Criteria for objective Bayesian model choice

Gonzalo García-Donato (UCLM)

Collaborators: MJ Bayarri[1], J Berger[2], A Forte[3]

Universidad de Castilla-La Mancha (Spain)
[1] Universidad de Valencia (Spain), [2] Duke University (USA), [3] Universidad Jaume I (Spain)

Madrid - November 2011

Estimation vs. Model selection

An experiment with outcome Y is of interest:

Estimation problems

Statistical model for Y is assumed known.

Estimation vs. Model selection

An experiment with outcome Y is of interest:

Estimation problems

Statistical model for Y is assumed known.

Model selection (MS) problems

The uncertainty about which model provides a better explanation for Y is explicitly considered.

Estimation vs. Model selection

An experiment with outcome Y is of interest:

Estimation problems

Statistical model for Y is assumed known.

Model selection (MS) problems

The uncertainty about which model provides a better explanation for Y is explicitly considered.

Key features of objective Bayesian MS, based on Bayes factors:

Estimation vs. Model selection

An experiment with outcome Y is of interest:

Estimation problems

Statistical model for Y is assumed known.

Model selection (MS) problems

The uncertainty about which model provides a better explanation for Y is explicitly considered.

Key features of objective Bayesian MS, based on Bayes factors:

- Results are highly sensitive to the choice of priors,

Estimation vs. Model selection

An experiment with outcome Y is of interest:

Estimation problems

Statistical model for Y is assumed known.

Model selection (MS) problems

The uncertainty about which model provides a better explanation for Y is explicitly considered.

Key features of objective Bayesian MS, based on Bayes factors:

- Results are highly sensitive to the choice of priors,
- sensitivity does not vanish as n grows (unlike the estimation scenario),

Estimation vs. Model selection

An experiment with outcome Y is of interest:

Estimation problems

Statistical model for Y is assumed known.

Model selection (MS) problems

The uncertainty about which model provides a better explanation for Y is explicitly considered.

Key features of objective Bayesian MS, based on Bayes factors:

- Results are highly sensitive to the choice of priors,
- sensitivity does not vanish as n grows (unlike the estimation scenario),
- improper priors cannot, in general, be used

Estimation vs. Model selection

An experiment with outcome Y is of interest:

Estimation problems

Statistical model for Y is assumed known.

Model selection (MS) problems

The uncertainty about which model provides a better explanation for Y is explicitly considered.

Key features of objective Bayesian MS, based on Bayes factors:

- Results are highly sensitive to the choice of priors,
- sensitivity does not vanish as n grows (unlike the estimation scenario),
- improper priors cannot, in general, be used
- which prior to be used is still an open question.

There have been...

...many efforts, over more than 30 years, to develop convincing objective priors for MS. A number of such proposals:

- the Intrinsic priors (Berger and Pericchi 1996; Moreno *et al.* 1998; O'Hagan 1997),

There have been...

...many efforts, over more than 30 years, to develop convincing objective priors for MS. A number of such proposals:

- the Intrinsic priors (Berger and Pericchi 1996; Moreno *et al.* 1998; O'Hagan 1997),
- the Expected posterior priors (Pérez and Berger 2002),

There have been...

...many efforts, over more than 30 years, to develop convincing objective priors for MS. A number of such proposals:

- the Intrinsic priors (Berger and Pericchi 1996; Moreno *et al.* 1998; O'Hagan 1997),
- the Expected posterior priors (Pérez and Berger 2002),
- the Integral priors (Cano *et al.* 2008),

There have been...

...many efforts, over more than 30 years, to develop convincing objective priors for MS. A number of such proposals:

- the Intrinsic priors (Berger and Pericchi 1996; Moreno *et al.* 1998; O'Hagan 1997),
- the Expected posterior priors (Pérez and Berger 2002),
- the Integral priors (Cano *et al.* 2008),
- the Divergence based priors (Bayarri and García-Donato 2008).

There have been...

...many efforts, over more than 30 years, to develop convincing objective priors for MS. A number of such proposals:

- the Intrinsic priors (Berger and Pericchi 1996; Moreno *et al.* 1998; O'Hagan 1997),
- the Expected posterior priors (Pérez and Berger 2002),
- the Integral priors (Cano *et al.* 2008),
- the Divergence based priors (Bayarri and García-Donato 2008).

Diversity is good, but up to a certain level!

We don't need another...prior

This lack of progress in reaching consensus resulted in our approaching the problem from a different direction: *is it possible a constructive minimum agreement?*

We don't need another...prior

This lack of progress in reaching consensus resulted in our approaching the problem from a different direction: *is it possible a constructive minimum agreement?*

Main motivation

Compiling+formalizing+completing the different criteria that have been deemed essential for MS priors, and seeing if these criteria can essentially determine the priors.

The problem

We observe a vector $\mathbf{y} \sim f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ of size n. The competing models are

$$M_0 : f_0(\mathbf{y} \mid \boldsymbol{\alpha}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_0), \quad M_1 : f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

for certain $\boldsymbol{\beta}_0$.

The problem

We observe a vector $\mathbf{y} \sim f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ of size n. The competing models are

$$M_0 : f_0(\mathbf{y} \mid \boldsymbol{\alpha}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_0), \qquad M_1 : f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

for certain $\boldsymbol{\beta}_0$.In testing notation

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0, \qquad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

The problem

We observe a vector $\mathbf{y} \sim f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ of size n. The competing models are

$$M_0 : f_0(\mathbf{y} \mid \boldsymbol{\alpha}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_0), \quad M_1 : f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

for certain $\boldsymbol{\beta}_0$.In testing notation

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0, \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

We base our response to the problem on the Bayes factor B_{10}: the ratio of prior marginals wrt the priors $\pi_1(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\pi_0(\boldsymbol{\alpha})$.

