, \mathbf{y})$ will consist in obtaining a sample value x from a Gamma $\Gamma(a, b)$ and transform it through $\sigma = x^{-(1+\beta)/2}$.

- Generation of samples from $p(\theta \mid \sigma, \beta, \mathbf{y})$.

It can be shown that, for $k \geq 1$, $\sum_{i=1}^{n} |\theta - y_i|^k \geq n\,|\theta - \overline{y}|^k$ then

$$p(\theta)\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left|\frac{y_i - \boldsymbol{\theta}}{\sigma}\right|^{2/(1+\beta)}\right] \leq p(\theta)\exp\left[-\frac{n}{2}\left|\frac{\overline{y} - \boldsymbol{\theta}}{\sigma}\right|^{2/(1+\beta)}\right]$$

If $p(\theta)$ is Uniform or noninformative, the superior boundary has the form of an Exponential Power distribution, thus a rejection method to obtain samples from θ can be used. Else, if $p(\theta)$ is bounded, we can also use the same rejection method. In other cases, ad-hoc methods of generation could be developped.

- Generation of samples from $p(\beta \mid \theta, \sigma, \mathbf{y})$

In this case

$$p(\beta \mid \theta, \sigma, \mathbf{y}) \propto 2^{-\frac{n(3+\beta)}{2}}\Gamma^{-n}[\frac{1}{2}(3+\beta)]\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left|\frac{y_i - \boldsymbol{\theta}}{\sigma}\right|^{2/(1+\beta)}\right]$$

This expression can be bounded by

$$2^{-3n/2}\Gamma^{-n}(\frac{3}{2})\exp(-(\tfrac{n}{2}\log(2))\beta)$$

so we can use again a rejection method to generate samples from $p(\beta \mid \theta, \sigma, \mathbf{y})$, taking samples from an exponential distribution truncated in $(-1, 1)$, and rejecting samples using the bound proposed.

In our experience with this approach, for moderate n Gibbs sampler seems to converge before 60 iterations, while in some complicated applications it takes about 300. When posterior mode of β approaches the extreme value -1, the rejection method proposed may take too long to converge. When in the initial runs the rejection algorithm does not accept at least 20% of the points generated, we replace it by the SIR method (Rubin, (1987)), and in some extreme cases by trapezoidal density estimation followed by rejection sampling.

2.2.2 Mixture representations

The Exponential Power distribution can be posed as a continuous mixture of Gamma and Uniform. (Walker and Gutiérrez-Peña,(1999)). This mixture representation of an $EP(Y \mid \theta, \sigma, \beta)$ distribution takes the form

$(Y \mid \theta, \sigma, \beta, U = u) \propto Uniform(\theta - \sigma(2u)^{\frac{\beta+1}{2}}, \theta + \sigma(2u)^{\frac{\beta+1}{2}})$ with $(U \mid \beta) \propto \Gamma(\frac{1}{2}(3+\beta), 1)$

where U is the mixing parameter.

This representation allows to approach the obtention of samples from the posterior $p(\theta, \beta, \sigma \mid \mathbf{y})$ from an Exponential Power model, through the Gibbs sampler. We will use the prior distributions $p(\theta)$ any density in $(-\infty, +\infty)$, $p(\beta)$ $\propto$ Uniform$(-1, 1)$, and $p(\sigma)$ $\propto$ $IG(a, b)$, an Inverse-Gamma distribution where a is little enough so that the prior distribution $p(\sigma)$ is approximately noninformative. We suppose all these prior distributions independent.

The likelihood of a sample $\mathbf{y}$, given the vector of mixing parameters $\mathbf{u} = (u_1, u_2, ..., u_n)$, is

$p(\mathbf{y} \mid \theta, \sigma, \beta, \mathbf{u}) = \sigma^{-n} 2^{-\frac{n(\beta+3)}{2}} \prod_{i=1}^{n} u_i^{-1}$, with $y_i \in [\{\theta - \sigma(2u_i)^{\frac{\beta+1}{2}}\}, \{\theta + \sigma(2u_i)^{\frac{\beta+1}{2}}\}]$.

In this framework, we have

$$p(\theta \mid \sigma, \beta, \mathbf{u}, \mathbf{y}) \propto p(\mathbf{y} \mid \theta, \sigma, \beta, \mathbf{u}) p(\boldsymbol{\theta}) \propto p(\theta), \theta \in [\max\{y_i - \sigma(2u_i)^{\frac{\beta+1}{2}}\}, \min\{y_i + \sigma(2u_i)^{\frac{\beta+1}{2}}\}].$$

We also get, for the remainder parameters,

$$\begin{aligned} \sigma \mid \theta, \beta, \mathbf{u}, \mathbf{y} \quad & \equiv IG(a+n, b), \sigma > \max\{\left(\frac{1}{2u_i}\right)^{\frac{\beta+1}{2}} |y_i - \theta|\} \\ \beta \mid \theta, \sigma, \mathbf{u}, \mathbf{y} \quad & \propto 2^{-\frac{n(3+\beta)}{2}} \Gamma^{-n}[\tfrac{1}{2}(3+\beta)], \beta \text{ such that } (2u_i)^{\frac{\beta+1}{2}} > \left(\frac{y_i - \theta}{\sigma}\right)^2, \forall i \\ u_i \mid \theta, \sigma, \beta, y_i \quad & \propto Exp(1), 2u_i > \left(\frac{(y_i - \theta)}{\sigma}\right)^{\frac{2}{1+\beta}} \end{aligned}$$

Successive implementations of the Gibbs sampler , given the sample $\mathbf{y}$, result in samples from $(\theta, \sigma, \beta, \mathbf{u} \mid \mathbf{y})$.

