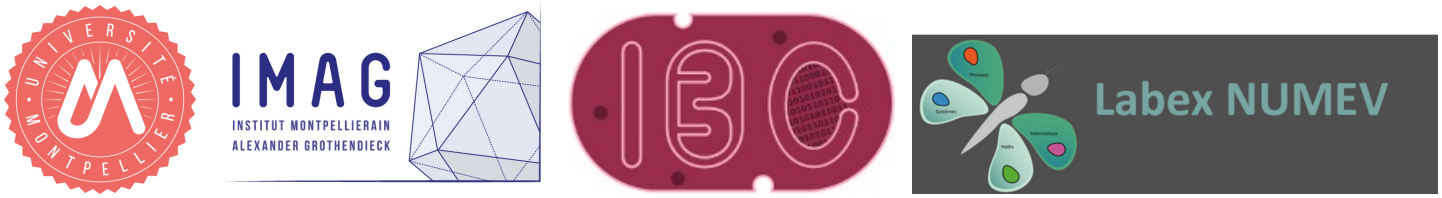
Bayesian paradigm, prior distributions
Bayes estimates, credible intervals
Bayesian discrimination between models

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck
Institut de Biologie Computationnelle (IBC)
Labex Numev



Joint work with Christian Robert

Plan

The Bayesian paradigm
Bayesian estimates
Conjugate prior
Noninformative prior
Jeffrey's prior
Bayesian credible intervals

The Model index as a Parameter
A key quantity, the integrated likelihood, also called the evidence
Why Bayesian inference embodies Occam's razor?
Bayesian test and Bayesian model choice: the same problem
The Bayes factor
The Ban on Improper Priors
Bayesian Model Averaging
Difficulties with the Bayesian model choice paradigm

The Bayesian paradigm

Bayes theorem = Inversion of probabilities

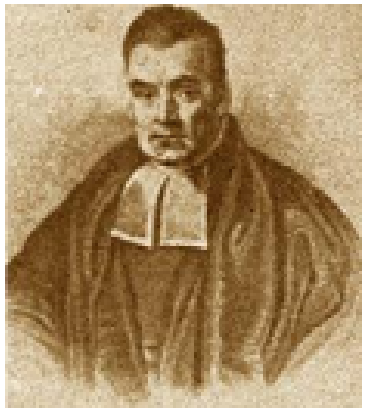
If A and B are events such that $\mathbb{P}(B) \neq 0$,
$\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ are related

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A) + \mathbb{P}(B|A^c)}$$

Reverend Thomas Bayes (ca. 1702–1761)

Presbyterian minister in Tunbridge Wells (Kent) from 1731, son of Joshua Bayes, nonconformist minister. Election to the *Royal Society* based on a tract of 1736 where he defended the views and philosophy of Newton.

Sole probability paper, *"Essay Towards Solving a Problem in the Doctrine of Chances"*, published posthumously in 1763 by Pierce and containing the seeds of *Bayes' Theorem*.



Subjectivism

Frank Plumpton Ramsey (1903 - 1930)

Bruno de Finetti (1906 - 1985)

Leonard Jimmie Savage (1921 - 1971)

Given an iid sample $\mathcal{D}_n = (x_1, \ldots, x_n)$ from a density $f(x|\theta)$, depending upon an unknown parameter $\theta \in \Theta$, the associated likelihood function is

$$\ell(\theta|\mathcal{D}_n) = \prod_{i=1}^{n} f(x_i|\theta)\,. \tag{1}$$

When $\mathcal{D}_n$ is a normal $\mathcal{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$, we get

$$\begin{aligned}
\ell(\theta|\mathcal{D}_n) &= \prod_{i=1}^{n} \exp\{-(x_i-\mu)^2/2\sigma^2\}/\sqrt{2\pi}\sigma \\
&\propto \exp\left\{-\sum_{i=1}(x_i-\mu)^2/2\sigma^2\right\}/\sigma^n \\
&\propto \exp\left\{-\left(n\mu^2 - 2n\bar{x}\mu + \sum_{i=1} x_i^2\right)/2\sigma^2\right\}/\sigma^n \\
&\propto \exp\left\{-\left[n(\mu-\bar{x})^2 + s^2\right]/2\sigma^2\right\}/\sigma^n,
\end{aligned}$$

$\bar{x}$ denotes the empirical mean and $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

In the Bayesian approach θ is considered as a random variable.