The problem

We observe a vector $\mathbf{y} \sim f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ of size n. The competing models are

$$M_0 : f_0(\mathbf{y} \mid \boldsymbol{\alpha}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}_0), \quad M_1 : f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

for certain $\boldsymbol{\beta}_0$.In testing notation

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0, \quad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

We base our response to the problem on the Bayes factor B_{10}: the ratio of prior marginals wrt the priors $\pi_1(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\pi_0(\boldsymbol{\alpha})$.
Without loss of generality we express

$$\pi_1(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \pi_1(\boldsymbol{\alpha})\pi_1(\boldsymbol{\beta} \mid \boldsymbol{\alpha}).$$

A needed consideration

Due to the nature of H_0 this problem is known in the literature as testing a "precise" or "punctual" hypothesis, which we interpret as the more real of $H_0^R : \boldsymbol{\beta} \approx \boldsymbol{\beta}^0$.

A needed consideration

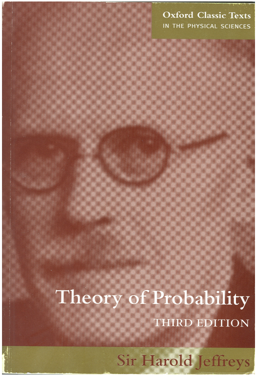
Due to the nature of H_0 this problem is known in the literature as testing a “precise” or “punctual” hypothesis, which we interpret as the more real of $H_0^R : \boldsymbol{\beta} \approx \boldsymbol{\beta}^0$.

Conditions under which testing H_0 is a valid approximation for H_0^R have been studied by Berger and Delampady (1987), Gómez-Villegas and Sánchez-Manzano (1992) and Verdinelly and Wasserman (1996).

Jeffreys' desiderata

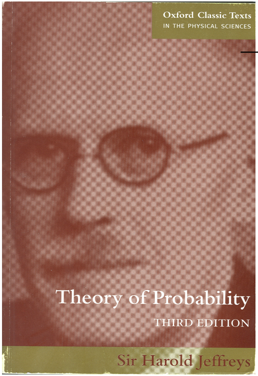
Jeffreys' desiderata

The approach of using criteria leading to priors is reminiscent of the Jeffreys (1961) approach to testing, wherein certain testing desiderata were presented and testing priors were derived from them.



Jeffreys' desiderata

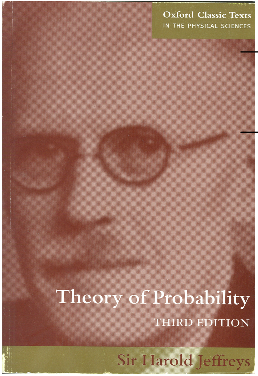
The approach of using criteria leading to priors is reminiscent of the Jeffreys (1961) approach to testing, wherein certain testing desiderata were presented and testing priors were derived from them.



$\longrightarrow$ see Robert *et al* (2009) for a comprehensive and modern review of Jeffreys' book.

Jeffreys' desiderata

The approach of using criteria leading to priors is reminiscent of the Jeffreys (1961) approach to testing, wherein certain testing desiderata were presented and testing priors were derived from them.

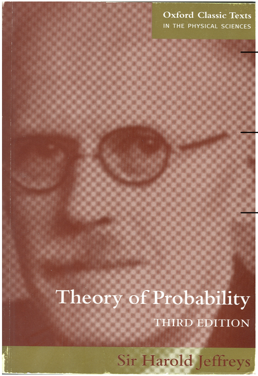


→ see Robert *et al* (2009) for a comprehensive and modern review of Jeffreys' book.

→ These arguments are often called Jeffreys' desiderata

Jeffreys' desiderata

The approach of using criteria leading to priors is reminiscent of the Jeffreys (1961) approach to testing, wherein certain testing desiderata were presented and testing priors were derived from them.



→ see Robert *et al* (2009) for a comprehensive and modern review of Jeffreys' book.

→ These arguments are often called Jeffreys' desiderata

→ These and related ideas have been repeatedly used to evaluate-guide-justify development of objective MS priors.

Jeffreys' desiderata (cont')

General problems

-
-

Testing whether β a normal mean is zero (σ unknown)

-
-

Jeffreys' desiderata (cont')

General problems

- If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are "orthogonal", then one can take $\pi_0(\boldsymbol{\alpha}) = \pi_1(\boldsymbol{\alpha})$. Due to its small impact on the Bayes factors he recommended an objective estimation prior.
-

Testing whether β a normal mean is zero (σ unknown)

-
-

Jeffreys' desiderata (cont')

General problems

- If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are "orthogonal", then one can take $\pi_0(\boldsymbol{\alpha}) = \pi_1(\boldsymbol{\alpha})$. Due to its small impact on the Bayes factors he recommended an objective estimation prior.
- The conditional prior $\pi_1(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$ should be proper and have heavy tails (he noted that this condition is closely related with what is nowadays known as information consistency).

Testing whether β a normal mean is zero (σ unknown)

-
-

Jeffreys' desiderata (cont')

General problems

- If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are "orthogonal", then one can take $\pi_0(\boldsymbol{\alpha}) = \pi_1(\boldsymbol{\alpha})$. Due to its small impact on the Bayes factors he recommended an objective estimation prior.
- The conditional prior $\pi_1(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$ should be proper and have heavy tails (he noted that this condition is closely related with what is nowadays known as information consistency).

Testing whether β a normal mean is zero (σ unknown)

- The conditional prior $\pi_1(\beta \mid \sigma)$ should be centered at zero and scaled by σ (from "considerations of similarity"),
-

Jeffreys' desiderata (cont')

General problems

- If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are "orthogonal", then one can take $\pi_0(\boldsymbol{\alpha}) = \pi_1(\boldsymbol{\alpha})$. Due to its small impact on the Bayes factors he recommended an objective estimation prior.
- The conditional prior $\pi_1(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$ should be proper and have heavy tails (he noted that this condition is closely related with what is nowadays known as information consistency).

