Divergence Based priors for the problem of hypothesis testing

gonzalo García-Donato and Susie Bayarri

May 22, 2009
Much of the material presented in this talk is based on the work of Jeffreys.
Much of the material presented in this talk is based on the work of Jeffreys.

- Sir Harold Jeffreys born in England in 1891 and died in 1989 (almost at the age of hundred). He was Professor of Astronomy at Cambridge.
Much of the material presented in this talk is based on the work of Jeffreys.

- Sir Harold Jeffreys born in England in 1891 and died in 1989 (almost at the age of hundred). He was Professor of Astronomy at Cambridge.
- His contributions to statistics through a Bayesian thinking are very popular (eg. Jeffreys’ prior) and are the basis of many important methods.
We focus here on his ideas on testing
We focus here on his ideas on testing

- specially on the solutions he provided to testing problem, but also and very importantly,
We focus here on his ideas on testing

- specially on the solutions he provided to testing problem, but also and very importantly,
- on his particular perception of the problem.
We focus here on his ideas on testing

- specially on the solutions he provided to testing problem, but also and very importantly,
- on his particular perception of the problem.

This has motivated the design of this presentation:
- PART I. Looking at testing problem through Jeffreys’ eyes,
We focus here on his ideas on testing

- specially on the solutions he provided to testing problem, but also and very importantly,
- on his particular perception of the problem.

This has motivated the design of this presentation:
- PART I. Looking at testing problem through Jeffreys’ eyes,
- PART II. Priors for testing,
We focus here on his ideas on testing

- specially on the solutions he provided to testing problem, but also and very importantly,
- on his particular perception of the problem.

This has motivated the design of this presentation:

- PART I. Looking at testing problem through Jeffreys’ eyes,
- PART II. Priors for testing,
- PART III. Generalizations of Jeffreys’ ideas,
We focus here on his ideas on testing

- specially on the solutions he provided to testing problem, but also and very importantly,
- on his particular perception of the problem.

This has motivated the design of this presentation:
- PART I. Looking at testing problem through Jeffreys’ eyes,
- PART II. Priors for testing,
- PART III. Generalizations of Jeffreys’ ideas,
- PART IV. Divergence based priors.
PART I. Looking at testing problem through Jeffreys’ eyes
Part I. Looking at testing problem through Jeffreys’ eyes

Testing a point null hypothesis

Let $y = (y_1, y_2, \ldots, y_n)$ be a random sample from $f(y \mid \theta, \nu)$. We are interested in testing the hypotheses

$$H_1 : \theta = \theta_0 \text{ (null)}, \quad \text{vs} \quad H_2 : \theta \neq \theta_0 \text{ (alternative)}.$$
Let $y = (y_1, y_2, \ldots, y_n)$ be a random sample from $f(y | \theta, \nu)$. We are interested in testing the hypotheses

$$H_1 : \theta = \theta_0 \text{ (null)}, \quad \text{vs} \quad H_2 : \theta \neq \theta_0 \text{ (alternative)}.$$

This is equivalent to selecting between the competing models

$$M_1 : f_1(y | \nu) = f(y | \theta_0, \nu) \quad \text{vs} \quad M_2 : f_2(y | \theta, \nu^*) = f(y | \theta, \nu^*).$$
Testing a point null hypothesis

Let \( y = (y_1, y_2, \ldots, y_n) \) be a random sample from \( f(y | \theta, \nu) \). We are interested in testing the hypotheses

\[
H_1 : \theta = \theta_0 \text{ (null)}, \quad \text{vs} \quad H_2 : \theta \neq \theta_0 \text{ (alternative)}.
\]

This is equivalent to selecting between the competing models

\[
M_1 : f_1(y | \nu) = f(y | \theta_0, \nu) \quad \text{vs} \quad M_2 : f_2(y | \theta, \nu^*) = f(y | \theta, \nu^*).
\]

Jeffreys (1961) calls \( \theta \) the new parameter;
PART I. Looking at testing problem through Jeffreys’ eyes

Testing a point null hypothesis

Let \( y = (y_1, y_2, \ldots, y_n) \) be a random sample from \( f(y | \theta, \nu) \). We are interested in testing the hypotheses

\[
H_1 : \theta = \theta_0 \text{ (null), vs } H_2 : \theta \neq \theta_0 \text{ (alternative)}.
\]

This is equivalent to selecting between the competing models

\[
M_1 : f_1(y | \nu) = f(y | \theta_0, \nu) \text{ vs } M_2 : f_2(y | \theta, \nu^*) = f(y | \theta, \nu^*).
\]

Jeffreys (1961) calls \( \theta \) the new parameter; \( \nu, \nu^* \) are the old (or common) parameters.
Initial considerations

- Due to the statement in $H_1$, this problem has received, in the literature, the name of “punctual”, “precise”, “sharp” hypothesis testing or significance tests (Jeffreys 1961).
Initial considerations

- Due to the statement in $H_1$, this problem has received, in the literature, the name of “punctual”, “precise”, “sharp” hypothesis testing or significance tests (Jeffreys 1961).

- A precise hypothesis is (in many situations) a convenient approximation of the more realistic hypothesis

$$H : \theta = \theta_0 \simeq H : \|\theta - \theta_0\| < \epsilon, \, \epsilon << .$$
Initial considerations

• Due to the statement in $H_1$, this problem has received, in the literature, the name of “punctual”, “precise”, “sharp” hypothesis testing or significance tests (Jeffreys 1961).

• A precise hypothesis is (in many situations) a convenient approximation of the more realistic hypothesis

$$H : \theta = \theta_0 \cong H : \|\theta - \theta_0\| < \epsilon, \epsilon << .$$

See Berger and Delampady (1987) for a quantification of this correspondence.
Initial considerations

- Due to the statement in $H_1$, this problem has received, in the literature, the name of “punctual”, “precise”, “sharp” hypothesis testing or significance tests (Jeffreys 1961).
- A precise hypothesis is (in many situations) a convenient approximation of the more realistic hypothesis

$$H : \theta = \theta_0 \approx H : \|\theta - \theta_0\| < \epsilon, \epsilon << .$$

See Berger and Delampady (1987) for a quantification of this correspondence.
- Common parameters $\nu$ in $M_1$ and $\nu^*$ in $M_2$ do not (necessarily) identify the same magnitude.
PART I. Looking at testing problem through Jeffreys’ eyes

Testing precise hypothesis is not an estimation problem
Testing precise hypothesis is not an estimation problem

Many authors, including Jeffreys, have repeatedly argued that estimation and testing problem are statistical problems of different nature.