In some sense, the likelihood function is transformed into a *posterior* distribution, which is a valid probability distribution on Θ

$$\pi(\theta|\mathcal{D}_n) = \frac{\ell(\theta|\mathcal{D}_n)\pi(\theta)}{\int \ell(\theta|\mathcal{D}_n)\pi(\theta)\,\mathrm{d}\theta}\,. \tag{2}$$

$\pi(\theta)$ is called the *prior* distribution and it has to be chosen to start the analysis.

The posterior density is a probability density on the parameter, which does not mean the parameter θ need be a genuine random variable.

This density is used as an inferential tool, not as a truthful representation.

Two motivations:

- the prior distribution summarizes the *prior information* on θ. However, the choice of $\pi(\theta)$ is often decided on practical grounds rather than strong subjective beliefs...
- the Bayesian approach provides a fully probabilistic framework for the inferential analysis, with respect to a reference measure $\pi(\theta)$.

Suppose $\mathcal{D}_n$ is a normal $\mathcal{N}(\mu, \sigma^2)$ sample of size n.

When σ^2 is known, if $\mu \sim \mathcal{N}\left(0, \sigma^2\right)$, then

$$\begin{aligned}
\pi(\mu|\mathcal{D}_n) &\propto \pi(\mu)\, \ell(\theta|\mathcal{D}_n) \\
&\propto \exp\{-\mu^2/2\sigma^2\} \exp\left\{-n(\bar{x}-\mu)^2/2\sigma^2\right\} \\
&\propto \exp\left\{-(n+1)\mu^2/2\sigma^2 + 2n\mu\bar{x}/2\sigma^2\right\} \\
&\propto \exp\left\{-(n+1)[\mu - n\bar{x}/(n+1)]^2/2\sigma^2\right\} ,
\end{aligned}$$

$$\boxed{\mu|\mathcal{D}_n \sim \mathcal{N}\left(n\bar{x}/(n+1), \sigma^2/(n+1)\right)}$$

When σ^2 is unknown, $\theta = (\mu, \sigma^2)$, if $\mu|\sigma^2 \sim \mathcal{N}\left(0, \sigma^2\right)$ and $\sigma^2 \sim \mathcal{IG}(1,1)$, then

$$\begin{aligned}
\pi((\mu,\sigma^2)|\mathcal{D}_n) &\propto \pi(\sigma^2) \times \pi(\mu|\sigma^2) \times f(\mathcal{D}_n|\mu,\sigma^2) \\
&\propto (\sigma^{-2})^{1/2+2} \exp\left\{-(\mu^2+2)/2\sigma^2\right\} \\
&\quad \times (\sigma^{-2})^{n/2} \exp\left\{-\left(n(\mu-\overline{x})^2 + s^2\right)/2\sigma^2\right\} \\
&\propto (\sigma^2)^{-(n+5)/2} \exp\left\{-\left[(n+1)(\mu - n\bar{x}/(n+1))^2 + (2+s^2)\right]/2\sigma^2\right\} \\
&\propto (\sigma^2)^{-1/2} \exp\left\{-(n+1)[\mu - n\bar{x}/(n+1)]^2/2\sigma^2\right\} . \\
&\quad \times (\sigma^2)^{-(n+2)/2-1} \exp\left\{-(2+s^2)/2\sigma^2\right\} ,
\end{aligned}$$

$$\boxed{\mu|\mathcal{D}_n, \sigma^2 \sim \mathcal{N}(n\bar{x}/(n+1), \sigma^2/(n+1))}$$

$$\boxed{\sigma^2|\mathcal{D}_n \sim \mathcal{IG}((n+2)/2, [2+s^2]/2)}$$

Variability in σ^2 induces more variability in μ, the marginal posterior in μ being then a Student's t distribution

$$\boxed{\mu|\mathcal{D}_n \sim \mathcal{T}\left(n+2, n\bar{x}/(n+1), (2+s^2)/(n+1)(n+2)\right)}$$

Bayesian estimates

For a given loss function $L\left(\theta, \hat{\theta}(\mathcal{D}_n)\right)$, we deduce a Bayesian estimate by minimizing the posterior expected loss:

$$\mathbb{E}^{\pi}_{\theta|\mathcal{D}_n}\left(L\left(\theta, \hat{\theta}(\mathcal{D}_n)\right)\right) .$$

To minimize the posterior expected loss is equivalent to minimize the Bayes risk, the frequentist risk integrated over the prior distribution.