In order to get samples from $(\beta \mid \theta, \sigma, \mathbf{u}, \mathbf{y})$ truncated in its bounds it is easy first to calculate the boundary region for β, and then use a variation of the rejection method presented in the previous section. Using Gibbs sampler also requires generation of samples from known distributions truncated in some regions• there already exist techniques to deal with them (Devroye, (1984)). A simplified version of this representation, taking β as a fixed value and changing it in a exploratory way, can be seen in (Choy, (1999)).

3 Applications to Normality checking

We are approaching the problem of evaluating data departure from Normality, through the use of the Exponential Power β parameter. This question may be approached as an hypothesis testing problem with $H_0 : \beta = 0$ facing $H_1 : \beta \neq 0$, as we know that for $\beta = 0$ the Exponential Power distribution agrees with the Normal distribution.

This problem can be seen as a Model Rejection problem. An approach introduced by Bernardo and Smith (1994) consists in choosing a discrepancy measure $\delta(\beta, \theta, \sigma)$,which measures the distance between likelihood functions $p(\mathbf{y} \mid \beta, \theta, \sigma)$ and $p(\mathbf{y} \mid 0, \theta, \sigma)$. Then the posterior expectation of this measure, $E_{\beta,\theta,\sigma|\mathbf{y}}[\delta(\beta, \theta, \sigma)]$,is computed. $E_{\beta,\theta,\sigma|\mathbf{y}}[\delta(\beta, \theta, \sigma)]$ can be seen as the difference between both models posterior utilities.

We will apply this idea using a range of discrepancy measures and investigate their behaviour through a simulation study. In order to give the same prior probability for all the models we pose, as has been done in previous sections, $p(\beta) \equiv U(-1, 1)$.

3.1 Kullback-Leibler distance

If we want to compare $p(\mathbf{y} \mid \beta, \theta, \sigma)$ y $p(\mathbf{y} \mid 0, \theta, \sigma)$, we can use the well-known Kullback-Leibler distance:

$$\delta_{KL}(\beta, \theta, \sigma) = \int p(\mathbf{y} \mid \beta, \theta, \sigma) \log \frac{p(\mathbf{y} \mid \beta, \theta, \sigma)}{p(\mathbf{y} \mid 0, \theta, \sigma)} d\mathbf{y}$$

In the particular case of the Exponential Power distribution, we can get its analytical form, which does not depend on (θ, σ):

$$\delta_{KL}(\beta, \theta, \sigma) = \delta_{KL}(\beta) = \frac{(1+\beta)}{2\Gamma(\frac{1}{2}(3+\beta))} \left[\Gamma(\tfrac{1+\beta}{2}) \log \frac{\Gamma(\frac{3}{2})}{2^{\frac{\beta}{2}} \Gamma(\frac{1}{2}(3+\beta))} - \Gamma(\tfrac{1}{2}(3+\beta)) + 2^{\beta}\Gamma(\tfrac{3}{2}(1+\beta)) \right]$$

In order to compute the posterior expectation of $\delta_{KL}(\beta, \theta, \sigma)$, analytical integration does no exist in a simple form, but we can obtain a Monte Carlo estimator based on samples generated from $p(\beta, \theta, \sigma \mid \mathbf{y})$, through:

$$E_{\beta,\theta,\sigma|\mathbf{y}}[\delta_{KL}(\beta, \theta, \sigma)] = E_{\beta,\theta,\sigma|\mathbf{y}}[\delta_{KL}(\beta)] \simeq \frac{1}{m} \sum_{j=1}^{m} \delta_{KL}(\beta_j)$$

where the vector $(\beta_j, \theta_j, \sigma_j)$ is generated from $p(\beta, \theta, \sigma \mid \mathbf{y})$ by means of the methods introduced in the previous section.

3.2 A discrepancy measure based on HPD regions

For the point null hypotheses testing $H_0 : \beta = 0$ facing $H_1 : \beta \neq 0$, a testing procedure can be developped constructing the Highest Posterior Density region (Berger, 1985) $R = \{\beta : p(\beta \mid \mathbf{y}) > k(\alpha)\}$, where $k(\alpha)$ is the largest constant such that $P_{\beta|\mathbf{y}}(\beta \in R) \geq 1 - \alpha$, α fixed sufficiently small. Then we accept H_0 if the point $\beta = 0$ falls into the HPD region R, and reject H_0 otherwise.

An evidence measure against data Normality can be computed finding the posterior region

$$C = \{\beta \mid p(\beta \mid \mathbf{y}) \geq p(0 \mid \mathbf{y})\}$$

and then computing the posterior probability $p(C \mid \mathbf{y})$. As $p(C \mid \mathbf{y})$ decreases, the evidence about the hypotheses $\beta = 0$ arising from the data increases.