Jeffreys, in his book *Theory of probability* right at the beginning of the chapter devoted to testing problems says
PART I. Looking at testing problem through Jeffreys’ eyes

Testing precise hypothesis is not an estimation problem

Many authors, including Jeffreys, have repeatedly argued that estimation and testing problem are statistical problems of different nature. Jeffreys, in his book *Theory of probability* right at the beginning of the chapter devoted to testing problems says

In the last chapters we were concerned with the estimation of the parameters in a law, the form of the law itself being given. We are now concerned with the more difficult question: in what circumstances do observations support a change of the form of the law itself?
(genuine) Testing

Testing (genuine) problems
The proposal $\theta = \theta_0$, is (a priori) plausible and represents, in a broad sense, a theory that needs to be tested.

As opposed to
(genuine) Testing versus estimation

Testing (genuine) problems
The proposal $\theta = \theta_0$, is (a priori) plausible and represents, in a broad sense, a theory that needs to be tested.

As opposed to

Estimation problems
There is no uncertainty about what is the right model and there is no a special value of $\theta$. 
Different problems

According to these ideas, it is not that surprising that the tools used for estimation problems do not work for testing problems:
Different problems

According to these ideas, it is not that surprising that the tools used for estimation problems do not work for testing problems:

- Improper priors produce arbitrary answers,
Different problems

According to these ideas, it is not that surprising that the tools used for estimation problems do not work for testing problems:

- Improper priors produce arbitrary answers,
- same (or worst) about proper flat priors,

\[ \text{Gamma}(0.00000000000000000000001, 0.00000000000000000000001) \]
Different problems

According to these ideas, it is not that surprising that the tools used for estimation problems do not work for testing problems:

- Improper priors produce arbitrary answers,
- same (or worst) about proper flat priors,

\[ \text{Gamma}(0.000000000000000000001, 0.0000000000000000000001) \]

- Credible intervals to ‘solve’ hypothesis testing.
Different problems

According to these ideas, it is not that surprising that the tools used for estimation problems do not work for testing problems:

- Improper priors produce arbitrary answers,
- same (or worst) about proper flat priors,

\[ \text{Gamma}(0.0000000000000000000001, 0.0000000000000000000000001) \]

- Credible intervals to ‘solve’ hypothesis testing.
Jeffreys’ tools

Jeffreys based the solution of the problem

\[ M_1 : f_1(y \mid \nu) \text{ vs } M_2 : f_2(y \mid \theta, \nu^*) , \]

on what we call now Bayes factors:

\[ B_{12} = \frac{m_1(y)}{m_2(y)} , \]

where, the prior marginals are

\[ m_1(y) = \int f_1(y \mid \nu)\pi_1(\nu) d\nu , \]

\[ m_2(y) = \int f_2(y \mid \theta, \nu^*)\pi_2(\theta, \nu^*) d\nu^* d\theta . \]
Other approaches

There are many other Bayesian measures used in the problem of hypothesis testing, like the BIC, DIC, etc.
Other approaches

There are many other Bayesian measures used in the problem of hypothesis testing, like the BIC, DIC, etc.

Principle (Berger y Pericchi 2001)
Testing and model selection methods should correspond, in some sense, to actual Bayes factors, arising from reasonable default prior distributions.
PART I. Looking at testing problem through Jeffreys’ eyes

Other approaches

There are many other Bayesian measures used in the problem of hypothesis testing, like the BIC, DIC, etc.

Principle (Berger y Pericchi 2001)
Testing and model selection methods should correspond, in some sense, to actual Bayes factors, arising from reasonable default prior distributions.

The question of what priors should be used is an open question, of which only partial answers are known.
PART II

Objective testing priors
The normal scenario

\[ Y_i \sim (iid) N(\mu, \sigma^2), \; i = 1, 2, \ldots, n, \; \sigma \text{ unknown.} \]

\[ H_1 : \mu = 0 \; (M_1 : N(0, \sigma^2)), \quad H_2 : \mu \neq 0 \; (M_2 : N(\mu, \tau^2)). \]

\[ B_{12} = \frac{\int p_1(y | \sigma) \pi_1(\sigma) d\sigma}{\int p_2(y | \mu, \tau) \pi_{2.1}(\tau) \pi(\mu | \tau) d\tau d\mu}. \]
Jeffreys’ proposal

\[ \pi_1(\sigma) = h(\sigma), \quad \pi_2(\mu, \tau) = h(\tau)\pi(\mu | \tau). \]
Jeffreys’ proposal

\[ \pi_1(\sigma) = h(\sigma), \quad \pi_2(\mu, \tau) = h(\tau)\pi(\mu \mid \tau). \]

Arguing, with respect to the common parameters \( \mu \) and \( \tau \) are orthogonal so \( \tau \) and \( \sigma \) have the same meaning, justifying the assignment of the same prior, \( h \) in this case.
Jeffreys’ proposal

\[ \pi_1(\sigma) = c\sigma^{-1}, \quad \pi_2(\mu, \tau) = c\tau^{-1}\pi(\mu | \tau). \]

Arguing, with respect to the common parameters \( \mu \) and \( \tau \) are orthogonal so \( \tau \) and \( \sigma \) have the same meaning, justifying the assignment of the same prior, \( h \) in this case. Bayes factor is quite robust to the election of \( h \), and hence we can use \( h(\sigma) = \sigma^{-1} \) (constants cancel out).
Jeffreys’ proposal

\[ \pi_1(\sigma) = c\sigma^{-1}, \quad \pi_2(\mu, \tau) = c\tau^{-1}\pi(\mu \mid \tau). \]