For instance, for the L_2 loss function, the corresponding Bayes optimum is the expected value of θ under the posterior distribution,

$$\hat{\theta}(\mathcal{D}_n) = \int \theta \, \pi(\theta|\mathcal{D}_n) \, \mathrm{d}\theta = \frac{\int \theta \, \ell(\theta|\mathcal{D}_n) \, \pi(\theta) \, \mathrm{d}\theta}{\int \ell(\theta|\mathcal{D}_n) \, \pi(\theta) \, \mathrm{d}\theta} \,, \tag{3}$$

for a given sample $\mathcal{D}_n$.

When no specific penalty criterion is available, the posterior expectation is often used as a default estimator, although alternatives are also available. For instance, the *maximum a posteriori estimator* (MAP) is defined as

$$\hat{\theta}(\mathcal{D}_n) \in \text{argmax}_\theta \quad \pi(\theta|\mathcal{D}_n)\,.$$

Similarity of with the maximum likelihood estimator: the influence of the prior distribution $\pi(\theta)$ on the estimate progressively disappears as the number of observations n increases.

Conjugate prior

The selection of the prior distribution is an important issue in Bayesian statistics.

When prior information is available about the data or the model, it can be used in building the prior.

In many situations, however, the selection of the prior distribution is quite delicate.

Since the choice of the prior distribution has a considerable influence on the resulting inference, this inferential step must be conducted with the utmost care.

Conjugate priors are such that the prior and posterior densities belong to the same parametric family.

An advantage when using a conjugate prior, is that one has to select only a few parameters to determine the prior distribution.

But the information known a priori may be either insufficient or incompatible with the structure imposed by conjugacy.

Justifications

- Limited/finite information conveyed by x
- Preservation of the structure of $\pi(\theta)$
- Device of virtual past observations
- Linearity of some estimators
- But mostly... tractability and simplicity
- First approximations to adequate priors, backed up by robustness analysis

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\rho(\sigma^2\mu + \tau^2 x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + x, \beta + n - x)$

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Negative Binomial	Beta	
$\mathcal{N}eg(m,\theta)$	$\mathcal{B}e(\alpha,\beta)$	$\mathcal{B}e(\alpha+m,\beta+x)$
Multinomial	Dirichlet	
$\mathcal{M}_k(\theta_1,\ldots,\theta_k)$	$\mathcal{D}(\alpha_1,\ldots,\alpha_k)$	$\mathcal{D}(\alpha_1+x_1,\ldots,\alpha_k+x_k)$
Normal	Gamma	
$\mathcal{N}(\mu,1/\theta)$	$\mathcal{G}a(\alpha,\beta)$	$\mathcal{G}(\alpha+0.5,\beta+(\mu-x)^2/2)$

Noninformative prior

Conjugate priors are nice to work with, but require hyperparameters's determination.

One can opt for a completely different perspective and rely on so-called *noninformative* priors that aim at attenuating the impact of the prior on the resulting inference.

These priors are fundamentally defined as coherent extensions of the uniform distribution.

For unbounded parameter spaces, the densities of noninformative priors actually may fail to integrate to a finite number and they are defined instead as positive measures.

Generalized Bayesian estimators with improper prior distributions

For instance, *location models*

$$x|\theta \sim f(x-\theta)$$

are usually associated with flat priors $\pi(\theta) = 1$, while *scale models*

$$x|\theta \sim \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$$

are usually associated with the log-transform of a flat prior, that is,

$$\pi(\theta) = 1/\theta\,.$$

In a more general setting, the noninformative prior favored by most Bayesians is the so-called *Jeffreys prior* which is related to the Fisher information matrix

$$I^F(\theta) = \mathrm{var}_\theta \left(\frac{\partial \log f(X|\theta)}{\partial \theta} \right)$$

by

$$\pi^J(\theta) = \left| I^F(\theta) \right|^{1/2},$$

where $|I|$ denotes the determinant of the matrix I.