The Highest Probability Density region (HPD) $1 - \alpha$ test, is equivalent to reject H_0 if $p(C \mid \mathbf{y}) > 1 - \alpha$.

To • x the notation for a discrepancy measure based on this probability, we write $p(C \mid \mathbf{y})$ as

$$p(C \mid \mathbf{y}) = \int \mathbf{I}\{\beta \mid p(\beta \mid \mathbf{y}) \geq p(0 \mid \mathbf{y})\} p(\beta, \theta, \sigma \mid \mathbf{y}) d\beta d\theta d\sigma$$

Then, if we define

$$\delta_{HPD}(\beta, \theta, \sigma) = \delta_{HPD}(\beta) = \begin{cases} 1 \text{ if } \dfrac{\int p(\mathbf{y} \mid \beta, \theta, \sigma) p(\theta, \sigma) d\theta d\sigma}{\int p(\mathbf{y} \mid 0, \theta, \sigma) p(\theta, \sigma) d\theta d\sigma} \geq 1 \\ 0 \text{ if } \dfrac{\int p(\mathbf{y} \mid \beta, \theta, \sigma) p(\theta, \sigma) d\theta d\sigma}{\int p(\mathbf{y} \mid 0, \theta, \sigma) p(\theta, \sigma) d\theta d\sigma} < 1 \end{cases}$$

we can pose

$$E_{\beta,\theta,\sigma|\mathbf{y}}[\delta_{HPD}(\beta, \theta, \sigma)] = E_{\beta,\theta,\sigma|\mathbf{y}}[\delta_{HPD}(\beta)] = p(C \mid \mathbf{y})$$

An interesting property of this method, apart from having a direct interpretation in terms of HPD regions, is that the expectation $E_{\beta,\theta,\sigma|\mathbf{y}}[\delta_{HPD}(\beta)]$ is already calibrated: since it is a posterior probability, it takes values in (0,1).

In order to estimate $E_{\beta,\theta,\sigma|\mathbf{y}}[\delta_{HPD}(\beta)]$ we will use the Monte Carlo approach of the posterior distribution. We compute

$p(0 \mid \mathbf{y})$; then we search, by numerical methods, all points β' apart from $\beta = 0$, such that $p(0 \mid \mathbf{y}) = p(\beta' \mid \mathbf{y})$, and compute the according HPD interval probability .If no point β' exists , $P(C \mid \mathbf{y}) = 1$ or 0 depending on $p(\beta \mid \mathbf{y}) > p(0 \mid \mathbf{y})$ for all β or conversely.

3.3 A discrepancy measure based on predictive distribution

(Walker and Gutiérrez-Peña (1999)) poses as a measure of evidence about the model $p(\mathbf{y} \mid \beta, \theta, \sigma)$ facing $p(\mathbf{y} \mid 0, \theta, \sigma)$ the difference of posterior expected utilities, based on the predictive distribution . A Monte Carlo estimation of its posterior expectation will be

$$E_{\beta,\theta,\sigma|\mathbf{y}}[\delta_P(\beta, \theta, \sigma)] \simeq \frac{1}{n} \sum_{i=1}^{n} \log \frac{\sum_{j=1}^{m} p(y_i \mid \beta_j, \theta_j, \sigma_j)}{\sum_{j=1}^{m} p(y_i \mid 0, \theta_j^*, \sigma_j^*)}$$

Where vectors (θ_j^*, σ_j^*) are generated from $p(0, \theta, \sigma \mid \mathbf{y})$,and vectors $(\beta_j, \theta_j, \sigma_j)$ are generated from $p(\beta, \theta, \sigma \mid \mathbf{y})$.

It would be useful to define a calibration for this measure of evidence about (β, θ, σ), as it is the case for the Kullback-Leibler measure.

3.4 Simulation study

A range of simulations has been carried out in order to compare the introduced evidence measures sensibility to data kurtosis deviance from normality.

Size $n = 100$ samples from an Exponential Power distribution with parameter β have been generated. We have computed the introduced discrepancy measures to evaluate posterior evidence against data Normality. This process has been replicated 30 times with 30 size $n = 100$ samples generated from Exponential Power distribution for each one of the following values of $\beta = -0.75$, -0.5, $-0.25, 0, 0.25, 0.5, 0.75$.

We obtain, for each β fixed in the data original distribution, a sample from each normality testing measure. We show in figure 1 Box plots for these samples, ordered by β, and having all measures resized to the (0,1) interval.

In order to have another comparison reference we also include Kolmogorov-Smirnov-Lilliefors test p-value, obtained over the same samples. We resize this p-value to ”$1 - p - value$” because our discrepancy measures take higher values as data moves away from Normality.

(figure 1 about here)

Discrepancy measures based on the frequentist Kolmogorov-Smirnov-Lilliefors test, posterior measure HPD and predictive distribution based measure give correct results, getting higher values as long as generated samples go further from Normality ($\beta = 0$). This happens whether data has a lower kurtosis than Normal distribution ($\beta \to -1$), or a higher kurtosis than Normal distribution ($\beta \to 1$). HPD seems to be more sensitive, and given it is already calibrated in (0,1), it will be used from now on as the reference measure.

In the case of Kullback-Leibler measure, we can see graphically the consequences of its asymmetry: this measure discriminate better samples with a higher kurtosis than Normal distribution, than samples with lower kurtosis than Normal distribution.

