# Reference Bayesian analysis for hierarchical models

Thaís C. O. Fonseca [*]       Helio S. Migon [†]       Heudson Mirandola [‡]

**Abstract**

In this paper, we propose an approach for constructing objective prior distributions for hierarchical models. In particular, we provide a flexible decomposition for the Fisher information tailored for hierarchical or multilevel models. This proposal takes advantage of the hierarchy and leads to an alternative way of computing Jeffreys priors for the hyperparameters and an upper bound for the prior information. While the Jeffreys prior gives the minimum information about parameters, the proposed bound gives an upper limit for the information put in the prior. Any prior with information above that limit may be considered too informative. From a practical point of view, the proposed prior may be evaluated computationally as part of a MCMC algorithm. This property might be essential for modeling setups with many levels in which analytic marginalization is not feasible. We illustrate the usefulness of our proposal with examples in mixture models, in model selection priors such as lasso and in the Student-t model.

Keywords: Invariant prior, information geometry, Kullback-Leibler distance

## 1   Introduction

There has been an increasing interest over the last decade in the construction of flexible and computational efficient models such as neural networks, clustering, multilevel, spatial random effects and mixture models. These models are setup in an hierarchical

---

[*]Department of Statistics, Universidade Federal do Rio de Janeiro, RJ, Brazil, thais@im.ufrj.br
[†]Department of Statistics, Universidade Federal do Rio de Janeiro, RJ, Brazil, migon@im.ufrj.br
[‡]Department of Mathematics, Universidade Federal do Rio de Janeiro, RJ, Brazil, mirandola@im.ufrj.br

fashion which naturally describe several real data characteristics. These models are designed to be flexible and complexity is take into account including extra levels in the hierarchy. Very often subjective prior elicitation for the hyperparameters is not trivial, with parameters being defined in low levels of the hierarchy and lacking practical interpretation. Frequently, the parameters are calibrated or estimated using empirical Bayes approaches. Recent work has focus attention in the specification of hyperpriors for parameters in low levels of the hierarchy such as the penalizing complexity priors of Simpson et al. (2017) which is weakly informative and penalizes parameter values far from the base model specification. In this context, objective prior specification may be an alternative for full Bayesian inference.

Consider $y$ the vector of observed responses and a probabilistic parametric model $\mathcal{M} = \{f(. \mid \theta), \theta \in \Theta\}$. Direct inference about $\theta$ using the integrated likelihood $f$ is not always trivial and levels of hierarchy are often introduced in the modelling to allow for feasible inferences regarding $\theta$ due to conditional independence given the latent variables. As follows, we assume that the mechanism which has generated the data is a data augmentation process: given $\theta$, a value of $w$ is selected from $f_2(w \mid \theta)$ and, given $w$, a value of $y$ is selected from $f_1(y \mid \theta, w)$. The model for $w \in \Re^d$ depends on the application and is often chosen to allow easier inferences about $\theta$ or due to easier interpretability. In this setup the model has two elements: the extended likelihood $f_1(y \mid w, \theta)$ and the model for the latent variable $f_2(w \mid \theta)$. The marginal likelihood for $\theta$ given the data is obtained by integration

$$f(y \mid \theta) = \int f_1(y \mid \theta, w) dF_2(w \mid \theta). \tag{1}$$

with $F_2$ the cumulative distribution associated with $f_2$. Notice, however, that the integration step is not usually desirable as in general flexible hierarchical modelling interest lies in making inference for the hidden effects $w$ as well as the parameters $\theta$. The introduction of latent variables in the inferential problem brings great benefits as very often the complete conditional distributions of $w \mid \theta, y$ have nice explicit forms. In an iterative algorithm as the one proposed by Tanner and Wong (1987), a sample is obtained from $p(w \mid \theta, y)$ and a sample from $p(\theta \mid w, y)$ is obtained conditional on the sampled value of $w$. Both complete conditional distributions are often easy to sample from.

2

Another important aspect of the hierarchical approach is that the model is usually a flexible version of a base model as discussed in Simpson et al. (2017). This naturally leads to ill behaved likelihoods as the lower levels in the hierarchy may converge to a constant and the model converges to the base model. This may happen with quite high probability if the sample size is not too large. This bad behaviour of the likelihood function is not corrected by reparametrization and the use of informative priors will mean that the inference made a priori and a posteriori will be approximately the same. Examples with this characteristic are the Student-t model (Fernandez and Steel, 1999; Fonseca et al., 2008), the skewed normal distribution (Liseo and Loperfido, 2006), mixture models (Bernardo and Giron, 1988; Grazian and Robert, 2018), the hyperbolic model (Fonseca et al., 2012). In these cases, the use of Jeffreys priors correct the ill behaved likelihood and provide better frequentist properties than the ones achieved by using informative priors.