Testing whether β a normal mean is zero (σ unknown)

- The conditional prior $\pi_1(\beta \mid \sigma)$ should be centered at zero and scaled by σ (from "considerations of similarity"),
- For $n = 1$ the Bayes factor should be one (since a single observation allows no discrimination between the two models).

Jeffreys' desiderata are

intuitively sensible but ad-hoc arguments: difficult to justify.

Jeffreys' desiderata are

intuitively sensible but ad-hoc arguments: difficult to justify.

We try to formalize the most general and compelling of the various criteria that have been suggested, together with a new criterion.

Jeffreys' desiderata are

intuitively sensible but ad-hoc arguments: difficult to justify.

We try to formalize the most general and compelling of the various criteria that have been suggested, together with a new criterion.

The resulting criteria can be organized into four blocks:

- I. Basic criteria,
- II. Consistency criteria,
- III. Predictive matching criteria,
- IV. Invariance criteria.

Jeffreys' desiderata are

intuitively sensible but ad-hoc arguments: difficult to justify.

We try to formalize the most general and compelling of the various criteria that have been suggested, together with a new criterion.

The resulting criteria can be organized into four blocks:

- I. Basic criteria,
- II. Consistency criteria,
- III. Predictive matching criteria,
- IV. Invariance criteria.

Few modern references that are relevant to the development of such criteria

Fernández *et al.* (2001); Berger and Pericchi (2001); Berger *et al.* (2003); Liang *et al.* (2008); Moreno *et al.* (2009); Casella *et al.* (2009)

I. Basic criteria

In words

Formally

I. Basic criteria

In words

The evidence provided by a MS procedure cannot depend on arbitrary constants

Formally

I. Basic criteria

In words

The evidence provided by a MS procedure cannot depend on arbitrary constants

Formally

Basic criterion

The conditional prior $\pi_1(\beta \mid \alpha)$ must be proper (integrating to one) and cannot be arbitrarily vague.

II.Consistency criteria

In words

Formally

II.Consistency criteria

In words

If the evidence in favor of one of the entertained models grows to infinite, the evidence provided by the MS procedure should 'grow' accordingly.

Formally

II.Consistency criteria

In words

If the evidence in favor of one of the entertained models grows to infinite, the evidence provided by the MS procedure should 'grow' accordingly.

Formally

- MS consistency criterion

If data **y** have been generated by M_i, then the posterior probability of M_i should converge in probability to 1 as $n \to \infty$.

II.Consistency criteria

In words

If the evidence in favor of one of the entertained models grows to infinite, the evidence provided by the MS procedure should 'grow' accordingly.

Formally

- MS consistency criterion

If data **y** have been generated by M_i, then the posterior probability of M_i should converge in probability to 1 as $n \to \infty$.

- Information consistency criterion

If $\Lambda_{10} \to \infty$ then B_{10} should also $\to \infty$.

Where Λ_{10} is the observed likelihood ratio for M_1 compared to M_0:

III. Predictive matching criteria

In words

Formally

III. Predictive matching criteria

In words

When the information in the data is extremely weak, the MS procedure should not be conclusive.

Formally

III. Predictive matching criteria

In words

When the information in the data is extremely weak, the MS procedure should not be conclusive.

Formally

Predictive matching criterion

• For samples $\mathbf{y}^*$ of 'minimal size', in comparing M_0 with M_1, one should have model selection priors such that $m_0(\mathbf{y}^*)$ and $m_1(\mathbf{y}^*)$ are close.

III. Predictive matching criteria

In words

When the information in the data is extremely weak, the MS procedure should not be conclusive.

Formally

Predictive matching criterion

- For samples $\mathbf{y}^*$ of 'minimal size', in comparing M_0 with M_1, one should have model selection priors such that $m_0(\mathbf{y}^*)$ and $m_1(\mathbf{y}^*)$ are close.
- Optimal is exact predictive matching: $m_0(\mathbf{y}^*) = m_1(\mathbf{y}^*)$.

Predictive matching

- Asking the priors for being appropriately ‘matched’ is a crucial aspect, specially when the models under comparison differ much in dimensionality.

Predictive matching

- Asking the priors for being appropriately 'matched' is a crucial aspect, specially when the models under comparison differ much in dimensionality.
- In Berger and Pericchi (2001), minimal sample size n^* was defined as the smallest sample size for which

$$0 < m_i \ (\mathbf{y}^*) < \infty,$$

for all i and when *objective estimation* priors π_i^N are used.

Predictive matching

- Asking the priors for being appropriately 'matched' is a crucial aspect, specially when the models under comparison differ much in dimensionality.
- In Berger and Pericchi (2001), minimal sample size n^* was defined as the smallest sample size for which

$$0 < m_i{}^N(\mathbf{y}^*) < \infty,$$

for all i and when *objective estimation* priors π_i^N are used.

Predictive matching

- Asking the priors for being appropriately 'matched' is a crucial aspect, specially when the models under comparison differ much in dimensionality.
- In Berger and Pericchi (2001), minimal sample size n^* was defined as the smallest sample size for which

$$0 < m_i{}^{N}(\mathbf{y}^*) < \infty,$$

for all i and when *objective estimation* priors π_i^N are used.
- We propose a different definition of *minimal size*.

Predictive matching: minimal size

We think that, in general, minimal sample size should be defined relative to the *model selection* priors being utilized:

Predictive matching: minimal size

We think that, in general, minimal sample size should be defined relative to the *model selection* priors being utilized:

Definition of Minimal training sample $\mathbf{y}^*$

for $\{M_1, \pi_1\}$ is a sample of minimal size $n^* \geq 1$ with

$$0 < m_i(\mathbf{y}^*) < \infty.$$

Crucial consequences:

Predictive matching: minimal size

We think that, in general, minimal sample size should be defined relative to the *model selection* priors being utilized:

Definition of Minimal training sample $\mathbf{y}^*$

for $\{M_1, \pi_1\}$ is a sample of minimal size $n^* \geq 1$ with

$$0 < m_i(\mathbf{y}^*) < \infty.$$

Crucial consequences:

- Because of *Basic criteria* this new n^* is smaller than the B&P01 n^*: the predictive matching criteria becomes a weaker condition.