With respect to the new parameter:
\( \pi(\mu \mid \tau) \) has to be proper, centered at zero (the null model), scaled by \( \tau \), symmetric around zero,
Jeffreys’ proposal

\[ \pi_1(\sigma) = c\sigma^{-1}, \quad \pi_2(\mu, \tau) = c\tau^{-1}\pi(\mu | \tau). \]

With respect to the new parameter:
\[ \pi(\mu | \tau) \] has to be proper, centered at zero (the null model), scaled by \( \tau \), symmetric around zero, with heavy tails.
Jeffreys’ proposal

\[ \pi_1(\sigma) = c\sigma^{-1}, \quad \pi_2(\mu, \tau) = c\tau^{-1} Ca(\mu | 0, \tau). \]

With respect to the new parameter:
\[ \pi(\mu | \tau) \] has to be proper, centered at zero (the null model), scaled by \( \tau \), symmetric around zero, with heavy tails. The simpler function with these properties is the \( Ca(0, \tau) \).
Jeffreys’ proposal have resisted the passage of time. It has been used as a benchmark for other priors. Why?

- It provides reasonable responses.
Jeffreys’ proposal have resisted the passage of time. It has been used as a benchmark for other priors. Why?

- It provides reasonable responses.
- It is information consistent: as the information against the null becomes overwhelming the Bayes factor against the null tends to infinity.
Jeffreys’ proposal have resisted the passage of time. It has been used as a benchmark for other priors. Why?

- It provides reasonable responses.
- It is information consistent: as the information against the null becomes overwhelming the Bayes factor against the null tends to infinity.
- Surprisingly the $Ca(0, \tau)$ has arisen, approximately, using other modern methodologies: the intrinsic priors (Berger and Pericchi, 1996) and the expected posterior priors (Pérez and Berger, 2002).
Diving a little bit deeper: a key slide
Diving a little bit deeper: a key slide

- Implicitly recognizes the null model as a source of objective information: $M_0$ gives the center and the scale. This should be interpreted as of unitary size,
Diving a little bit deeper: a key slide

- Implicitly recognizes the null model as a source of objective information: $M_0$ gives the center and the scale. This should be interpreted as of unitary size,

...the mere fact that it it has been suggested that the parameter is zero corresponds to some presumption that it is fairly small.
Diving a little bit deeper: a key slide

- Implicitly recognizes the null model as a source of objective information: $M_0$ gives the center and the scale. This should be interpreted as of unitary size,

  ...the mere fact that it has been suggested that the parameter is zero corresponds to some presumption that it is fairly small.

- The importance of the tails (information consistency).
PART III

Generalizations of Jeffreys’ ideas
Jeffreys’ proposals are very specific to the problem at hand (a normal mean). They do not arise through a rule which can be reasonably applied to other testing scenarios.
Jeffreys' proposals are very specific to the problem at hand (a normal mean). They do not arise through a rule which can be reasonably applied to other testing scenarios.

- Is it always a Cauchy a sensible prior?
Jeffreys’ proposals are very specific to the problem at hand (a normal mean). They do not arise through a rule which can be reasonably applied to other testing scenarios.

- Is it always a Cauchy a sensible prior?
- What is a good scale?
Jeffreys’ proposals are very specific to the problem at hand (a normal mean). They do not arise through a rule which can be reasonably applied to other testing scenarios.

- Is it always a Cauchy a sensible prior?
- What is a good scale?
- What about problems with multivariate parameters or problems with some design?

Although not too many, there are some interesting ideas in the literature about the extension of Jeffreys’ ideas. We start focusing on priors for new parameters. These ideas are the basis for the full problem with nuisance parameters (more on this at the end).
Jeffreys’ proposals are very specific to the problem at hand (a normal mean). They do not arise through a rule which can be reasonably applied to other testing scenarios.

- Is it always a Cauchy a sensible prior?
- What is a good scale?
- What about problems with multivariate parameters or problems with some design?

Although not too many, there are some interesting ideas in the literature about the extension of Jeffreys’ ideas.
Jeffreys’ proposals are very specific to the problem at hand (a normal mean). They do not arise through a rule which can be reasonably applied to other testing scenarios.

- Is it always a Cauchy a sensible prior?
- What is a good scale?
- What about problems with multivariate parameters or problems with some design?

Although not too many, there are some interesting ideas in the literature about the extension of Jeffreys’ ideas.

We start focusing on priors for new parameters. These ideas are the basis for the full problem with nuisance parameters (more on this at the end)
Idea 1: Jeffreys extending Jeffreys

An omnipresent quantity in Jeffreys' work is the divergence:
\[ D = \int \log \frac{f_1}{f_2} \, df_1 - df_2. \]

For the general problem:
\[ H_1: y \text{iid} \sim f(y|\theta_0) \]
\[ H_2: y \text{iid} \sim f(y|\theta), \quad \theta \in \mathbb{R}, \]
Jeffreys proposed the rule
\[ \pi_J(\theta) = \frac{1}{\pi d \tan^{-1}\left(\frac{D[\theta, \theta_0]}{2}\right)}, \]
where \( D[\theta, \theta_0] \) is the divergence between \( f(y|\theta_0) \) and \( f(y|\theta) \).
Idea 1: Jeffreys extending Jeffreys

An omnipresent quantity in Jeffreys’ work is the divergence:

\[ D = \int \log \frac{f_1}{f_2} d(f_1 - f_2). \]
Idea 1: Jeffreys extending Jeffreys

An omnipresent quantity in Jeffreys’ work is the divergence:

\[ D = \int \log \frac{f_1}{f_2} d(f_1 - f_2). \]

For the general problem:

\[ H_1 : y \text{ iid } \sim f(y \mid \theta_0), \quad H_2 : y \text{ iid } \sim f(y \mid \theta), \quad \theta \in \mathcal{R}, \]