The objetive Bayes approach has the pionner works of Bayes (1763) and Laplace (1812) which attempted to provide default inferences by setting uniform prior for unknown parameters. However, these proposals are not invariant to transformations. Jeffreys (1946) introduced invariant priors based on distances between distributions. In particular, the distances considered for prior derivation were the squared Hellinger and the Kullback-Leibler distances. Both distances behave locally as the Fisher information. George and McCulloch (1993) considered a large class of invariant priors based on discrepancy measures. The motivation of these authors was to think in terms of the parametric family $\mathcal{M}$ instead of the parameter domain $\Theta$. Thus, the prior weight assigned to a neighbor value of $\theta$, say $\theta^*$, depends directly on the discrepancy between the parametric family members $f(. \mid \theta)$ and $f(. \mid \theta^*)$.

In the context of hierarchical models, it is natural that the information to estimate $\theta$ contained in the complete data $(y, w)$ is larger than the information contained in the incomplete data $y$. That is, if $w$ and $y$ were known it would be easier to make inference about $\theta$. A way to measure information in an inferential problem is through discrepancy functions such as Fisher information. Notice that the Fisher information about $\theta$ obtained from the complete data problem must be larger than the one obtained from the incomplete data problem.

In this work, we present a Fisher information decomposition for a general hierar-

chical problem which allows for computing the information about $\theta$ in the incomplete model $f$ and in the extended model $f_1$ and $f_2$ overcoming the marginalization step. From the proposed decomposition we obtain an alternative way to compute the Jeffreys prior for $\theta$ and an upper bound for this prior. The Jeffreys rule prior gives the minimum prior information for inference about $\theta$. On the other hand, any prior which gives more information than the upper bound may be too informative.

As follows we propose a Fisher information decomposition, the upper bound for the Jeffreys prior of $\theta$ and an alternative way to compute the usual Jeffreys rule prior for an hierarchical model.

Many alternative hierarchical models, often explored in the current literature will be presented in the next Sections. The first is a discrete mixture model (Bernardo, Giron, 1988).

# 2 Fisher matrix decomposition and proposed non-informative prior

The usual Jeffreys rule prior is obtained as $\pi_y(\theta) \propto \sqrt{\det(I_y(\theta))}$, with $I_y(\theta) = [I_y(\theta)_{ij}]$ being the Fisher information matrix of the parametric model $f(y \mid \theta)$,

$$I_y(\theta)_{ij} = -E_{f(y|\theta)}\left[\frac{\partial^2}{\partial\theta_i\theta_j} \log f(y \mid \theta)\right]. \tag{2}$$

Also, let $I_y(\theta \mid w)$ denote the Fisher information matrix of the model $f(y \mid w, \theta)$. Similarly, we define $I_w(\theta)$, $I_{y,w}(\theta)$ and $I_w(\theta \mid y)$. Our first result is an alternative method to compute the Fisher information matrix based on the hierarchical structure of the model (1).

**Theorem 2.1.** *The Fisher information matrices of the hierarchical model* (1) *can be decomposed as*

$$I_{y,w}(\theta) = I_y(\theta) + E_y[I_w(\theta \mid y)] = I_w(\theta) + E_w[I_y(\theta \mid w)]. \tag{3}$$

For the sake of simplicity, in the expression above, $E_y$ denotes $E_{f(y|\theta)}$ and $E_w$ denotes $E_{f_2(w|\theta)}$. If $E_y[I_w(\theta \mid y)]$ is not easily computed then we may take advantage of the hierarchy, which leads to $E_{f(y|\theta)}[I_w(\theta \mid y)] = E_{f_2(w|\theta)}\{E_{f_1(y|w,\theta)}[I_w(\theta \mid y)]\}$.

To prove Theorem 2.1, we observe that the Fisher information matrix can be derived from the Kulback-Leibler (KL) divergence. Namely, for any parametric model $p(y \mid \theta)$ it holds

$$KL[p(y \mid \theta) : p(y \mid \theta + d\theta)] = \frac{1}{2} d\theta^T I_y(\theta) \, d\theta + O(|d\theta|^3). \qquad (4)$$

In other words, the Fisher information matrix $I(\theta) = [I(\theta)_{ij}]$ can be obtained as the Hessian of the KL divergence

$$I(\theta)_{ij} = \frac{\partial^2}{\partial \theta^i \partial \theta^j} KL[p(y \mid \theta) : p(y \mid \theta')] \mid_{\theta'=\theta} . \qquad (5)$$

The proposition below is the main ingredient to derive (3).