Predictive matching: minimal size

We think that, in general, minimal sample size should be defined relative to the *model selection* priors being utilized:

Definition of Minimal training sample $\mathbf{y}^*$

for $\{M_1, \pi_1\}$ is a sample of minimal size $n^* \geq 1$ with

$$0 < m_i(\mathbf{y}^*) < \infty.$$

Crucial consequences:

- Because of *Basic criteria* this new n^* is smaller than the B&P01 n^*: the predictive matching criteria becomes a weaker condition.
- In problems with more than 2 competing models (e.g variable selection) the concept of minimal size is almost insensitive to the dimension of the largest model.

IV. Invariance criteria

In words

Formally

IV. Invariance criteria

In words

If models posses an invariance structure, this should be preserved after marginalization

Formally

IV. Invariance criteria

In words

If models posses an invariance structure, this should be preserved after marginalization

Formally

Invariance criterion

If M_0 and M_1 are invariant under certain group of transformations G_0, then the conditional distribution, $\pi_1(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$, should be chosen in such a way that the conditional marginal distribution

$$f_1^I(\mathbf{y} \mid \boldsymbol{\alpha}) = \int f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})\, \pi_1(\boldsymbol{\beta} \mid \boldsymbol{\alpha})\, d\boldsymbol{\beta},$$

is also invariant under G_0.

Invariance criterion: first important consequence (In case of existence of such structure)

- Note: G_0 is a group of transformations relevant for the null model M_0.

Hence

Invariance criterion: first important consequence (In case of existence of such structure)

- Note: G_0 is a group of transformations relevant for the null model M_0.

Hence

invariance criterion can be understood as a formalization of the Jeffreys' requirement that the prior for a non-null parameter should be "centered at the simple model" (will become apparent in the examples).

Invariance criterion: second important consequence(In case of existence of such structure)

....it is about the tricky question of determining the priors for the common model parameters $\pi_0(\boldsymbol{\alpha})$ and $\pi_1(\boldsymbol{\alpha})$.

Invariance criterion: second important consequence(In case of existence of such structure)

....it is about the tricky question of determining the priors for the common model parameters $\pi_0(\boldsymbol{\alpha})$ and $\pi_1(\boldsymbol{\alpha})$.

With invariance criterion, the problem becomes transformed in one with competing models:

$$f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_0(\boldsymbol{\alpha}) \qquad \textit{vs} \quad f_1^I(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_1(\boldsymbol{\alpha})$$

with the *same dimension* and *sharing a common invariance structure*.

Invariance criterion: second important consequence(In case of existence of such structure)

....it is about the tricky question of determining the priors for the common model parameters $\pi_0(\boldsymbol{\alpha})$ and $\pi_1(\boldsymbol{\alpha})$.

With invariance criterion, the problem becomes transformed in one with competing models:

$$f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_0(\boldsymbol{\alpha}) \qquad \textit{vs} \;\; f_1^I(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_1(\boldsymbol{\alpha})$$

with the *same dimension* and *sharing a common invariance structure*.
In this situation choosing

Invariance criterion: second important consequence(In case of existence of such structure)

....it is about the tricky question of determining the priors for the common model parameters $\pi_0(\boldsymbol{\alpha})$ and $\pi_1(\boldsymbol{\alpha})$.

With invariance criterion, the problem becomes transformed in one with competing models:

$$f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_0(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha}) \;\; vs \;\; f_1^I(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_1(\boldsymbol{\alpha})$$

with the *same dimension* and *sharing a common invariance structure*.

In this situation choosing

Invariance criterion: second important consequence(In case of existence of such structure)

....it is about the tricky question of determining the priors for the common model parameters $\pi_0(\boldsymbol{\alpha})$ and $\pi_1(\boldsymbol{\alpha})$.

With invariance criterion, the problem becomes transformed in one with competing models:

$$f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_0(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha}) \;\; vs \;\; f_1^I(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_1(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha})$$

with the *same dimension* and *sharing a common invariance structure*. In this situation choosing

Invariance criterion: second important consequence(In case of existence of such structure)

....it is about the tricky question of determining the priors for the common model parameters $\pi_0(\boldsymbol{\alpha})$ and $\pi_1(\boldsymbol{\alpha})$.

With invariance criterion, the problem becomes transformed in one with competing models:

$$f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_0(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha}) \;\; vs \;\; f_1^I(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_1(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha})$$

with the *same dimension* and *sharing a common invariance structure*.

In this situation choosing where $\pi^H(\cdot)$ is the right-Haar density of G_0

Invariance criterion: second important consequence(In case of existence of such structure)

....it is about the tricky question of determining the priors for the common model parameters $\pi_0(\boldsymbol{\alpha})$ and $\pi_1(\boldsymbol{\alpha})$.

With invariance criterion, the problem becomes transformed in one with competing models:

$$f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_0(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha}) \;\; vs \;\; f_1^I(\mathbf{y} \mid \boldsymbol{\alpha}), \pi_1(\boldsymbol{\alpha}) = \pi^H(\boldsymbol{\alpha})$$

with the *same dimension* and *sharing a common invariance structure*.

In this situation choosing where $\pi^H(\cdot)$ is the right-Haar density of G_0

Berger *et al* (1998)

ensures, under commonly satisfied conditions, exact predictive matching.

Problem 1

Suppose **y** is an iid sample of a normal population with σ unknown and the hypotheses about the mean

$$H_0 : \mu = 0, \qquad H_1 : \mu \neq 0.$$

The priors $\pi_0(\sigma)$ and $\pi_1(\mu, \sigma) = \pi_1(\mu \mid \sigma)\pi_1(\sigma)$ needs to be assigned.