4 Application to robustness and model comparison problems

In this section we shall apply the Exponential Power distribution to some Bayesian models, replacing the usually Normal errors with Exponential Power distributed errors. In order to determine whether this choice is appropiate, we use the HPD evidence measure as well as a of model's predictive effectivity measure.

4.1 Model predictive evaluation

Gelfand, Dey and Chang, (1992) propose the choice between two non linear models through the use of the predictive distribution, in a cross validation perspective. Given $\mathbf{y}_{(r)}$, the vector $\mathbf{y}$ without the rth observation, we compute $d_r = \hat{E}(h(Y_r)|\ \mathbf{y}_{(r)})$, where $h(Y_r)$ is a diagnostic function which measures model fitting for each observation. An intuitive possibility is to take $h(y_r) = y_r - Y_r$,so that $d_r = y_r - \hat{E}(Y_r|\ \mathbf{y}_{(r)})$. The sum $\sum(d_r)^2$ may be seen as a measure of model's predictive effectivity.

If response variable y and regression variables x_i are related by some function named g :

$$y_i = g(\mathbf{x}_i|\ \boldsymbol{\theta})+\varepsilon_i$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$, y $\mathbf{x}_i = (x_1, .., x_k)_i$,and ε is a vector of n independent and $PE(0, \sigma, \beta)$ distributed random errors ,posterior parameter distribution takes the form

$$p(\boldsymbol{\theta}, \boldsymbol{\sigma}, \beta \mid \mathbf{y}) \propto [w(\beta)]^n \sigma^{-(n+1)} \exp\left[-c(\beta) \sum_{i=1}^{n} \left| \frac{y_i - g(\mathbf{x}_i|\ \boldsymbol{\theta})}{\sigma} \right|^{2/(1+\beta)} \right]$$

We can estimate posterior conditional densities for parameters $\boldsymbol{\theta}, \boldsymbol{\sigma}, \beta$ for every regular function $g(\mathbf{x}_i|\ \boldsymbol{\theta})$. However, specific functions $g(\mathbf{x}_i|\ \boldsymbol{\theta})$ are needed if we want to simulate samples from posterior distributions using direct Gibbs sampler or the EP mixture representation, because development of these methods depends on.the $g(\mathbf{x}_i|\ \boldsymbol{\theta})$ function used in each particular case. Both procedures will be applied in the next sections. We use density Monte Carlo estimations in section 4.2 and Gibbs sampler in section 4.3.

Computation of $d_r = y_r - \widehat{E}(Y_r|\ \mathbf{y}_{(r)})$ can be done through MonteCarlo estimations. We also take profit of the integration over σ property in order to make computations simpler.

4.2 Applications to non linear models

Gelfand, Dey and Chang, (1992) compare in their work two models, the logistic one : $y = \theta_0(1 + \theta_1\theta_2^x)^{-1} + \varepsilon$, and the Gompertz model: $y = \theta_0 e^{-\theta_1\theta_2^x} + \varepsilon$ in an application, where Y represents onion bulbs weight measured over the increasing time X, in a series of $n = 15$ observations. In that paper, ε were distributed as Normal. In this work we suppose ε are Exponential Power distributed.

In both logistic and Gompertz models, we use the reparametrizations: $\theta_1' = \log(\theta_1)$ y $\theta_2' = \log(\theta_2/(1 - \theta_2))$. We take all these prior distributions to be Uniform over large intervals, so they are approximately noninformative. We have then

$p(\theta_0, \theta_1', \theta_2', \sigma, \beta) \propto 1/\sigma$.

4.2.1 Logistic model

Posterior distributions for the parameters related to the ε error term, σ and β, are shown in figure2.

(Figure 2 about here)

Posterior distribución $p(\beta \mid \mathbf{y})$ has its mode at $\beta = 1$, far from $\beta = 0$ value attached to Normality, and the introduced measure HPD=0.87, so it seems appropiate to use the Exponential Power model instead of the usual Normal model . We have also considered the Bayesian analysis with $\beta = 0$ fixed, that is, taking the error term ε as Normally distributed. Table 1 displays parameters posterior means and modes for both models EP and Normal:

(Table 1 about here)

4.2.2 Gompertz model

In this model it looks also appropiate to use the Exponential Power model, since posterior distribution $p(\beta| \mathbf{y})$ mode is far from the Normality value $\beta = 0$,as we can see in Table 2. Measure HPD=0.91 in this case.

(Table 2 about here)

4.2.3 Comparison of predictive fitting for both models

We can assess predictive efectivity for both models through the estimator $\hat{E}(Y_r| \mathbf{y}_{(r)})$, obtaining results displayed in Table 3:

(Table 3 about here)

We see that Exponential Power model seems to fit slightly better for both Logistic and Gompertz models , something expected after the study of β posterior distribution in both cases. Logistic model fits better than Gompertz model in both Exponential Power errors model and Normal errors model .

4.3 Application to a longitudinal data model

Han and Carlin (2001) compare two models in an AIDS longitudinal data clinical trial. Data from this experiment has also been studied in Carlin and Louis (2000) . CD4 lymphocite counts Y_i were measured for each subject at the 0, 2, 6, 12 and 18 months visits. Two different treatments and two different baseline conditions (AIDS diagnostic or not) were considered.