Jeffreys proposed the rule
Idea 1: Jeffreys extending Jeffreys

An omnipresent quantity in Jeffreys’ work is the divergence:

\[ D = \int \log \frac{f_1}{f_2} d(f_1 - f_2). \]

For the general problem:

\[ H_1 : y \ iid \sim f(y \mid \theta_0), \quad H_2 : y \ iid \sim f(y \mid \theta), \quad \theta \in \mathcal{R}, \]

Jeffreys proposed the rule

\[ \pi^J(\theta) = \frac{1}{\pi} \frac{d}{d\theta} \tan^{-1}(D[\theta, \theta_0])^{1/2}, \]

where \( D[\theta, \theta_0] \) is the divergence between \( f(y \mid \theta_0) \) and \( f(y \mid \theta) \).
Jeffreys extending Jeffreys (cont’)

\[ \pi^J(\theta) \text{ reduces to Jeffreys Cauchy proposal when } \theta \text{ is a normal mean.} \]

Noticeable, when \(|\theta - \theta_0|\) is small,

\[ \pi^J(\theta) \approx \frac{1}{\pi} (1 + D[\theta, \theta_0])^{-1} \pi^{NJ}(\theta), \]

where \(\pi^{NJ}(\theta)\) is Jeffreys’ (estimation) prior (i.e. the squared root of the expected Fisher information).
Jeffreys extending Jeffreys (cont’)

\( \pi^J(\theta) \) reduces to Jeffreys Cauchy proposal when \( \theta \) is a normal mean. Noticeable, when \( |\theta - \theta_0| \) is small,

\[
\pi^J(\theta) \approx \frac{1}{\pi} (1 + D[\theta, \theta_0])^{-1} \pi^{NJ}(\theta),
\]

where \( \pi^{NJ}(\theta) \) is Jeffreys’ (estimation) prior (i.e. the squared root of the expected Fisher information).

Unfortunately

Note that \( \pi^J \) can lead to improper priors and at least in principle can not be applied for multivariate parameters. However, the approximation was a main inspiration for the definition of DB priors, with clear similarities between them.

For the linear model \( y \sim N(\alpha X_1 + \beta X, \sigma^2 I) \), \( X_1^t X = 0 \), with hypotheses

\[
H_1 : \beta = 0, \quad H_2 : \beta \neq 0,
\]

they propose

\[
\beta \sim Ca_k(0, n\sigma^2 (X^t X)^{-1}),
\]

the matrix \((X^t X)^{-1}\) ‘suggested by the information matrix’.

This is a proposal followed by many others: Liang et al (2008), Bayarri and García-Donato (2007).

Along these same lines, for a general multidimensional parameter $\beta$, Kass and Wasserman (1995) propose

$$\beta \sim Ca_k(0, I_\beta(\alpha, \beta = 0)^{-1}),$$

where $I_\beta$ is the block of the Fisher information matrix corresponding to $\beta$ (one observation).
Notice

- These generalizations focus on the Fisher information, a second order approximation of the Kullback-Leibler divergence. Jeffreys did a similar exercise.
Notice

- These generalizations focus on the Fisher information, a second order approximation of the Kullback-Leibler divergence. Jeffreys did a similar exercise.
- They are primarily concerned with location-type parameters.
Notice

- These generalizations focus on the Fisher information, a second order approximation of the Kullback-Leibler divergence. Jeffreys did a similar exercise.
- They are primarily concerned with location-type parameters.

Our work retakes the essence of the Jeffreys’ ideas using directly divergence measures as a central quantity.
PART IV
Divergence based priors
Motivation: $\theta$ is a location parameter (in $R$)

\[ H_1 : \theta = \theta_0 \quad \text{vs.} \quad H_2 : \theta \neq \theta_0 \]
Motivation: $\theta$ is a location parameter (in $\mathbb{R}$)

$H_1 : \theta = \theta_0$ vs. $H_2 : \theta \neq \theta_0$

$M_1 : f_1(y) = f(y | \theta_0)$ vs. $M_2 : f_2(y | \theta) = f(y | \theta)$

where $f$ is a given model.
Motivation: $\theta$ is a location parameter (in R)

\[ H_1 : \theta = \theta_0 \quad \text{vs.} \quad H_2 : \theta \neq \theta_0 \]

\[ M_1 : f_1(y) = f(y | \theta_0) \quad \text{vs.} \quad M_2 : f_2(y | \theta) = f(y | \theta), \]

where $f$ is a given model.

Let $D[\theta, \theta_0]$ be a measure of unitary ‘distance’ between $M_1$ and $M_2$. We prefer use the Kullback-Leibler divergence divided by $n^*$.

\[ n^* D[\theta, \theta_0] = \int \log \frac{f(y | \theta_0)}{f(y | \theta)} (f(y | \theta) - f(y | \theta_0)) \, dy. \]
Motivation: $\theta$ is a location parameter (in $\mathbb{R}$)

$$H_1 : \theta = \theta_0 \quad \text{vs.} \quad H_2 : \theta \neq \theta_0$$

$$M_1 : f_1(y) = f(y \mid \theta_0) \quad \text{vs.} \quad M_2 : f_2(y \mid \theta) = f(y \mid \theta),$$

where $f$ is a given model.

- Let $D[\theta, \theta_0]$ be a measure of unitary ‘distance’ between $M_1$ and $M_2$. We prefer use the Kullback-Leibler divergence divided by $n^*$.

$$n^* D[\theta, \theta_0] = \int \log \frac{f(y \mid \theta_0)}{f(y \mid \theta)} (f(y \mid \theta) - f(y \mid \theta_0)) \, dy.$$

- Let $h_q(\cdot)$ be a real valued decreasing function, possibly indexed by a real argument $q > 0$. Simple possibilities:

$$h(x) = e^{-qx}, \quad h_q(x) = (1 + x)^{-q}, \quad h_q(x) = (1 + x/q)^{-q}.$$
Definition

\[ \pi_D(\theta) \propto h[q(D[\theta, \theta_0])], \]
being \( q \) such that \( \pi_D \) is proper.