**Proposition 2.1.** *Consider the probability distributions $f(y,w)$ and $g(y,w)$, where $y, w$ are random vectors. The Kullback-Leibler divergence satisfy:*

$$KL[f(y,w) : g(y,w)] = KL[f(w) : g(w)] + E_{f(w)}[KL[f(y \mid w) : g(y \mid w)]$$
$$= KL[f(y) : g(y)] + E_{f(y)}[KL[f(w \mid y) : g(w \mid y)]. \qquad (6)$$

Proposition 2.1 follows by a direct computation. In fact, the Kullback-Leibler divergence of $f(y,w) = f(y \mid w)f(w)$ and $g(y,w) = g(y \mid w)g(w)$ is given by

$$KL[f(y,w) : g(y,w)] = \int_{\mathcal{Y} \times \mathcal{W}} f(y,w) \log(f(y,w)/g(y,w)) dy dw$$
$$= \int_{\mathcal{W}} f(w) \int_{\mathcal{Y}} f(y \mid w)(\log(f(y \mid w)/g(y \mid w)) + \log(f(w)/g(w))$$
$$= \int_{\mathcal{W}} f(w) \log(f(w)/g(w)) + \int_{\mathcal{W}} f(w) \int_{\mathcal{Y}} f(y \mid w) \log(f(y \mid w)/g(y \mid w))$$
$$= KL[f(w) : g(w)] + E_{f(w)}[KL[f(y \mid w) : g(y \mid w)] \qquad (7)$$

The first term of the third equality holds since $\int_{\mathcal{Z}} f(z \mid w) dz = 1$. The second equality in (6) holds similarly. $\qquad \square$

Now, we will prove Theorem 2.1. For the hierarchical model (1), we consider the likelihood $f(y,w \mid \theta)$ of $\theta$ given the joint random vector $(y,w)$ and the complete conditional posterior distribution $f(w \mid y, \theta)$. Applying Proposition 2.1, we obtain

$$KL[f(y,w \mid \theta) : f(y,w \mid \theta')] = KL[f_2(w \mid \theta) : f_2(w \mid \theta')] + E_{w\mid\theta}[KL[f_1(y \mid w, \theta) : f_1(y \mid w, \theta')]$$
$$= KL[f(y \mid \theta) : f(y \mid \theta')] + E_{y\mid\theta}[KL[f(w \mid y, \theta) : f(w \mid y, \theta')].$$

Taking the Hessian of the above equalities at $\theta' = \theta$, we obtain (3). $\qquad \square$

5

We can use Theorem 2.1 to compute the Fisher information matrices of high-level hierarchical models:

$$y \sim f(y \mid w_k, \dots, w_1, \theta),$$
$$w_k \sim f_k(w_k \mid w_{k-1}, \dots, w_1, \theta),$$
$$\dots,$$
$$w_1 \sim f_1(w_1 \mid \theta). \tag{8}$$

We write $w = (w_1, \dots, w_k)$ and $\mathbf{w_t} = (w_1, \dots, w_t)$, for $t = 1, \dots, k$. Applying recursively Theorem 2.1, it holds

$$I_{y,w}(\theta) = I_y(\theta) + E_y[I_{w_1}(\theta \mid y)] + E_{y,\mathbf{w_1}}[I_{w_2}(\theta \mid y, \mathbf{w_1})] + \dots + E_{y,\mathbf{w_{k-1}}}[I_{w_k}(\theta \mid y, \mathbf{w_{k-1}})]$$
$$= I_{w_1}(\theta) + E_{\mathbf{w_1}}[I_{w_2}(\theta \mid \mathbf{w_1})] + \dots + E_{\mathbf{w_{k-1}}}[I_{w_k}(\theta \mid \mathbf{w_{k-1}})] + E_w[I_y(\theta \mid w)].$$

As a particular case, consider the hierarchy model:

$$y \mid w \sim f_1(y \mid w) \quad \text{and} \quad w \mid \theta \sim f_2(w \mid \theta). \tag{9}$$

Since $f_1(y \mid w)$ does not depend on $\theta$, it holds that $I_y(\theta \mid w) = 0$. Applying Theorem 2.1 we obtain $I_{y,w}(\theta) = I_w(\theta) + E_w[I_y(\theta \mid w)] = I_w(\theta)$. Thus, it holds

**Corollary 2.1.** *For the hierarchical model is $y \mid w \sim f_1(y \mid w)$ and $w \mid \theta \sim f_2(w \mid \theta)$, the posterior Fisher information matrices satisfy*

$$I_y(\theta) = I_w(\theta) - E_y[I_w(\theta \mid y)]. \tag{10}$$

*In particular, if we assume the Jeffreys' prior $\pi_w(\theta)$ of the model $w \mid \theta$ is proper then the Jeffreys' prior $\pi_y(\theta)$ of the marginal likelihood distribution $y \mid \theta$ is also proper.*

It is worth to emphasize that both parâmeters $w$ and $\theta$ can be multivariate. In this case, we can also use of Theorem 2.1 to give a lower bound of the Jeffreys rule prior $\pi_y(\theta) \propto \det(I_y(\theta))^{1/2}$ in terms of the hierarchy. Namely, we prove the following:

**Theorem 2.2.** *Under the hypotheses of Theorem 2.1, the usual Jeffreys rule priors satisfy the inequalities as follows. If the parâmeter $\theta$ is multivariate (dimension $\geq 2$), then*

$$\det(I_{y,w}(\theta))^{\frac{1}{2}} \geq \det(I_y(\theta))^{\frac{1}{2}} + \det\left(E_y[I_w(\theta \mid y)]\right)^{\frac{1}{2}}. \tag{11}$$

*Proof.* Assume the dimension of the parameter space $\Theta = [\theta]$ is $n \geq 2$. The Minkowski determinant theorem (see for instance Theorem 4.1.8 pg 115 of Marcus and Minc (1964)) states, for any $n \times n$ symmetric and positive semi-definite matrices $A$ and $B$,

$$\det(A+B)^{\frac{1}{n}} \geq \det(A)^{\frac{1}{n}} + \det(B)^{\frac{1}{n}}. \tag{12}$$

Fisher information matrices are symmetric and positive semi-definite. By (3) and (12),

$$\det(I_{y,w}(\theta))^{\frac{1}{n}} = \det\left(I_y(\theta) + E_y[I_w(\theta \mid y)]\right)^{\frac{1}{n}}$$
$$\geq \det(I_y(\theta))^{\frac{1}{n}} + \det(E_y[I_w(\theta \mid y)])^{\frac{1}{n}}. \tag{13}$$

And, since $\frac{n}{2} \geq 1$,

$$\det(I_{y,w}(\theta))^{\frac{1}{2}} = \left(\det(I_y(\theta))^{\frac{1}{n}} + \det(E_y[I_w(\theta \mid y)])^{\frac{1}{n}}\right)^{\frac{n}{2}}$$
$$\geq \det(I_y(\theta))^{\frac{1}{2}} + \det(E_y[I_w(\theta \mid y)])^{\frac{1}{2}}.$$

Theorem 2.2 is proved. $\qquad\square$

The examples below will show, in many other situations, how to conveniently decompose $w$ in order to compute the Fisher information matrix of hierarchical models.

# 3 Example 1: Mixture model with known components

Mixture models are a useful representation of several applications such as gene expression, classification, etc. A simple mixture model with known components (Bernardo and Giron, 1988) may be written as

$$f(y \mid \theta) = \sum_{j=1}^{p} \theta_j g_j(y), \tag{14}$$

with $S^n = \{\theta = [\theta_0, \ldots, \theta_p]^T \mid \theta_k > 0 \text{ and } \sum_k \theta_k = 1\}$ a simplex defining the weights in the mixture and $g_k(y)$, $k = 0, \ldots, p$ completely specified density functions with known parameters. We set $w$ assuming values in $\{0, \ldots, p\}$ with categorical distribution $f_2(w = k \mid \theta) = \theta_k$, with $k \in \{0, \ldots, p\}$. For each $w \in \{0, \ldots, p\}$, we fix a probability distribution $f_1(y \mid w) = g_w(y)$. The two-level hierarchical model $f_1(y \mid w)$ and $f_2(w \mid \theta)$

yields the mixture of probabilities $f(y \mid \theta) = \sum_w \theta_w g_w(y)$. The Fisher information matrix of the mixture model $f(y \mid \theta)$ is

$$I_y(\theta)_{ij} = \int (g_i(y) - g_0(y))(g_j(y) - g_0(y)) \frac{1}{f(y \mid \theta)} dy, \tag{15}$$

with $i, j = 1, \ldots, n$. It is not a simple task to show the properness of Jeffreys' prior $\pi_y(\theta)$. In fact, this result was obtained by Grazian and Robert (2018). Corollary 2.1 gives an alternative proof of this result as we discuss as follows.

The Fisher information matrix based on the model $f_2(w \mid \theta)$, $I_w(\theta) = [I_w(\theta)_{ij}]$, with $i, j = 1, \ldots, p$, is given by

$$I_w(\theta)_{ij} = \frac{1}{\theta_0} + \frac{1}{\theta_i} \delta_{ij},$$

where $\delta_{ij}$ is the Kronecker delta. This results in the upper bound $I_y(\theta) < I_w(\theta)$.

It is well known the Jeffreys' prior of the categorical model is proper with normalizing constant $\int \det(I_w(\theta))^{\frac{1}{2}} d\theta = \frac{1}{2^{p+1}} \mathrm{vol}(S_2^p) = \frac{\pi^{\frac{p+1}{2}}}{\Gamma(\frac{p+1}{2})}$, where $S_2^p \subset \mathbb{R}^{p+1}$ denotes the round sphere in $\mathbb{R}^{p+1}$ of radius 2. By Corollary 2.1, the Jeffreys' prior of the incomplete model based on $I_y(\theta)$ is also proper.