We desire to compare models M_1 and M_2, where $M_1 : Y_i = X_i\mathbf{a} + W_i\mathbf{b}_i + \varepsilon_i$ and $M_2 : Y_i = P_i\mathbf{c} + Q_i\mathbf{d}_i + \nu_i$, $i = 1,..,n$, where (X_i, W_i) in model M_1 and (P_i,Q_i) in model M_2 are observation matrices that represent, respectively, treatments and baseline conditions, and timepoints (see Han and Carlin (2001) for details). In model M_1, $\mathbf{a}$ is a 9 terms fixed effects vector and $\mathbf{b}_i$ are 3 terms random effects vectors. In model M_2, $\mathbf{c}$ is a 6 terms fixed effects vector and $\mathbf{d}_i$ are 2 terms random effects vectors. Model M_1 supposes a change in the slope of Y two months after baseline timepoint, while model M_2 assumes the same slope from baseline timepoint, along all the study.

Following Zeger and Karim (1992) work, we choose an Uniform over a large region, (an approximately noninformative prior) for $\mathbf{a}$. For $\mathbf{b}_i$ vectors we use a Multivariate Normal prior $N(0, \mathbf{V})$, where $\mathbf{V}^{-1}$ is Wishart $W((\rho R)^{-1}, \rho)$ and R and ρ are the same values suggested by Carlin and Louis (2000) for this data.

As in the previous application, we use an Exponential Power distribution for the random errors, $\varepsilon_{ij} \propto PE(0,\sigma^2 I, \beta)$ independent and identically distributed. We have $y_{ij} = \mathbf{x}_{ij}\mathbf{a} + \mathbf{w}_{ij}\mathbf{b}_i + \varepsilon_{ij}$ for each patient i and time point j, where $i = 1, ..., n$ and $j = 1, ..., s_i$, with s_i being the number of observations taken for the i^{th} patient.Usually $s_i = 5$, and observations are taken at time points t =0,2,,6,12,18 months from the beginning of the treatment, but there are many missing observations at last time points. In order to compare both models M_1 and M_2, we compute the predictive mean Monte Carlo estimator

$$\hat{E}(Y_{ij}|\ \mathbf{y}_{(ij)}) = \frac{1}{m}\sum_{k=1}^{m}(\mathbf{x}_{ij}\mathbf{a}^{(k)}+\mathbf{w}_{ij}\mathbf{b}_i^{(k)})$$

For each observation, we compute $d_{ij} = y_{ij} - \hat{E}(Y_{ij}|\ \mathbf{y}_{(ij)})$ .Then we compute $\sum d_{ij}^2$ as a measure of predictive fitting.

We present the Bayesian development for model M_1, taking into account there are no important variations with respect to the M_2 analysis, where we change matrices X_i and W_i for P_i and Q_i respectively, and vectors $\mathbf{a}$ and $\mathbf{b}_i$ for $\mathbf{c}$ and $\mathbf{d}_i$ respectively.

4.3.1 Posterior distribution. Sample generation

We want to generate samples from $p(\sigma, \beta, \mathbf{a}, \mathbf{b} \mid \mathbf{y}_{\mathbf{(ij)}})$. We use the Gibbs sampler, generating samples for each conditional distribution.

- Samples from $p(\sigma|\beta, \mathbf{a}, \mathbf{b}, \mathbf{y})$

Taking $a = \dfrac{\sum_{i=1}^{n} s_i(1+\beta)}{2}$ and $b = \dfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{s_i}|y_{ij}{}_{-}(\mathbf{x}_{ij}\mathbf{a} + \mathbf{w}_{ij}\mathbf{b}_i)|^{2/(1+\beta)}$, we generate a value q wih gamma $\Gamma(a, b)$ distribution and we compute $\sigma = q^{-(1+\beta)/2}$.

It can be shown that σ is distributed as $p(\sigma|\beta, \mathbf{a}, \mathbf{b}, \mathbf{y})$.

- Samples from $p(\beta|\ \boldsymbol{\sigma}, \mathbf{a}, \mathbf{b}, \mathbf{y})$

In this case, posterior β distribution is bounded by the expression

$2^{-3\Sigma s_i/2}[\Gamma(\dfrac{3}{2})]^{-\Sigma s_i}\exp(-(\frac{\Sigma s_i}{2}\log(2))\beta)$ so we use a rejection method as the one introduced in the first section..

Samples from $p(\mathbf{a} \mid \boldsymbol{\sigma},\beta,\mathbf{b},\mathbf{y})$

In this case, $\mathbf{a}=(a_1,...,a_9)$, so we use the Gibbs sampler taking samples from each conditional $p(a_k| \ \boldsymbol{\sigma},\beta,\mathbf{a}_{(k)},\mathbf{b},\mathbf{y})$, noting by $\mathbf{a}_{(k)}$ the $(a_1,...,a_9)$ vector without its kth term.