Interpretation
\( \pi_D \) is a robust prior which uses \( M_1 \) as a 'guess' for \( M_2 \): those values of \( \theta \) such that \( f_1(y) \) and \( f_2(y|\theta) \) are close, have the greatest probability.

This idea goes a step beyond other proposals, under which the prior is (an inverse) function of \( (\theta - \theta_0)^2 \) or \( I(\theta_0)^* (\theta - \theta_0)^2 \).
PART IV. Divergence based priors

Definition

Define (θ location)

\[ \pi^D(\theta) \propto h_q(D[\theta, \theta_0]), \]

being \( q \) such that \( \pi^D \) is proper.
PART IV. Divergence based priors

Definition

Define \((\theta \text{ location})\)

\[ \pi^D(\theta) \propto h_q(D[\theta, \theta_0]), \]

being \(q\) such that \(\pi^D\) is proper.

Interpretation

\(\pi^D\) is a robust prior which uses \(M_1\) as a ‘guess’ for \(M_2\): those values of \(\theta\) such that \(f_1(y)\) and \(f_2(y | \theta)\) are close, have the greatest probability.

This idea goes a step beyond other proposals, under which the prior is (an inverse) function of \((\theta - \theta_0)^2\) or \(I(\theta_0) \ast (\theta - \theta_0)^2\).
Basic property

\[ \pi^D \] is a unimodal density, with mode at \( \theta_0 \), symmetric around \( \theta_0 \).
Testing a normal mean \( H_1 : \mu = 0 \), revisited (\( \sigma \) known)

We have \( D[\mu, 0] = \frac{\mu^2}{\sigma^2} \), and hence

\[
\pi^D(\mu) = \frac{\Gamma(q)}{\Gamma(q - .5) \sigma \sqrt{\pi}} (1 + \frac{\mu^2}{\sigma^2})^{-q}, \quad q > 0.5
\]
PART IV. Divergence based priors

Testing a normal mean $H_1 : \mu = 0$, revisited ($\sigma$ known)

We have $D[\mu, 0] = \mu^2 / \sigma^2$, and hence

$$
\pi^D(\mu) = \frac{\Gamma(q)}{\Gamma(q - .5)\sigma\sqrt{\pi}} \left(1 + \frac{\mu^2}{\sigma^2}\right)^{-q}, \quad q > 0.5
$$
Testing a normal mean $H_1 : \mu = 0$, revisited ($\sigma$ known)

We have $D[\mu, 0] = \mu^2 / \sigma^2$, and hence

$$\pi^D(\mu) = \frac{\Gamma(q)}{\Gamma(q - .5)\sigma\sqrt{\pi}} \left(1 + \frac{\mu^2}{\sigma^2}\right)^{-q}, \quad q > 0.5$$
Testing a normal mean $H_1 : \mu = 0$, revisited ($\sigma$ known)

We have $D[\mu, 0] = \mu^2 / \sigma^2$, and hence

$$
\pi^D(\mu) = \frac{\Gamma(q)}{\Gamma(q - 0.5)\sigma\sqrt{\pi}} \left(1 + \frac{\mu^2}{\sigma^2}\right)^{-q}, \quad q > 0.5
$$
Testing a normal mean $H_1 : \mu = 0$, revisited ($\sigma$ known)

We have $D[\mu, 0] = \mu^2 / \sigma^2$, and hence

$$
\pi^D(\mu) = \frac{\Gamma(q)}{\Gamma(q - .5)\sigma\sqrt{\pi}} \left(1 + \frac{\mu^2}{\sigma^2}\right)^{-q}, \quad q > 0.5
$$
Scale parameters: introducing the general rule

The above proposal can be easily applied to scale parameters, $\sigma$:

$M_1 : f_1(y) = f(y \mid \sigma_0)$ vs. $M_2 : f_2(y \mid \sigma) = f(y \mid \sigma)$. 

Reparameterize in terms of $\theta = \log \sigma$, a location parameter, derive the DB prior $\pi_D(\theta) \propto h_q(\mathcal{D}^* [\theta, \log \sigma_0])$, and transform it to the original parameterization. One obtains $\pi_D(\sigma) \propto h_q(\mathcal{D}^* [\log \sigma, \log \sigma_0]) \sigma$. Note, if $\mathcal{D}$ is invariant $\pi_D(\sigma) \propto h_q(\mathcal{D}^* [\sigma, \sigma]) \pi_N(\sigma)$. 

gonzalo garcía-Donato and susie Bayarri ()

DB priors

May 22, 2009 30 / 46
Scale parameters: introducing the general rule

The above proposal can be easily applied to scale parameters, $\sigma$:

$$M_1 : f_1(y) = f(y \mid \sigma_0) \text{ vs. } M_2 : f_2(y \mid \sigma) = f(y \mid \sigma).$$

Reparameterize in terms of $\theta = \log \sigma$, a location parameter, derive the DB prior

$$\pi^D(\theta) \propto h_q(D^*[\theta, \log \sigma_0]),$$

and transform it to the original parameterization.
The above proposal can be easily applied to scale parameters, \( \sigma \):

\[
M_1 : f_1(y) = f(y | \sigma_0) \ vs. \ M_2 : f_2(y | \sigma) = f(y | \sigma).
\]

Reparameterize in terms of \( \theta = \log \sigma \), a location parameter, derive the DB prior

\[
\pi^D(\theta) \propto h_q(D^*[\theta, \log \sigma_0]),
\]

and transform it to the original parameterization. One obtains

\[
\pi^D(\sigma) \propto h_q(D^*[\log \sigma, \log \sigma_0]) \frac{1}{\sigma}.
\]
Scale parameters: introducing the general rule

The above proposal can be easily applied to scale parameters, $\sigma$:

$$M_1 : f_1(y) = f(y \mid \sigma_0) \quad \text{vs.} \quad M_2 : f_2(y \mid \sigma) = f(y \mid \sigma).$$