Suppose $\mathcal{D}_n$ is a normal $\mathcal{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$.
The Fisher information matrix leads to the Jeffreys prior $\pi^J(\theta) = 1/\sigma^3$ and

$$\begin{aligned}\pi((\mu,\sigma^2)|\mathcal{D}_n) &\propto (\sigma^{-2})^{(3+n)/2} \exp\left\{-\left(n(\mu-\overline{x})^2 + s^2\right)/2\sigma^2\right\} \\ &\propto \sigma^{-1} \exp\left\{-n(\mu-\bar{x})^2/2\sigma^2\right\} \times (\sigma^2)^{-(n+2)/2} \exp\left\{\frac{-s^2}{2\sigma^2}\right\},\end{aligned}$$

$$\boxed{\mu|\sigma^2, \mathcal{D}_n \sim \mathcal{N}\left(\bar{x}, \sigma^2/n\right)}$$

$$\boxed{\sigma^2|\mathcal{D}_n \sim \mathcal{IG}\left(n/2, s^2/2\right)}$$

$$\boxed{\mu|\mathcal{D}_n \sim \mathcal{T}\left(n, \bar{x}, s^2/n^2\right))}$$

Bayesian Credible Intervals

Since the Bayesian approach processes θ as a random variable, a natural definition of a confidence region on θ is to determine $C(\mathcal{D}_n)$ such that

$$\pi(\theta \in C(\mathcal{D}_n)|\mathcal{D}_n) = 1 - \alpha \tag{4}$$

where α is a predetermined level.

The integration is done over the parameter space, rather than over the observation space.

The quantity $1 - \alpha$ thus corresponds to the probability that a random θ belongs to this set $C(\mathcal{D}_n)$, rather than to the probability that the random set contains the "true" value of θ.

Given this drift in the interpretation of a confidence set is called a *credible set* by Bayesians.

A standard credible set corresponds to the values of θ with the highest posterior values,

$$C(\mathcal{D}_n) = \{\theta;\, \pi(\theta|\mathcal{D}_n) \geq k_\alpha\} \ ,$$

where k_α is determined by the coverage constraint (4). This region is called the *highest posterior density* (HPD) region.

Once again, suppose $\mathcal{D}_n$ is a normal $\mathcal{N}(\mu, \sigma^2)$ sample of size n and $\theta = (\mu, \sigma^2)$.

$$\boxed{\mu|\sigma^2, \mathcal{D}_n \sim \mathcal{N}\left(\bar{x}, \sigma^2/n\right)}$$

$$\boxed{\sigma^2|\mathcal{D}_n \sim \mathcal{IG}\left(n/2, s^2/2\right)}$$

$$\boxed{\mu|\mathcal{D}_n \sim \mathcal{T}\left(n, \bar{x}, s^2/n^2\right))}$$

Therefore, the credible interval of probability $1-\alpha$ on μ is

$$\left[\bar{x} - t_{1-\alpha/2,n}\sqrt{(n-1)s^2/n^2}, \bar{x} + t_{1-\alpha/2,n}\sqrt{(n-1)s^2/n^2}\right].$$

The Model index as a Parameter

When are comparing models with indices $k = 1, 2, \ldots, J$, we introduce a model indicator $\mathfrak{M}$ taking values in $\{1, 2, \ldots, J\}$ and representing the index of the "true" model.

If $\mathfrak{M} = k$, the data $\mathcal{D}_n$ are generated from a statistical model $\mathfrak{M}_k$ with likelihood $\ell_k(\theta_k|\mathcal{D}_n)$ and parameter $\theta_k \in \Theta_k$.

Bayes procedures will depend on the posterior probabilities in the model space

$$\mathbb{P}^{\pi}(\mathfrak{M} = k|\mathcal{D}_n)\,,$$

The prior π is defined over the collection of model indices, $\{1, 2, \ldots, J\}$, and, conditionally on the model index $\mathfrak{M}$, on the corresponding parameter space, Θ_k.

Choice of the prior model probabilities $\mathbb{P}^\pi(\mathfrak{M} = k)$:

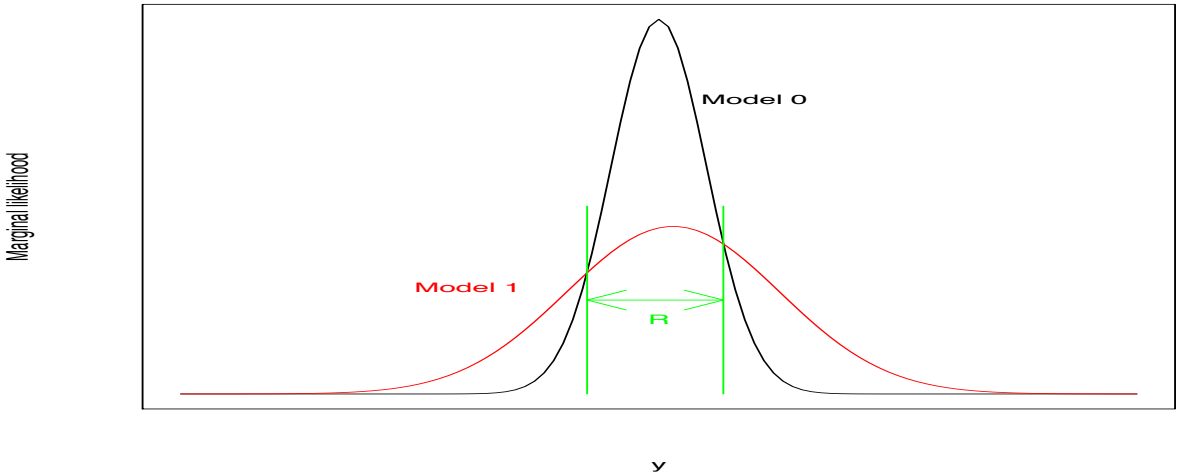
- in some cases, there is experimental or subjective evidence about those probabilities,
- typically, we are forced to settle for equal weights $\mathbb{P}^\pi(\mathfrak{M} = k) = 1/J$.