To sample values from $p(a_k| \ \boldsymbol{\sigma},\beta,\mathbf{a}_{(k)},\mathbf{b},\mathbf{y})$, noting

$$z_{ij}^k=\left[\frac{y_{ij}-\mathbf{w}_{ij}\mathbf{b}_i-\sum\limits_{l\neq k}x_{ij}^l a_l}{x_{ij}^k}\right]$$

we have

$$p(a_k| \ \boldsymbol{\sigma},\beta,\mathbf{a}_{(k)},\mathbf{b}_\mathbf{i},\mathbf{y}) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{s_i}\left|\frac{(a_k-z_{ij}^k)}{\sigma/x_{ij}^k}\right|^{2/(1+\beta)}\right]$$

Defining

$\bar{z}^k=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{s_i}z_{ij}^k$ and $\sigma_{\max}=\dfrac{\sigma}{Min\{x_{ij}^k\}}$,we can see that (2) is bounded by

$$\exp\left[-\frac{\Sigma s_i}{2}\left|\frac{a_k-\bar{z}^k}{\sigma_{\max}}\right|^{2/(1+\beta)}\right]$$

Applying then the rejection method introduced in Section 1 to get samples a_k. We repeat this computations over the $a_1,...,a_9$ a sufficient number of iterations, obtaining finally a sample from $p(\mathbf{a} \mid \boldsymbol{\sigma},\beta,\mathbf{b},\mathbf{y}_{(\mathbf{ij})})$.

- Samples from $p(\mathbf{b}_i|\sigma,\beta,\mathbf{a},\mathbf{V},\mathbf{y})$

We have $p(\mathbf{b}_i|\sigma,\beta,\mathbf{a},\mathbf{V},\mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\sigma},\beta,\mathbf{a},\mathbf{b}_\mathbf{i})p(\mathbf{b}_\mathbf{i} \mid \mathbf{V})$.

We need , for every $i=1,...,n$, a sample from $\mathbf{b}_\mathbf{i}=(b_{i1},b_{i2},b_{i3})$. We implement a method similar to the one used to get samples from $\mathbf{a}$, since $p(\mathbf{b}_\mathbf{i} \mid \mathbf{V})$ is bounded. In this case a sufficient number of Gibbs sampler iterations with each $\mathbf{b}_\mathbf{i}=(b_{i1},b_{i2},b_{i3})$ for each observation is necessary, taking a lot of computing time.

- Samples from $p(\mathbf{V}^{-1}| \ \mathbf{b})$

We know this posterior distribution is also Wishart, so we use methods that already exist to sample from it (see Carlin and Louis (2000)).

4.3.2 Results

We display the results for model $M_1: Y_i=X_i\mathbf{a}+W_i\mathbf{b}_i+\varepsilon_i$. We have realized 500 iterations of Gibbs sampler, rejecting the first 100. In figure 3 we show posterior distribution histograms for $a_1, b_{8,1}, \sigma$ and β. β posterior distribution leads to a HPD=0.93 value, rejecting error normality.

(figure 3 about here)

In Table 4 we present $a_1, ..., a_9$,$b_{8,1}$, σ and β posterior modes . We also append Carlin and Louis (2000) results obtained with Normal errors model ($\beta = 0$).

(Table 4 about here)

Results are generally similar for fixed effects parameters $a_1, ..., a_9$ and for the random effects parameters $b_{8,1}$, in spite of the difference between Exponential Power and Normal models, and having different computational approaches. As it was the case for the models shown in the previous section, σ is the parameter more affected by the introduction of β as a random variable.

In order to compare models $M_1 : Y_i = X_i\mathbf{a} + W_i\mathbf{b}_i + \varepsilon_i$ and $M_2 : Y_i = P_i\mathbf{c} + Q_i\mathbf{d}_i + \nu_i$, taking $\beta = 0$ (Normal errors model) and β random (Exponential Power model), we compute $d_{ij} = y_{ij} - \hat{E}(Y_{ij}|\ \mathbf{y}_{(ij)})$ for each y_{ij} available observation. Table 5 displays predicted values for the first 6 cases and the fitting measure $\sum\sum d_{ij}^2$ for each of the four different models, based on the 1405 available observations.

(Table 5 about here)

We see that model M_2 has a predictive fit better than model M_1. These results agree with previous work (see Carlin and Louis (2000) and Han and Carlin (2001)). For each of these two models using β as a random parameter leads to higher predictive precision than using the Normal model $\beta = 0$. This difference is clearly bigger in model M_2.

5 Conclusions

The use of the Exponential Power distribution family leads to more robust models in many applications, at the expense of technical and computational complications. In the models presented centralization parameters are less affected than scale parameter σ when we introduce β as a random parameter. In some applications, the use of this family will be worth while depending on factors like data deviation from normality or model's predictive effectivity ,aspects we can evaluate using the techniques exposed in this work. From the computational point of view, many applications would need ad-hoc methods in order to work with this distribution, while some of the tools introduced in this work can serve as a basis. Sometimes , posterior distributions for θ and β can have very high kurtosis and low variability. In this few extreme cases we have corrected the exposed techniques applying SIR and trapezoidal rejection methods.

Acknowledgements

We thank the editor and referee for their very helpful comments, which have considerably improved the article.

This research was supported by grant DGESIC PB98-0797 awarded by the MEC from Spain.

6 References

Berger, J.O. (1985) Statistical Decision Theory and Bayesian Analysis. Second Edition. New York:Wiley.