In the simpler model $w \mid \theta \sim Bern(\theta)$, one has $I_w(\theta) = \theta^{-1}(1-\theta)^{-1}$. Firstly, we take advantage of the hierarchy to obtain conjugated posterior for $w \mid y, \theta$. The complete conditional posterior distribution of $w$ is

$$w \mid y, \theta \sim Bern\left(\frac{\theta g_1(y)}{\theta g_1(y) + (1-\theta) g_2(y)}\right).$$

The Fisher information of $w \mid \theta, y$ about $\theta$ is given by

$$I_w(\theta \mid y) = \frac{g_1(y)/\theta}{\theta g_1(y) + (1-\theta) g_2(y)} + \frac{g_2(y)/(1-\theta)}{\theta g_1(y) + (1-\theta) g_2(y)} + \frac{(g_1(y) - g_2(y))^2}{[\theta g_1(y) + (1-\theta) g_2(y)]^2}.$$

Expectation of $I_w(\theta \mid y)$ with respect to the model $f(y \mid \theta)$ is obtained computationaly in the estimation algorithm. The computation of $E[I_w(\theta \mid y)]$ leads to less error then directy computation of $I_y(\theta)$. The resulting Jeffreys rule prior for the mixture model (14) is then

$$\pi(\theta) = \left\{ \theta^{-1}(1-\theta)^{-1} - E[I_w(\theta \mid y)] \right\}^{1/2}. \tag{16}$$

The resulting prior must be integrable as $I_w(\theta)^{1/2} = \theta^{-1/2}(1-\theta)^{-1/2}$ also is (Bernardo and Giron, 1988; Grazian and Robert, 2018). This result is immediate from Corollary 2.1.

# 4 Example 2: Student-t model with unknown degrees of freedom

Let $y \mid \theta \sim \text{ST}(\theta)$ be a standard Student-t model with fixed mean and precision and unknown degrees of freedom $\nu$. This model robustfy the Gaussian model by allowing for fatter tails. Several authors have dealt with reference prior specification for the tail parameter $\nu$ (Fonseca et al., 2008; Villa and Walker, 2014; Simpson et al., 2017). This model may be rewritten in an hierarchical setting as

$$Y|w \sim N(0, 1/w),$$

$$w \mid \theta \sim Ga(\theta/2, \theta/2).$$

This model has two levels of hierarchy with $\theta$ appearing in the second level only, as in (9). In this case, by Corollary 2.1, the Fisher decomposition simplifies to

$$I_y(\theta) = I_w(\theta) - E_y[I_w(\theta \mid y)]. \tag{17}$$

The first term is obtained from the latent variable prior distribution $f_2(w \mid \theta)$ which is a gamma distribution and twice differentiation and expectation with respect to $w \mid \theta$ leads to

$$I_w(\theta) = \frac{1}{4} \, \psi^{(2)}\left(\frac{\theta}{2}\right) - \frac{1}{2\theta}. \tag{18}$$

The second part in (17) is computed based on the complete condicional posterior distribution of the latent variable $w$ which is $w \mid \theta, y \sim Ga((\theta+1)/2, (y^2 + \theta)/2)$. Twice differentiation and taking expectation in the distribution of $w \mid \theta, y$ results in

$$I_w(\theta \mid y) = \frac{1}{4} \, \psi^{(2)}\left(\frac{\theta+1}{2}\right) + \frac{\theta+1}{2(\theta + y^2)^2} - \frac{1}{\theta + y^2}.$$

Now we take expectations with respect to the model $f(y \mid \theta)$. For this model the result is analitic and given by

$$E[I_w(\theta \mid y)] = \frac{1}{4} \, \psi^{(2)}\left(\frac{\theta+1}{2}\right) + \frac{\theta+2}{2\theta(\theta+3)} - \frac{1}{\theta+1}. \tag{19}$$

Equations (18) and (19) together result in the usual Jeffreys rule prior for the degrees of freedom of the Student-t model. The Jeffreys prior is bounded above by the function $p(\theta) = kI_w(\theta)^{1/2}$ which is not integrable. However, $I_y(\theta) \asymp \mathcal{O}(\theta^{-4})$ as $\theta \to \infty$. Thus, the Jeffreys prior for the degree of freedom of the Student-t model is proper.

# 5    Example 3: Hyperbolic model

This example illustrates the use of our proposed decomposition when all levels of hierarchy depend on the parameter of interest. In particular, we present results for the Hyperbolic model which is an extension of the Gaussian model which allows for asymmetric behavior. The Jeffreys rule prior for this model has been proposed by Fonseca et al. (2012), however, the prior is only obtained computationally and no propertness of the resulting proposal is proved.