Problem 1

Suppose **y** is an iid sample of a normal population with σ unknown and the hypotheses about the mean

$$H_0 : \mu = 0, \qquad H_1 : \mu \neq 0.$$

The priors $\pi_0(\sigma)$ and $\pi_1(\mu, \sigma) = \pi_1(\mu \mid \sigma)\pi_1(\sigma)$ needs to be assigned.

Basic criterion: $\pi_1(\mu \mid \sigma)$ must be proper and not arbitrarily vague.

Invariance criterion

Note that M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = g\mathbf{y}$.

Invariance criterion

Note that M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = g\mathbf{y}$.

Result

$$\mathbf{y} \sim f_1^I(\mathbf{y} \mid \sigma) = \int f_1(\mathbf{y} \mid \sigma, \mu)\pi_1(\mu \mid \sigma) d\mu$$

is invariant under the action of G_0 if and only if $\pi_1(\mu \mid \sigma) = \frac{1}{\sigma}\, h(\frac{\mu}{\sigma})$.

Invariance criterion

Note that M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = g\mathbf{y}$.

Result

$$\mathbf{y} \sim f_1^I(\mathbf{y} \mid \sigma) = \int f_1(\mathbf{y} \mid \sigma, \mu)\pi_1(\mu \mid \sigma) d\mu$$

is invariant under the action of G_0 if and only if $\pi_1(\mu \mid \sigma) = \frac{1}{\sigma}\, h(\frac{\mu}{\sigma})$.

This result gives a characterization for choosing $\pi_1(\mu \mid \sigma)$

- scaled by σ,
- centered at zero (the null model).

Invariance criterion

Note that M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = g\mathbf{y}$.

Result

$$\mathbf{y} \sim f_1^I(\mathbf{y} \mid \sigma) = \int f_1(\mathbf{y} \mid \sigma, \mu)\pi_1(\mu \mid \sigma)d\mu$$

is invariant under the action of G_0 if and only if $\pi_1(\mu \mid \sigma) = \frac{1}{\sigma}\, h(\frac{\mu}{\sigma})$.

This result gives a characterization for choosing $\pi_1(\mu \mid \sigma)$

- scaled by σ,
- centered at zero (the null model).

or equivalently a characterization of Jeffreys' *considerations of similarity*.

Predictive matching

The minimal size (new definition) associated with

$$\{M_1, \pi_1(\mu, \sigma) = \sigma^{-1}\, h(\mu/\sigma)\pi_1(\sigma)\},$$

Predictive matching

The minimal size (new definition) associated with

$$\{M_1, \pi_1(\mu, \sigma) = \sigma^{-1}\, h(\mu/\sigma)\pi_1(\sigma)\}, \text{ is } n^* = 1.$$

Result

If in addition $\pi_0(\sigma) = \pi^H(\sigma)$ and $\pi_1(\sigma) = \pi^H(\sigma)$ where $\pi^H(\sigma) = 1/\sigma$ (ie the right-Haar measure for G_0) then the resulting MS procedure is exact predictive matching (under the weak assumption of even h).

Proof.

Jeffreys (1961) (a very ingenious change of variable), generalized by Berger *et al.* (1998) using group invariance theory. □

Predictive matching

Using

$$\pi_0(\sigma) = \sigma^{-1}, \;\; \pi_1(\mu, \sigma) = \sigma^{-1}\sigma^{-1}h(\mu/\sigma) \tag{1}$$

is of course the basis for Jeffreys' proposal.

Predictive matching

Using

$$\pi_0(\sigma) = \sigma^{-1}, \;\; \pi_1(\mu, \sigma) = \sigma^{-1}\sigma^{-1}h(\mu/\sigma) \tag{1}$$

is of course the basis for Jeffreys' proposal.

A subtle difference

Here (1) becomes justified (almost characterized) through invariance and predictive matching while Jeffreys justified its use on the grounds of orthogonality (here unneeded).

Predictive matching

Using

$$\pi_0(\sigma) = \sigma^{-1}, \;\; \pi_1(\mu, \sigma) = \sigma^{-1}\sigma^{-1}h(\mu/\sigma) \tag{1}$$

is of course the basis for Jeffreys' proposal.

A subtle difference

Here (1) becomes justified (almost characterized) through invariance and predictive matching while Jeffreys justified its use on the grounds of orthogonality (here unneeded).

- Consistency criterion

It is well known (e.g. Jeffreys (1961); Fernández *et al.* 2001; Liang *et al.* 2008) that, in this case, a density h with heavy tails (no moments) ensures consistency.

Problem 2

Suppose **y** is an iid sample of a normal population with μ unknown and the hypotheses about the standard deviation

$$H_0 : \sigma = \sigma_0, \qquad H_1 : \sigma \neq \sigma_0,$$

where σ_0 is certain positive number.
The priors $\pi_0(\mu)$ and $\pi_1(\mu, \sigma) = \pi_1(\sigma \mid \mu)\pi_1(\mu)$ needs to be assigned.

Problem 2

Suppose **y** is an iid sample of a normal population with μ unknown and the hypotheses about the standard deviation

$$H_0 : \sigma = \sigma_0, \qquad H_1 : \sigma \neq \sigma_0,$$

where σ_0 is certain positive number.
The priors $\pi_0(\mu)$ and $\pi_1(\mu, \sigma) = \pi_1(\sigma \mid \mu)\pi_1(\mu)$ needs to be assigned.

- Basic criterion: $\pi_1(\sigma \mid \mu)$ must be proper and not arbitrarily vague.

Invariance

In this case M_0 and M_1 are invariant under the group $G_0 = \{g \in \mathcal{R}\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = \mathbf{y} + g\mathbf{1}_n$.

Result

Invariance

In this case M_0 and M_1 are invariant under the group $G_0 = \{g \in \mathcal{R}\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = \mathbf{y} + g\mathbf{1}_n$.