Reparameterize in terms of $\theta = \log \sigma$, a location parameter, derive the DB prior

$$\pi^D(\theta) \propto h_q(D^*[\theta, \log \sigma_0]),$$

and transform it to the original parameterization. One obtains

$$\pi^D(\sigma) \propto h_q(D^*[\log \sigma, \log \sigma_0]) \frac{1}{\sigma}.$$

Note, if $D$ is invariant

$$\pi^D(\sigma) \propto h_q(D[\sigma, \sigma_0])\pi^N(\sigma).$$
PART IV. Divergence based priors

The general case (without nuisance par.): construction
The general case (without nuisance par.): construction

\[ M_1 : f_1(y) = f(y | \theta_0) \text{ vs. } M_2 : f_2(y | \theta) = f(y | \theta), \]

with \( \theta \) a \( k \)-dimensional parameter. Let \( \pi^N(\theta) \) the reference prior.
The general case (without nuisance par.): construction

\[ M_1 : f_1(y) = f(y \mid \theta_0) \text{ vs. } M_2 : f_2(y \mid \theta) = f(y \mid \theta), \]

with \( \theta \) a \( k \)-dimensional parameter. Let \( \pi^N(\theta) \) the reference prior.

Parameterize in terms of \( \xi = g(\theta) \) such that \( \pi^N(\xi) = 1 \). Remarkably \( \xi \) behaves, asymptotically, as location parameter (Bernardo 2005).
The general case (without nuisance par.): construction

\[ M_1 : f_1(y) = f(y \mid \theta_0) \text{ vs. } M_2 : f_2(y \mid \theta) = f(y \mid \theta), \]

with \( \theta \) a \( k \)-dimensional parameter. Let \( \pi^N(\theta) \) the reference prior.

- Parameterize in terms of \( \xi = g(\theta) \) such that \( \pi^N(\xi) = 1 \). Remarkably \( \xi \) behaves, asymptotically, as location parameter (Bernardo 2005).
- Apply the definition of DB priors for location parameters.
The general case (without nuisance par.): construction

\[ M_1 : f_1(y) = f(y \mid \theta_0) \quad \text{vs.} \quad M_2 : f_2(y \mid \theta) = f(y \mid \theta), \]

with \( \theta \) a \( k \)-dimensional parameter. Let \( \pi^N(\theta) \) the reference prior.

- Parameterize in terms of \( \xi = g(\theta) \) such that \( \pi^N(\xi) = 1 \). Remarkably \( \xi \) behaves, asymptotically, as location parameter (Bernardo 2005).

- Apply the definition of DB priors for location parameters.

- Obtain the prior in the original parameterization (apply the transformation of variables formula).
The general case (without nuisance par.): construction

\[ M_1 : f_1(y) = f(y | \theta_0) \quad \text{vs.} \quad M_2 : f_2(y | \theta) = f(y | \theta), \]

with \( \theta \) a \( k \)-dimensional parameter. Let \( \pi^N(\theta) \) the reference prior.

- Parameterize in terms of \( \xi = g(\theta) \) such that \( \pi^N(\xi) = 1 \). Remarkably \( \xi \) behaves, asymptotically, as location parameter (Bernardo 2005).
- Apply the definition of DB priors for location parameters.
- Obtain the prior in the original parameterization (apply the transformation of variables formula).
General definition

\[ \pi_D(\theta) \propto h q(D[\theta, \theta_0]) \pi_N(\theta). \]
Of course, all these steps are not at all needed. By construction, we have arrived at the general definition:

$$\pi^D(\theta) \propto h_q(D[\theta, \theta_0])\pi^N(\theta).$$
Properties: The DB priors in words

- They capture the information contained into the problem, making use of the distance between the models.
Properties: The DB priors in words

- They capture the information contained into the problem, making use of the distance between the models.
- The information contained is equivalent to the information in a sample of unitary size.
Properties: The DB priors in words

- They capture the information contained into the problem, making use of the *distance* between the models.
- The information contained is equivalent to the information in a sample of unitary size.
- DB priors generalize the work of Jeffreys (1961) and Zellner and Siow (1980).
Properties: More theoretically

- If $\theta$ is a location parameter then $\pi^D$ approximately behaves as a Student centered at $\theta_0$ and scaled by the Fisher information matrix evaluated at $\theta_0$. 

- $\pi^D$ is invariant under reparameterizations.

- $\pi^D$ satisfies the 'sufficiency principle' (e.g. do not change if the problem is reduced via sufficient statistics).

For the example analyzed it is consistent in information.
Properties: More theoretically

- If $\theta$ is a location parameter then $\pi^D$ approximately behaves as a Student centered at $\theta_0$ and scaled by the Fisher information matrix evaluated at $\theta_0$.
- $\pi^D$ is invariant under reparameterizations.
Properties: More theoretically

- If $\theta$ is a location parameter then $\pi^D$ approximately behaves as a Student centered at $\theta_0$ and scaled by the Fisher information matrix evaluated at $\theta_0$.
- $\pi^D$ is invariant under reparameterizations.
- $\pi^D$ satisfies the ‘sufficiency principle’ (e.g. do not change if the problem is reduced via sufficient statistics).
Properties: More theoretically

- If $\theta$ is a location parameter then $\pi^D$ approximately behaves as a Student centered at $\theta_0$ and scaled by the Fisher information matrix evaluated at $\theta_0$.
- $\pi^D$ is invariant under reparameterizations.
- $\pi^D$ satisfies the ‘sufficiency principle’ (e.g. do not change if the problem is reduced via sufficient statistics).
- For the example analyzed it is consistent in information
Properties: More theoretically

- If $\theta$ is a location parameter then $\pi^D$ approximately behaves as a Student centered at $\theta_0$ and scaled by the Fisher information matrix evaluated at $\theta_0$.
- $\pi^D$ is invariant under reparameterizations.
- $\pi^D$ satisfies the ‘sufficiency principle’ (e.g. do not change if the problem is reduced via sufficient statistics).
- For the example analyzed it is consistent in information
- The (DB) Bayes factor can be expressed as

$$B_{21}^D = B_{21}^N \times \text{Correction factor (FC)},$$

where FC is a expectation with respect to the posterior distribution.
The choice of $q$

Clearly, $\pi^D$ depends on the election of $q$, a parameter which needs to be assigned. This parameter regulates the heaviness of the tails. Our proposal is

$$q = q_0 + 0.5,$$

with

$$q_0 = \inf\{q > 0 : \pi^D(\theta) \text{ is proper}\}.$$

This reproduces the Jeffreys-Zellner-Siow priors in the normal scenario.
### DB priors in practice

**Leaving aside formal criteria...**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Standard type</th>
<th>Improper likelihood</th>
<th>Irregular models</th>
</tr>
</thead>
</table>
Leaving aside formal criteria...

the best way of knowing if a proposal is reasonable is studying its
behaviour in scenarios, representative of the practice of statistics.
DB priors in practice

Leaving aside formal criteria...
the best way of knowing if a proposal is reasonable is studying its behaviour in scenarios, representative of the practice of statistics.