A random variable is said to have a Hyperbolic distribution with parameters $\alpha$, $\beta$, $\delta$ e $\mu$ when its density is given by

$$g(y; \alpha, \beta, \delta, \mu) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2 - \beta^2})} \exp\left\{-\alpha\sqrt{\delta^2 + (y-\mu)^2} + \beta(y-\mu)\right\} \qquad (20)$$

with $y, \mu \in \Re$, $\delta > 0$ e $|\beta| < \alpha$ and $K_1(.)$ the modified Bessel function of the tird kind with index 1. This model can be written alternatively as a mixture of a Gaussian and a GIG[1] distribution, that is,

$$f(y|\lambda, \alpha, \beta, \delta, \mu) = \int_0^\infty f_1(y|\mu, \beta, w) f_2(w|\lambda, \delta, \alpha, \beta) dw, \qquad (21)$$

with $f_1$ the Gaussian distribution with mean $\mu + \beta w$ and variance $w$ and $f_2$ the Generalized Inverse Gaussian distribution (GIG) with parameters 1, $\sqrt{\alpha^2 - \beta^2}$ and $\delta$.

In particular, consider the standard Hyperbolic model (20) for $\mu = 0$, $\delta = 1$ and $\alpha^2 = 2\beta^2$ which corresponds to an asymmetric density function for $y$. Let $y \mid \theta \sim \text{Hyp}(\sqrt{2\theta^2}, \theta, 1, 0)$. This model may be rewritten as

$$y|w, \theta \sim N(\theta w, w),$$

$$w \mid \theta \sim GIG(1, \theta, 1).$$

Now we illustrate the usefulness of the proposed decomposition when both levels of hierarchy depend on $\theta$. By Theorem 2.1, the Fisher decomposition is given by

$$I_y(\theta) = I_w(\theta) - E[I_w(\theta \mid y)] + E[I_y(\theta \mid w)]. \qquad (22)$$

---

[1]If $W \sim GIG(\rho, \gamma, \kappa)$ then its density is given by

$$c(\rho, \gamma, \kappa) \, w^{\rho-1} \exp\left\{-\frac{1}{2}(\kappa^2 w^{-1} + \gamma^2 w)\right\}, \ \kappa, \gamma \geq 0, \ \rho \in \Re, \ w > 0,$$

with $c(\rho, \gamma, \kappa) = \frac{(\gamma/\kappa)^\rho}{2K_\rho(\gamma\kappa)}$. Jørgensen (1982) presents details about this class.

In this case the latent variable $w$ has a GIG prior distribution and differentiating $\log f_2(w \mid \theta)$ and taking expectations with respect to $w \mid \theta$ leads to

$$I_w(\theta) = S_1(\theta), \tag{23}$$

with $S_1(\theta) = \frac{K_3(\theta)}{K_1(\theta)} - R_1^2(\theta)$, and $R_1(\theta) = \frac{K_2(\theta)}{K_1(\theta)}$. Notice that in this example the first level of hierarchy $f_1(y \mid w, \theta)$ also depends on $\theta$ thus, to obtain the upper bound for the Jeffreys prior, we need to compute $E[I_y(\theta \mid w)]$. In the first level $y$ has a Gaussian distribution, thus differentiating $\log f_1(y \mid w, \theta)$ and taking expectations with respect to $y \mid w, \theta$ leads to

$$E[I_y(\theta \mid w)] = \frac{1}{\theta} R_1(\theta), \tag{24}$$

The resulting upper bound for the Jeffreys rule prior is given by the integrable function

$$\pi(\theta) = k \left\{ S_1(\theta) + \frac{1}{\theta} R_1(\theta) \right\}^{1/2}.$$

The second term in the right hand side of (22) is computed based on the complete condicional posterior distribution of the latent variable $w$ which in this model is also a GIG distribution given by

$$w \mid \theta, y \sim GIG(1/2, \sqrt{2\theta^2}, \sqrt{y^2 + 1}).$$

Twice differentiation and taking expectation in the distribution of $w \mid \theta, y$ results in

$$I_w(\theta \mid y) = 2(y^2 + 1)S_{1/2}(\theta\sqrt{2(y^2 + 1)}).$$

Now we take expectations with respect to the model $f(y \mid \theta)$ to obtain the Jeffreys rule prior for the asymmetry parameter in the Hyperbolic model $y \mid \theta \sim \text{Hyp}(\sqrt{2\theta^2}, \theta, 1, 0)$.

$$\pi(\theta) = \left\{ S_1(\theta) + \frac{1}{\theta} R_1(\theta) - E[I_w(\theta \mid y)] \right\}^{1/2}. \tag{25}$$