Result

$$\mathbf{y} \sim f_1(\mathbf{y} \mid \mu) = \int f_1(\mathbf{y} \mid \sigma, \mu)\pi_1(\sigma \mid \mu)d\sigma$$

is invariant under the action of G_0 *(and hence the priors satisfy the criterion)* if and only if $\pi_1(\sigma \mid \mu) = h(\sigma)$.

Hence:

Invariance

In this case M_0 and M_1 are invariant under the group $G_0 = \{g \in \mathcal{R}\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = \mathbf{y} + g\mathbf{1}_n$.

Result

$$\mathbf{y} \sim f_1(\mathbf{y} \mid \mu) = \int f_1(\mathbf{y} \mid \sigma, \mu)\pi_1(\sigma \mid \mu) d\sigma$$

is invariant under the action of G_0 (and hence the priors satisfy the criterion) if and only if $\pi_1(\sigma \mid \mu) = h(\sigma)$.

Hence:$\pi_1(\sigma \mid \mu)$ must not depend on μ.

Predictive matching

Again the minimal size (new definition) associated with

$$\{M_1, \pi_1(\mu, \sigma) = h(\sigma)\pi_1(\mu)\}, \qquad \text{is} \qquad n^* = 1.$$

Result

Predictive matching

Again the minimal size (new definition) associated with

$$\{M_1, \pi_1(\mu, \sigma) = h(\sigma)\pi_1(\mu)\}, \qquad \text{is} \qquad n^* = 1.$$

Result

If we take $\pi_0(\mu) = \pi^H(\mu)$ and $\pi_1(\mu) = \pi^H(\mu)$ where $\pi^H(\mu) = 1$ (ie the right-Haar measure for G_0), then the resulting procedure is exact predictive matching.

Consistency criteria

It can be seen that the observed likelihood ratio $\Lambda_{10} \to \infty$ if and only if $n \geq 2$ and either $S \to \infty$ or $S \to 0$.

Result

Consistency criteria

It can be seen that the observed likelihood ratio $\Lambda_{10} \to \infty$ if and only if $n \geq 2$ and either $S \to \infty$ or $S \to 0$.

Result

- *If $S \to \infty$ then $B_{10} \to \infty$ (independently of h),*

Consistency criteria

It can be seen that the observed likelihood ratio $\Lambda_{10} \to \infty$ if and only if $n \geq 2$ and either $S \to \infty$ or $S \to 0$.

Result

- *If $S \to \infty$ then $B_{10} \to \infty$ (independently of h),*
- *If $S \to 0$ then $B_{10} \to \infty$ for all $n \geq 2$ if*

$$\int_0^\infty \sigma^{1/2}\, h(\sigma)\, d\sigma = \infty.$$

Consistency criteria

It can be seen that the observed likelihood ratio $\Lambda_{10} \to \infty$ if and only if $n \geq 2$ and either $S \to \infty$ or $S \to 0$.

Result

- *If $S \to \infty$ then $B_{10} \to \infty$ (independently of h),*
- *If $S \to 0$ then $B_{10} \to \infty$ for all $n \geq 2$ if*

$$\int_0^\infty \sigma^{1/2}\, h(\sigma)\, d\sigma = \infty.$$

Note: this is a stronger requirement than having no moments and is not met, for instance, by the conjugate prior.

Problem 3

Consider the Gamma density with mean μ and shape parameter α:

$$Ga(y \mid \alpha, \mu) = \left(\frac{\alpha}{\mu}\right)^{\alpha} \Gamma(\alpha)^{-1} \, y^{\alpha-1} \, e^{-y\alpha/\mu}.$$

Now suppose that $\mathbf{y}$ is an iid sample of a gamma population with mean μ unknown and the hypotheses about the shape parameter

$$H_0 : \alpha = \alpha_0, \qquad H_1 : \alpha \neq \alpha_0,$$

where α_0 is certain positive number.
The priors $\pi_0(\mu)$ and $\pi_1(\mu, \alpha) = \pi_1(\alpha \mid \mu)\pi_1(\mu)$ needs to be assigned.

Problem 3

Consider the Gamma density with mean μ and shape parameter α:

$$Ga(y \mid \alpha, \mu) = \left(\frac{\alpha}{\mu}\right)^{\alpha} \Gamma(\alpha)^{-1} y^{\alpha-1} e^{-y\alpha/\mu}.$$

Now suppose that $\mathbf{y}$ is an iid sample of a gamma population with mean μ unknown and the hypotheses about the shape parameter

$$H_0 : \alpha = \alpha_0, \qquad H_1 : \alpha \neq \alpha_0,$$

where α_0 is certain positive number.
The priors $\pi_0(\mu)$ and $\pi_1(\mu, \alpha) = \pi_1(\alpha \mid \mu)\pi_1(\mu)$ needs to be assigned.

- Basic criterion: $\pi_1(\alpha \mid \mu)$ must be proper and not arbitrarily vague.

Invariance

Here M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = g\mathbf{y}$.

Result

Invariance

Here M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = g\mathbf{y}$.

Result

$$\mathbf{y} \sim f_1(\mathbf{y} \mid \mu) = \int f_1(\mathbf{y} \mid \alpha, \mu)\pi_1(\alpha \mid \mu) d\alpha$$

is invariant under the action of G_0 (and hence the priors satisfy the criterion) if and only if

$$\pi_1(\alpha \mid \mu) = h(\alpha).$$

Hence:

Invariance

Here M_0 and M_1 are invariant under the group $G_0 = \{g \in (0, \infty)\}$ with action over $\mathbf{y}$ as $g(\mathbf{y}) = g\mathbf{y}$.