Scenario

- Standard type
- Improper likelihood
- Irregular models
The problem with nuisance parameters:
some few words

\[ M_1 : f_1(y \mid \nu_1) = f(y \mid \theta_0, \nu) \quad \text{vs.} \quad M_2 : f_2(y \mid \theta, \nu^*) = f(y \mid \theta, \nu^*), \]
The problem with nuisance parameters: some few words

\[ M_1 : f_1(y \mid \nu_1) = f(y \mid \theta_0, \nu) \text{ vs. } M_2 : f_2(y \mid \theta, \nu^*) = f(y \mid \theta, \nu^*), \]

Assume that \( \theta \) and \( \nu^* \) are orthogonal.
The problem with nuisance parameters: some few words

\[ M_1 : f_1(y \mid \nu_1) = f(y \mid \theta_0, \nu) \text{ vs. } M_2 : f_2(y \mid \theta, \nu^*) = f(y \mid \theta, \nu^*), \]

- Assume that \( \theta \) and \( \nu^* \) are orthogonal.
- Use \( \pi_1(\nu) = \pi_1^N(\nu) \) and,
The problem with nuisance parameters: some few words

\[ M_1 : f_1(y \mid \nu_1) = f(y \mid \theta_0, \nu) \text{ vs. } M_2 : f_2(y \mid \theta, \nu^*) = f(y \mid \theta, \nu^*), \]

- Assume that \( \theta \) and \( \nu^* \) are orthogonal.
- Use \( \pi_1(\nu) = \pi^N_1(\nu) \) and,
- \( \pi_2(\theta, \nu^*) = \pi^N_1(\nu^*)\pi^D(\theta \mid \nu^*), \) where

\[
\pi^D(\theta \mid \nu) \propto h_q\left(D[(\theta, \theta_0) \mid \nu]\right) \pi^N(\theta \mid \nu) .
\]
DB priors have a simple form and are easy to explain!
To be highlighted...

- DB priors have a simple form and are easy to explain!
- They have appealing theoretical properties and in standard scenarios DB priors produce similar answers to other well established methods.
To be highlighted...

- DB priors have a simple form and are easy to explain!
- They have appealing theoretical properties and in standard scenarios DB priors produce similar answers to other well established methods.
- In complex scenarios, where the other proposals fail, DB priors are well defined and produce quite reasonable results.
DB priors have a simple form and are easy to explain!

They have appealing theoretical properties and in standard scenarios DB priors produce similar answers to other well established methods.

In complex scenarios, where the other proposals fail, DB priors are well defined and produce quite reasonable results.

DB priors are quite general and, we hope, can be considered a reasonable extension of Jeffreys’ ideas on testing.
Full technical details available at:

Journal of the Royal Statistical Society

70, Part 5, pp. 981–1003

Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing

M. J. Bayarri
University of Valencia, Spain

and G. García-Donato
University of Castilla-La Mancha, Albacete, Spain

[Received May 2007. Revised March 2008]

Summary. We introduce objective proper prior distributions for hypothesis testing and model selection based on measures of divergence between the competing models; we call them divergence-based (DB) priors. DB priors have simple forms and desirable properties like information (finite sample) consistency and are often similar to other existing proposals like intrinsic priors. Moreover, in normal linear model scenarios, they reproduce the Jeffreys–Zellner–Siow priors exactly. Most importantly, in challenging scenarios such as irregular models and mixture models, DB priors are well defined and very reasonable, whereas alternative proposals are not. We derive approximations to the DB priors as well as Markov chain Monte Carlo and asymptotic expressions for the associated Bayes factors.

Keywords: Bayes factors; Information consistency; Intrinsic priors; Irregular models; Kullback–Leibler divergence; Mixture models

gonzalo garcía-Donato and susie Bayarri () DB priors May 22, 2009 40 / 46
Full technical details available at:

Journal of the Royal Statistical Society

70, Part 5, pp. 981–1003

Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing

M. J. Bayarri

University of Valencia, Spain

and G. Garcia-Donato

University of Castilla-La Mancha, Albacete, Spain

[Received May 2007. Revised March 2008]

Summary. We introduce objective proper prior distributions for hypothesis testing and model selection based on measures of divergence between the competing models; we call them divergence-based (DB) priors. DB priors have simple forms and desirable properties like information (finite sample) consistency and are often similar to the existing proposals like intrinsic priors. Moreover, in normal linear models, they reproduce the Jeffreys-Zellner-Siow priors exactly. Most importantly, in challenging scenarios, such as irregular models and mixture models, DB priors are well defined and very reasonable, whereas alternative proposals are not. We derive approximations to the DB priors as well as Markov chain Monte Carlo and asymptotic expressions for the associated Bayes factors.

Keywords: Bayes factors; Information consistency; Intrinsic priors; Irregular models; Kullback-Leibler divergence; Mixture models
Simple example: credible intervals and testing

Suppose:
- From a sample $y_i \sim N(\mu, 1)$, of size $n = 50$ I get $\bar{y} = 0.25$ from which I derive the 95\% credible interval using $\pi(\mu) = 1$. 

I am interested in knowing if $H_0: \mu = 0$ = 'no effect' (although I did not explicitly say).
From the interval above, I would decide Bayesianly, to reject $H_0$ (say has a probability which is less than 0.05). But $\text{Prob}(H_0 | y) = 0.66$.