# 6 Example 5: Hierarchical priors for variable selection in regression analysis

This example illustrates the use of hierarchical models for variable selection in regression analysis and a convenient use of our proposed Fisher decomposition. Consider the usual linear regression model for predictor variables $X$ and responses $y$

$$y \mid w_1 \sim N(Xw_1, \sigma^2 I_n), \tag{26}$$

11

where $y$ is the $n \times 1$ vector of responses, $X$ is the $n \times p$ matrix of covariates, and $w_1$ is the $n \times 1$ vector of regression coefficients. Consider the problem of variable selection in that context, that is, if $p$ is large it is desirable to find some $w_1$'s equal or close to zero. Tibshirani (1996) proposed a method called lasso, least absolute shrinkage selection operator, which is able to produce coefficients exactly equal to zero in regression models. The lasso estimate $\hat{w}_1$ is defined by

$$\hat{w}_1 = \arg\min \left\{ \sum_{i=1}^n \left( y_i - \sum_{j=1}^p w_{1,j} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p |w_{1,j}| \leq t. \qquad (27)$$

The parameter $t$, the Lasso parameter, controls the amount of shrinkage that is applied to the estimates. Let $\hat{w}_{1,j}^0$ be the full least square estimates and let $t_0 = \sum_{j=1}^p |\hat{w}_{1,j}^0|$. Values of $t < t_0$ will cause shrinkage of the solutions towards zero. This method aims to improve prediction accuracy and to be more interpretable, as we may focus only in the strongest effects. However, estimation of $t$ is not an easy task, Tibshirani (1996) comments that since the lasso estimate is a non-linear and non-differentiable function of the response, it is difficult to obtain an accurate estimate of its standard error.

The lasso constraint $\sum_{j=1}^p |w_{1,j}| \leq t$ is equivalent to the addition of a penalty term $\gamma \sum |w_{1,j}|$ to the residual sum of squares, that is, we should minimize

$$\sum_{i=1}^n \left( y_i - \sum_{j=1}^p w_{1,j} x_{ij} \right)^2 + \gamma \sum_{j=1}^p |\beta_j|. \qquad (28)$$

In the Bayesian context, this is equivalent to the use of independent Laplace priors for the regression coefficients (Park and Casella, 2008), that is,

$$p(w_{1,j}) = \frac{\gamma}{2} \exp\left\{-\gamma |w_{1,j}|\right\}. \qquad (29)$$

We may derive the Lasso estimate as the posterior mode under the prior (29) for the $w_1$'s. The choice of the Bayesian Lasso parameter is also not trivial and several methods has been proposed to estimate this parameter such as cross validation and empirical Bayes, however, these methods are often unstable. It is pointed out by Park and Casella (2008) that the standard error obtained for the lasso parameter are not fully satisfactory.

Mallick and Yi (2014) rewrites the Laplace prior for $w_1$ in a convenient hierarchical way which allows for analytical full conditional distributions of the latent variables

and model parameters. The lasso prior is obtained as an uniform scale mixture by considering the conditional setting

$$w_{1,j} \mid w_{2,j} \quad \sim \quad Unif(-\sigma w_{2,j}, \sigma w_{2,j}) \tag{30}$$

$$w_{2,j} \mid \theta \quad \sim \quad Ga(2,\theta) \tag{31}$$

As follows we obtain the Jeffreys prior for $\theta$ based on the Fisher decomposition (9) which simplifies to

$$I_y(\theta) = I_{w_2}(\theta) - E_y[I_w(\theta \mid y)], \tag{32}$$

with $w = (w_1, w_2)$. The Fisher information for the bottom level in the hierarchy is easily obtained as $I_{w_2}(\theta) = \frac{2p}{\theta^2}$. In order to obtain $E_y[I_w(\theta \mid y)]$ we take advantage of the known complete conditional distribution of $w \mid \theta, y$ which is presented in Mallick and Yi (2014) as

$$w_{1,j} \mid w_{2,j}, \theta, y \quad \sim \quad N(\hat{w}_1^{OLS}, \sigma^2(X'X)^{-1}) \prod_{j=1}^{p} I(|w_{1,j}| \leq \sigma w_{2,j}) \tag{33}$$

$$w_{2,j} \mid \theta, y \quad \sim \quad Exp(\theta), \tag{34}$$

with $\hat{w}_1^{OLS}$ the ordinary least squared estimator of $w_1$ in the usual regression model (26). The distribution of $w_{1,j} \mid w_{2,j}, \theta, y$ does not depend on $\theta$. Thus, $E_y[I_w(\theta \mid y)]$ depend on the distribution of $w_{2,j} \mid \theta, y$ only and is trivialy obtained as $E_y[I_w(\theta \mid y)] = \frac{p}{\theta^2}$. The final Jeffreys prior has closed form given by

$$\pi(\theta) \propto \theta^{-1}. \tag{35}$$

# 7 Example 5: Two level Hierarchical Normal model

The usual random effects models depends strongly on the estimation of the unknown variances in each level of the hierarchy. The signal to noise ratio has been studied in several areas of applications. As follows we consider the two level hierarchical random effects model in its simplest version as a prototypical example.

$$y_i \quad = \quad \omega_i + \epsilon_{1i}$$

$$\omega_i \quad = \quad \mu + \epsilon_{2i}$$

with $\epsilon_{1i} \sim N[0, \delta^{-1}]$, $\epsilon_{2i} \sim N[0, \phi^{-1}]$ and $(\delta^{-1}, \phi^{-1}) \in \Re^+ \times \Re^+$. Assume that $\delta$ is known. In this setting $y|\omega, \delta) \sim N[\omega, \delta^{-1}]$, $\omega|\mu \sim N[\mu, \phi^{-1}]$ and integrating $\omega$ out yields the marginal model $y|\omega \sim N[\mu, \delta^{-1} + \phi^{-1}]$. This model may be seen as the simplest version of several hierarchical models which depends on random effects estimation. It is worth noting that in this example $y \perp \phi|\omega$, so the second term in the left hand side of the information identity is null.