Result

$$\mathbf{y} \sim f_1(\mathbf{y} \mid \mu) = \int f_1(\mathbf{y} \mid \alpha, \mu)\pi_1(\alpha \mid \mu) d\alpha$$

is invariant under the action of G_0 (and hence the priors satisfy the criterion) if and only if

$$\pi_1(\alpha \mid \mu) = h(\alpha).$$

Hence:$\pi_1(\alpha \mid \mu)$ must not depend on μ.

Predictive matching criterion

The minimal size (new definition) associated with

$$\{M_1, \pi_1(\mu,\alpha) = h(\alpha)\pi_1(\mu)\}, \qquad \text{is} \qquad n^* = 1.$$

Result

Predictive matching criterion

The minimal size (new definition) associated with

$$\{M_1, \pi_1(\mu, \alpha) = h(\alpha)\pi_1(\mu)\}, \qquad \text{is} \qquad n^* = 1.$$

Result

If we take $\pi_0(\mu) = \pi^H(\mu)$ and $\pi_1(\mu) = \pi^H(\mu)$ where $\pi^H(\mu) = 1/\mu$ is the right-Haar measure for G_0, then exact predictive matching criterion is satisfied.

Consistency criteria

In this case the observed likelihood ratio Λ_{10} has a more involved expression, $\Lambda_{10} = \Lambda_{10}(n, \bar{y}^g, \bar{y})$ where $\bar{y}^g$ is the geometric mean.

Consistency criteria

In this case the observed likelihood ratio Λ_{10} has a more involved expression, $\Lambda_{10} = \Lambda_{10}(n, \bar{y}^g, \bar{y})$ where $\bar{y}^g$ is the geometric mean.

Still it can be proved that if $\bar{y}^g/\bar{y} \to 1$ then $\Lambda_{10} \to \infty$ (more of such situations?).

Result

Consistency criteria

In this case the observed likelihood ratio Λ_{10} has a more involved expression, $\Lambda_{10} = \Lambda_{10}(n, \bar{y}^g, \bar{y})$ where $\bar{y}^g$ is the geometric mean.

Still it can be proved that if $\bar{y}^g/\bar{y} \to 1$ then $\Lambda_{10} \to \infty$ (more of such situations?).

Result

If $\bar{y}^g/\bar{y} \to 1$ then $B_{10} \to \infty$ if

$$\int_0^\infty \alpha^{1/2}\, h(\alpha)\, d\alpha = \infty$$

Priors that satisfy the criteria

- Pr1. $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ (μ is a normal mean):

$$\pi_0(\sigma) = \sigma^{-1}, \;\; \pi_1(\mu, \sigma) = \sigma^{-2} h(\mu/\sigma)$$

  with h proper (not vague), even and $\int x h(x)\, dx = \infty$

- Pr2. $H_0 : \sigma = \sigma_0$ vs. $H_1 : \sigma \neq \sigma_0$ (σ is a normal sd):

$$\pi_0(\mu) = 1, \;\; \pi_1(\mu, \sigma) = h(\sigma)$$

  with h proper (not vague) and $\int \sqrt{x} h(x)\, dx = \infty$

- Pr3. $H_0 : \alpha = \alpha_0$ vs. $H_1 : \alpha \neq \alpha_0$ (α is a gamma shape):

$$\pi_0(\mu) = 1, \;\; \pi_1(\mu, \alpha) = h(\alpha)$$

  with h proper (not vague) and $\int \sqrt{x} h(x)\, dx = \infty$

General definition

For the problem

$$M_0 : f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \quad M_1 : f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

Bayarri and García-Donato (2008) proposed the Divergence-Based priors:

$$\pi_1^D(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) \propto g_q\Big(D(\boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\alpha})\Big)\, \pi_1^N(\boldsymbol{\beta} \mid \boldsymbol{\alpha}),$$

where

- D is some 'distance' between f_1 and f_0,

General definition

For the problem

$$M_0 : f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \qquad M_1 : f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

Bayarri and García-Donato (2008) proposed the Divergence-Based priors:

$$\pi_1^D(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) \propto g_q\Big(D(\boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\alpha})\Big)\, \pi_1^N(\boldsymbol{\beta} \mid \boldsymbol{\alpha}),$$

where

- D is some 'distance' between f_1 and f_0,
- g_q is a real value decreasing function indexed by a parameter $q > 0$, and

General definition

For the problem

$$M_0 : f_0(\mathbf{y} \mid \boldsymbol{\alpha}), \quad M_1 : f_1(\mathbf{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}),$$

Bayarri and García-Donato (2008) proposed the Divergence-Based priors:

$$\pi_1^D(\boldsymbol{\beta} \mid \boldsymbol{\alpha}) \propto g_q\Big(D(\boldsymbol{\beta}, \boldsymbol{\beta}_0, \boldsymbol{\alpha})\Big)\, \pi_1^N(\boldsymbol{\beta} \mid \boldsymbol{\alpha}),$$

where

- D is some ‘distance’ between f_1 and f_0,
- g_q is a real value decreasing function indexed by a parameter $q > 0$, and
- $\pi_1^N(\boldsymbol{\beta} \mid \boldsymbol{\alpha})$ is an objective estimation prior (possibly improper).

DB priors: recommended ingredients

This definition defines a vast family of prior distributions (depending on D, h_q and π_1^N).