Come back!
Simple example: credible intervals and testing

Suppose:

- From a sample $y_i \sim N(\mu, 1)$, of size $n = 50$ I get $\bar{y} = 0.25$ from which I derive the 95% credible interval using $\pi(\mu) = 1$.
- I obtain CI$=[0.11,0.39]$. 
Simple example: credible intervals and testing

Suppose:
- From a sample $y_i \sim N(\mu, 1)$, of size $n = 50$ I get $\bar{y} = 0.25$ from which I derive the 95% credible interval using $\pi(\mu) = 1$.
- I obtain CI=[0.11, 0.39].
- I am interested in knowing if $H : \mu = 0 = \text{‘no effect’}$ (although I did not explicitly say).
Simple example: credible intervals and testing

Suppose:

- From a sample $y_i \sim N(\mu, 1)$, of size $n = 50$ I get $\bar{y} = 0.25$ from which I derive the 95% credible interval using $\pi(\mu) = 1$.
- I obtain $\text{CI} = [0.11, 0.39]$.
- I am interested in knowing if $H: \mu = 0 = \text{‘no effect’}$ (although I did not explicitly say).
- From the interval above, I would decide Bayesianly, to reject $H$ (say has a probability which is less than 0.05). But
Suppose:

- From a sample \( y_i \sim N(\mu, 1) \), of size \( n = 50 \) I get \( \bar{y} = 0.25 \) from which I derive the 95\% credible interval using \( \pi(\mu) = 1 \).
- I obtain CI=[0.11,0.39].
- I am interested in knowing if \( H : \mu = 0 \) = ‘no effect’ (although I did not explicitly say).
- From the interval above, I would decide Bayesianly, to reject \( H \) (say has a probability which is less than 0.05). But

\[
\text{Prob}(H \mid y) = 0.66
\]
Standard scenarios

- In linear models, the DB priors coincide with the proposals of Zellner and Siow (1980, 1984).
PART IV. Divergence based priors

Standard scenarios

- In linear models, the DB priors coincide with the proposals of Zellner and Siow (1980, 1984).

- In other traditional problems (Bernoulli, exponential, Poisson or Normal), DB priors produce Bayes factors very similar to the intrinsic.
PART IV. Divergence based priors

Standard scenarios

• In linear models, the DB priors coincide with the proposals of Zellner and Siow (1980, 1984).

• In other traditional problems (Bernoulli, exponential, Poisson or Normal), DB priors produce Bayes factors very similar to the intrinsic.

\[ H_1 : (\mu, \sigma) = (0, 1) \text{ vs. } H_2 : (\mu, \sigma) \neq (0, 1) \]
Scenario with improper likelihood

Let $f(y \mid \theta, p) = p \, N(y \mid 0, 1) + (1 - p) \, N(y \mid \theta, 1)$, \,(p \text{ known})

The problem: $H_1 : \theta = 0$ vs. $H_2 : \theta \neq 0$. 
Scenario with improper likelihood

Let \( f(y | \theta, p) = p \, N(y | 0, 1) + (1 - p) \, N(y | \theta, 1), \) \((p\) known) \(\)

The problem: \( H_1 : \theta = 0 \) vs. \( H_2 : \theta \neq 0. \)

- Intrinsic Bayes factors do not exist. Berger and Pericchi (2001) ‘recommend’ using \( \pi_2^{BP}(\theta) = Ca(\theta|0, 1). \)
Scenario with improper likelihood

Let \( f(y \mid \theta, p) = p N(y \mid 0, 1) + (1 - p) N(y \mid \theta, 1), \) \((p \text{ known})\)

The problem: \( H_1 : \theta = 0 \) vs. \( H_2 : \theta \neq 0. \)

- Intrinsic Bayes factors do not exist. Berger and Pericchi (2001) ‘recommend’ using \( \pi_2^{BP}(\theta) = Ca(\theta|0, 1). \)
- The DB prior does not have a closed form, although a good approximation is

\[
\pi_2^D(\theta) \approx Ca(\theta|0, 1/(1 - p)).
\]
Comparisons

$\pi^D$ (solid line), $Ca(0, 1/1 - p)$ (dashed) and $\pi_{BP}^{2} (\mu) = Ca(\mu|0, 1)$ (dots).
Scenario with irregular models

\[ f(y \mid \theta) = \exp\{- (y - \theta)\}, \quad y > \theta, \]
and the problem \( H_1 : \theta = \theta_0 \) vs. \( H_2 : \theta > \theta_0 \).
PART IV. Divergence based priors

**Scenario with irregular models**

Sea

\[ f(y | \theta) = \exp\{-\(y - \theta\)\}, \quad y > \theta, \]

and the problem \( H_1 : \theta = \theta_0 \) vs. \( H_2 : \theta > \theta_0 \).

- These models pose in serious problems to the intrinsic priors. They conjecture that it is

\[ \pi_{2I}^A(\theta) = (-e^{\theta-\theta_0} \log(1 - e^{\theta_0-\theta}) - 1), \quad \theta > \theta_0. \]
PART IV. Divergence based priors

Scenario with irregular models

Sea

\[ f(y \mid \theta) = \exp\{-(y - \theta)\}, \quad y > \theta, \]

and the problem \( H_1 : \theta = \theta_0 \) vs. \( H_2 : \theta > \theta_0 \).

- These models pose in serious problems to the intrinsic priors. They conjecture that it is

\[ \pi_{2}^{AI}(\theta) = \left( -e^{\theta - \theta_0} \log(1 - e^{\theta_0 - \theta}) - 1 \right), \quad \theta > \theta_0. \]

- DB prior is

\[ \pi_{2}^{D}(\theta) = \left( 1 + 2(\theta - \theta_0) \right)^{-3/2}, \quad \theta > \theta_0. \]
Comparisons ($\theta_0 = 0$)

For, $\pi^D_2$ (solid line) y $\pi^{AI}_2$ (dashed line)