Bernardo J.M.and Smith, A.F.(1993) Bayesian Theory. Wiley and Sons.

Box G.E.P. and Tiao, G. C. (1973) Bayesian Inference in Statistical Analysis. Addison-Wesley.

Carlin, B.P. and Louis, T.A. (2000) Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall.

Choy, S.T.B.(1999) Discussion-Robustifying Bayesian procedures Bayesian Statistics 6, 685-710. Oxford University. Press.

Devroye, L. (1984) Non Uniform Random Variable Generation. Springer-Verlag.

Gelfand, A.E. and Ghosh, S.K. (1998) Model Choice: a minimum posterior predictive loss approach.. Biometrika(1998) 85, 1-11.

Gelfand, A. E., Dey, D.K,.and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Bayesian Statistics 4. Oxford University. Press.

Gómez, E., Gómez-Villegas,M.A. and Marín,J.M. (1998) A multivariate generalization of the exponential family of distributions. Communications in Statistics. Theory and methods, 27, 3, 589-600.

Goutis, C, and. Robert, C.P. (1998) Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. Biometrika(1998) 85, 29-37.

Han, C.and Carlin, B.P. (2001) Markov Chain Monte Carlo methods for computing bayes factors: a comparative review. Journal of the American Statistical Association, 96, 1122-1132.

Marin, J.M. (1998) Ph Thesis. Facultad de matemáticas. Universidad Computense de Madrid.

Natarajan , R.and Kass, R. (2000) Reference Bayesian methods for generalised linear mixed models. Journal of the American Statistical Association, 95, 227-237

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. J. Wiley, New York.

Vounatsou, P. Smith, A.F.M.and Choy, S.T.B.(1998) Bayesian robustness for location and scale parameters using simulation.Imperial College Technical Report Series TR 96-21.

Walker, S.G.and Gutierrez-Peña, E. (1999). Robustifying Bayesian procedures. Bayesian Statistics 6, 685-710. Oxford University. Press.

Zeger, S. L.and Karim, M. R. (1991) Journal of the American Statistical Association, 86, 79-86.

Tables

	Exponential Power model					Normal model			
	θ_0	θ'_1	θ'_2	σ	β	θ_0	θ'_1	θ'_2	σ
Posterior mode	699.75	4.37	0.0158	11.3	1	697	4.35	0.015	25.6
Posterior mean	702.31	4.44	0.016	13.5	0.8	702.30	4.39	0.016	27.3

Table 1. Posterior means and modes. Logistic model

	Exponential Power model					Normal model			
	θ_0	θ'_1	θ'_2	σ	β	θ_0	θ'_1	θ'_2	σ
Posterior mode	726.5	2.55	0.55	15.8	0.8	721.5	2.55	0.55	33
Posterior mean	723.8	2.567	0.543	19.5	0.75	721.37	2.57	0.54	33.8

Table 2. Posterior means and modes. Gompertz model

	Logistic model		Gompertz model	
	Exponential Power	Normal	Exponential Power	Normal
$\sum(d_r)^2$	9593.30	9646.06	14632.62	14993.61
$\sum\|d_r\|$	258.06	260.68	370.67	377.23

Table 3. Predictive fit measures for both models

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	$b_{8,1}$	σ	β
random β	10.1	-1.25	-0.75	0.15	1.21	-0.22	-4.31	-0.60	0.50	-5.86	0.51	1
β=0.	9.93	-0.04	-0.16	0.004	0.309	-0.34	-4.29	-0.32	0.35	-7.5	1.83	

Table 4. Posterior modes for some parameters. M_1model

Model M_1, random β			Model M_1, $\beta = 0$	
y_{ij}	$\hat{E}(Y_{ij}\| \mathbf{y}_{(ij)})$	$\|d_{ij}\|$	$\hat{E}(Y_{ij}\| \mathbf{y}_{(ij)})$	$\|d_{ij}\|$
10.67	1.72	8.94	1.68	8.99
8.42	11.82	3.4	12.87	4.44
9.43	-4.71	14.15	-1.86	11.30
6.32	11.91	5.59	2.13	4.18
8.12	12.04	3.91	13.07	4.95
4.58	5.93	1.35	6.50	1.92
...	...	...	...	...
$\sum d_{ij}^2$		76010		77961

Model M_2, random β			Model M_2, $\beta = 0$	
y_{ij}	$\hat{E}(Y_{ij}\| \mathbf{y}_{(ij)})$	$\|d_{ij}\|$	$\hat{E}(Y_{ij}\| \mathbf{y}_{(ij)})$	$\|d_{ij}\|$
10.67	7.31	3.36	5.84	4.82
8.42	7.99	0.42	10.08	1.66
9.43	9.30	0.12	4.71	4.72
6.32	6.79	0.47	7.77	1.44
8.12	5.65	2.46	5.69	2.42
4.58	3.80	0.78	5.10	0.52
	...	...	...	...
$\sum d_{ij}^2$		5159		15865

Table 5. Predictions for the 6 first cases and predictive fit measure for both models

Figure Captions

figure 1. Box plots of measures of discrepancy obtained in 30 size $n = 100$ samples from Exponential Power distribution with parameters $\beta = -0.75$, -0.5, $-0.25, 0, 0.25, 0.5, 0.75$.

figure 2. Posterior distributions for σ and β . Logistic model.