DB priors: recommended ingredients

This definition defines a vast family of prior distributions (depending on D, h_q and π_1^N).
Below the author's specific recommendations:

DB priors: recommended ingredients

This definition defines a vast family of prior distributions (depending on D, h_q and π_1^N).
Below the author's specific recommendations:

- $D =$ symmetrized Kullback-Leibler divergence divided by n,

DB priors: recommended ingredients

This definition defines a vast family of prior distributions (depending on D, h_q and π_1^N).
Below the author's specific recommendations:

- $D =$ symmetrized Kullback-Leibler divergence divided by n,
- $g_q(x) = (1 + x)^{-q}$ (has polynomial tails),

DB priors: recommended ingredients

This definition defines a vast family of prior distributions (depending on D, h_q and π_1^N).
Below the author's specific recommendations:

- $D =$ symmetrized Kullback-Leibler divergence divided by n,
- $g_q(x) = (1 + x)^{-q}$ (has polynomial tails),
- π_1^N the reference prior of Berger and Bernardo (1992),

DB priors: recommended ingredients

This definition defines a vast family of prior distributions (depending on D, h_q and π_1^N).
Below the author's specific recommendations:

- $D =$ symmetrized Kullback-Leibler divergence divided by n,
- $g_q(x) = (1 + x)^{-q}$ (has polynomial tails),
- π_1^N the reference prior of Berger and Bernardo (1992),
- (partly our intuition)

$$q = \frac{1}{2} + \inf\{q > 0 \ : \ \pi_1^D() \text{ is proper}\}.$$

DB priors, the examples and the criteria

- For the problems shown, DB priors lead to proposals that fully satisfy with criteria,

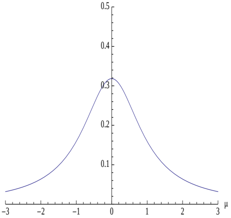
DB priors, the examples and the criteria

- For the problems shown, DB priors lead to proposals that fully satisfy with criteria,
- we expect this happening with broad generality (formal proofs are work in progress).

Problem 1: normal mean with σ unknown

In this case

$$\pi_1^D(\mu \mid \sigma) = Cauchy(\mu \mid 0, \sigma).$$

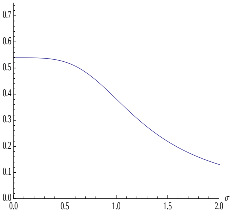


Coincides with Jeffreys' famous proposal.

Problem 2: normal standard deviation normal with μ unknown

In this case

$$\pi_1^D(\sigma \mid \mu) = \frac{\sqrt{\pi}}{4\Gamma(5/4)^2} \frac{1}{\sigma} \left(\frac{\sigma_0^2}{\sigma^2} + \frac{\sigma^2}{\sigma_0^2} \right)^{-1/2}.$$



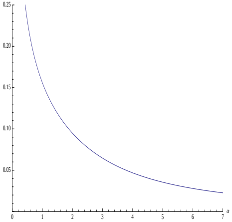
$(\sigma_0 = 1)$

Problem 3: gamma shape parameter (mean μ unknown)

In this case

$$\pi_1^D(\alpha \mid \mu) \propto \Big(1+(\alpha-\alpha_0)(\log(\frac{\alpha}{\alpha_0})+\psi(\alpha)-\psi(\alpha_0))\Big)^{-1/2}(\psi^{(1)}(\alpha)-\alpha^{-1})^{1/2},$$

where ψ and $\psi^{(1)}$ are the digamma and trigamma functions respectively.



$(\alpha_0 = 3)$

Thanks!

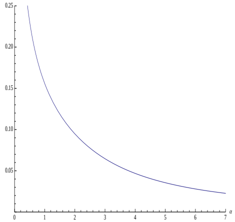
Thanks!

Thanks!

Problem 3: an educative radiography of $\pi_1^D(\alpha \mid \mu)$

The problem $H_0 : \alpha = 3$ vs. $H_1 : \alpha \neq 3$.

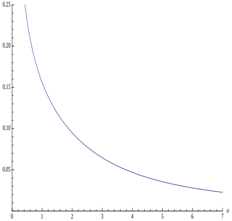
$$\pi_1^D(\alpha \mid \mu) \quad = \quad c(\alpha_0) D(\alpha, \alpha_0)^{-1/2} \quad \times \quad \pi^N(\alpha \mid \mu)$$



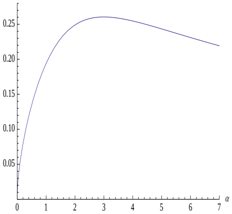
Problem 3: an educative radiography of $\pi_1^D(\alpha \mid \mu)$

The problem $H_0 : \alpha = 3$ vs. $H_1 : \alpha \neq 3$.

$$\pi_1^D(\alpha \mid \mu) \quad = \quad c(\alpha_0)D(\alpha, \alpha_0)^{-1/2} \quad \times \quad \pi^N(\alpha \mid \mu)$$



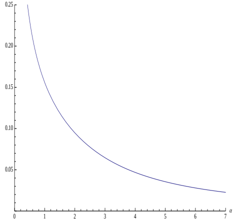
$=$



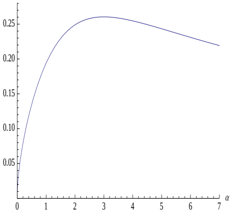
Problem 3: an educative radiography of $\pi_1^D(\alpha \mid \mu)$

The problem $H_0 : \alpha = 3$ vs. $H_1 : \alpha \neq 3$.

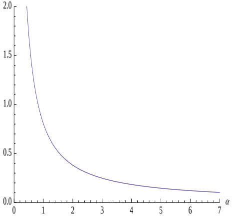
$$\pi_1^D(\alpha \mid \mu) \quad = \quad c(\alpha_0) D(\alpha, \alpha_0)^{-1/2} \quad \times \quad \pi^N(\alpha \mid \mu)$$



$=$



$\times$



Invariance criterion: surprising facts

- $\pi^H(\alpha)$ is typically improper (and hence could be multiplied by an arbitrary constant) and yet, if the same $\pi^H(\alpha)$ is used for all marginal models, the prior is appropriately calibrated across models in the strong sense of exact predictive matching.
- For invariant models, the combination of the Invariance criterion and (exact) Predictive matching criterion allows complete specification of the prior for α in all models and this argument does not require orthogonality, which, since Jeffreys (1961), has been viewed as a necessary condition to say that one can use a common prior for α in different models.
- For those concerned with the use of improper priors: the use of any approximating series of proper priors for $\pi^H(\alpha)$ will, in the limit, yield Bayes factors equal to that obtained directly from $\pi^H(\alpha)$.