The prior and posterior densities of $\omega$ are both Gaussian distributions with the same general form

$$\omega|\phi, y \sim N[a(\phi), A(\phi)]$$

with $a(\phi) = 0$ and $A(\phi) = \frac{1}{\phi}$ in the prior distribution and $a(\phi) = \frac{y}{\phi+1}$ and $A(\phi) = \frac{1}{\phi+1}$ in the posterior distribution, where, without loss of generality, we suppose that $\mu = 0$ and $\delta = 1$. The log conditional density is

$$l(\phi) = -\frac{1}{2}log(2\pi) - \frac{1}{2}log(A(\phi)) - \frac{1}{2A}(\omega - a(\phi))^2$$

Twice differentiation and expectation in the distribution of $\omega$ results in

$$E(-l''(\phi)) = \frac{1}{2}\left(\frac{A'(\phi)}{A(\phi)}\right)^2 + \frac{(a(\phi))^2}{A(\phi)}$$

In the prior distribution $A'(\phi) = -\frac{1}{\phi^2}$ and it follows that $I_\omega(\phi) = \frac{1}{2\phi^2}$. In the posterior distribution $a'(\phi) = -\frac{y}{(1+\phi)^2}$, $A'(\phi) = -\frac{1}{(1+\phi)^2}$ and $I_\omega(\phi \mid y) = \frac{1}{2(1+\phi)^2} + \frac{y^2}{(1+\phi)^3}$. Taking expectation in the marginal data distribution leads to

$$E_y[I_\omega(\phi|y)] = \frac{2 + \phi}{2\phi(\phi + 1)^2} \tag{36}$$

Finally, applying the proposed Fisher decomposition

$$I_y(\phi) = I_w(\phi) - E[I_w(\phi \mid y)] = \frac{1}{2}\left[\frac{1}{\phi^2} - \frac{\phi + 2}{\phi(\phi + 1)^2}\right] \tag{37}$$

In this example, the marginal model is easily obtained and usual differentiation of the log likelihood leads to the same result obtained in (37).

# References

Bernardo, J. M. and Giron, F. J. (1988). "A Bayesian analysis of simple mixture problems." *Bayesian Statistics*, 3, 67–78.

Fernandez, C. and Steel, M. F. J. (1999). "Multivariate Student-t regression models: Pitfalls and inference." *Biometrika*, 86, 153–167.

Fonseca, T. C. O., Ferreira, M. A. R., and Migon, H. S. (2008). "Objective Bayesian analysis for the Student-t regression model." *Biometrika*, 95, 325–333.

Fonseca, T. C. O., Migon, H. S., and Ferreira, M. A. R. (2012). "Bayesian analysis based on the Jeffreys prior for the hyperbolic distribution." *Brazilian Journal of Probability and Statistics*, 26, 327–343.

George, E. I. and McCulloch, R. (1993). "On obtaining invariant prior distributions." *Journal of statistical planning and inference*, 37, 169–179.

Grazian, C. and Robert, C. P. (2018). "Jeffreys priors for mixture estimation: Properties and alternatives." *Computational Statistics & Data Analysis*, 121, 149–163.

Jeffreys, H. (1946). "An invariant form for the prior probability in estimation problems." *Proceedings of the Royal Statistical Society London*.

Liseo, B. and Loperfido, N. (2006). "Default Bayesian analysis of the skew-normal distribution." *Journal of Statistical Planning and Inference*, 136, 373–389.

Mallick, H. and Yi, N. (2014). "A New Bayesian Lasso." *Stat Interface*, 7, 571–582.

Marcus, M. and Minc, H. (1964). *A Survey of Matrix Theory and Matrix Inequalities*. Allyn and Bacon Inc., Boston.

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103, 482, 681–686.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sorbye, S. H. (2017). "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science*, 32, 1–28.

Tanner, M. A. and Wong, W. H. (1987). "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association*, 82, 528–540.

Tibshirani, R. (1996). "Regression shirinkage and selection via the Lasso." *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Villa, C. and Walker, S. G. (2014). "Objective Prior for the Number of Degrees of Freedom of a t Distribution." *Bayesian Analysis*, 9, 197–220.