figure 3. Posterior distribution for parameters $a_1, b_{8,1}, \sigma$ and β. M_1 model

Figures

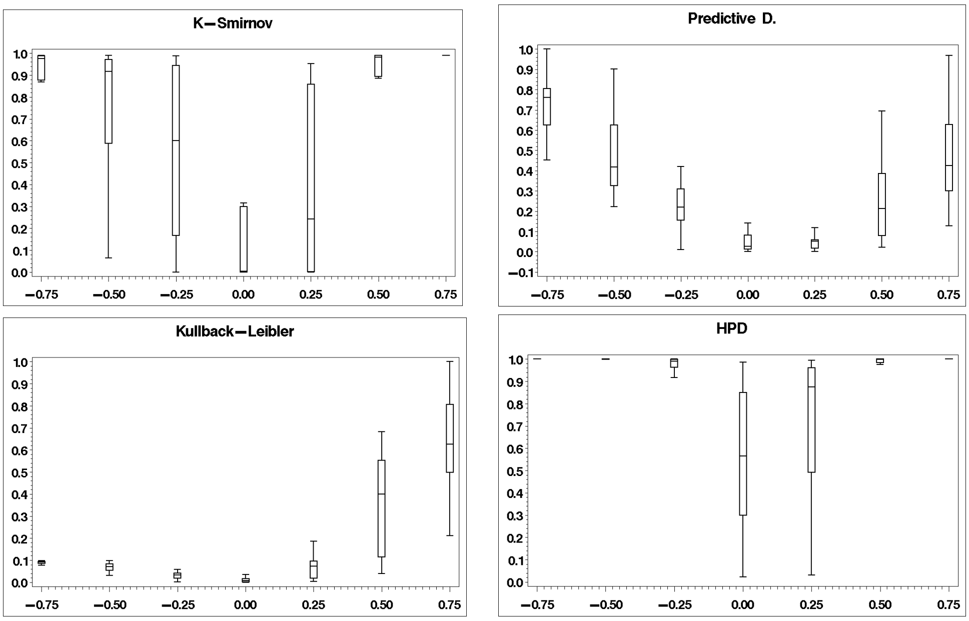
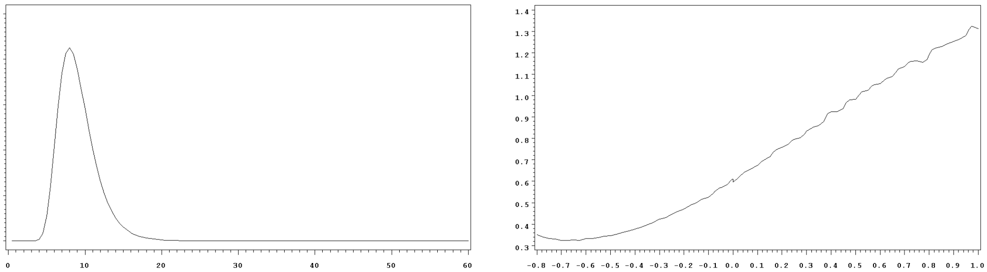
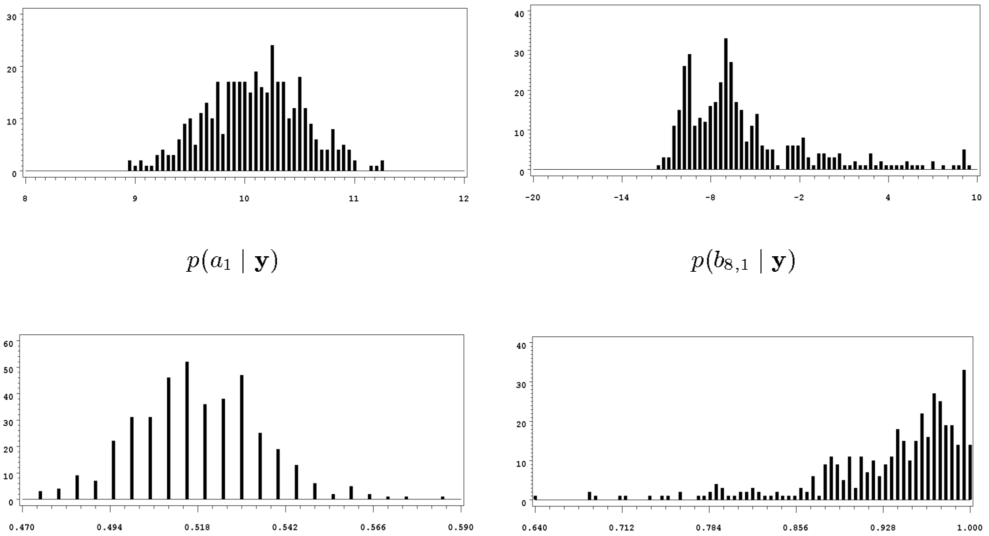


Figure 1



$p(\sigma \mid \mathbf{y})$ $p(\beta \mid \mathbf{y})$

Figure 2



$p(a_1 \mid \mathbf{y})$

$p(b_{8,1} \mid \mathbf{y})$

$p(\sigma \mid \mathbf{y})$

$p(\beta \mid \mathbf{y})$

Figure 3