A key quantity, the integrated likelihood, also called the evidence

$$\mathbb{P}^{\pi}(\mathfrak{M} = k|\mathcal{D}_n) \propto \mathbb{P}^{\pi}(\mathfrak{M} = k) \int \ell_k(\theta_k|\mathcal{D}_n)\pi_k(\theta_k)\,\mathrm{d}\theta_k\,.$$

$\mathbb{P}^{\pi}(\mathfrak{M} = k|\mathcal{D}_n)$ is the core object in Bayesian model choice, the default procedure is to select the model with the highest posterior probability. That corresponds to 0-1 loss function.

Why Bayesian inference embodies Occam's razor?



This graph gives the basic intuition for why complex models can turn out to be less probable.

The horizontal axis represents the space of possible data sets. Bayes' theorem rewards models in proportion to how much they predicted the data that occurred. These predictions are quantized by a normalized probability distribution.

A simple model, like Model 0, makes only a limited range of predictions; a more powerful model, like Model 1, that has, for example, more free parameters, is able to predict a greater variety of data sets.

Suppose that equal prior probabilities have been assigned to the two models. Then, if the data set falls in region R, the less powerful model will be the more probable model.

The marginal likelihood corresponds to a penalized likelihood!

The BIC information criterium comes from an asymptotic Laplace approximation of the evidence.

Bayesian test and Bayesian model choice: the same problem

For instance, given a single observation $x \sim \mathcal{N}(\mu, \sigma^2)$ from a normal model where σ^2 is known,

If $\mu \sim \mathcal{N}(\xi, \tau^2)$, the posterior distribution $\mu|x \sim \mathcal{N}(\xi(x), \omega^2)$ with

$$\xi(x) = \frac{\sigma^2\xi + \tau^2 x}{\sigma^2 + \tau^2} \quad \text{and} \quad \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

If the question of interest is to decide whether μ is negative or positive, we can directly compute

$$\begin{aligned} \mathbb{P}^{\pi}(\mu < 0|x) &= \mathbb{P}^{\pi}\left(\frac{\mu - \xi(x)}{\omega} < \frac{-\xi(x)}{\omega}\right) \\ &= \Phi\left(-\xi(x)/\omega\right) . \end{aligned} \tag{5}$$

where Φ is the normal cdf.

This computation does not seem to follow from the principles we just stated but it is only a matter of perspective.

We can derive the priors on both models from the original prior.

Deriving this posterior probability indeed means that, a priori, μ is negative with probability $\mathbb{P}^{\pi}(\mu < 0) = \Phi(-\xi/\tau)$ and that, in this model, the prior on μ is the truncated normal

$$\pi_1(\mu) = \frac{\exp\{-(\mu-\xi)^2/2\tau^2\}}{\sqrt{2\pi}\tau\Phi(-\xi/\tau)}\,\mathbb{I}_{\mu<0}\,,$$

while μ is positive with probability $\Phi(\xi/\tau)$ and, in this second model, the prior on μ is the truncated normal

$$\pi_2(\mu) = \frac{\exp\{-(\mu-\xi)^2/2\tau^2\}}{\sqrt{2\pi}\tau\Phi(\xi/\tau)}\,\mathbb{I}_{\mu>0}\,.$$

The Bayes factor

$$B_{21}^{\pi}(\mathcal{D}_n) = \frac{\mathbb{P}^{\pi}(\mathfrak{M}=2|\mathcal{D}_n)/\mathbb{P}^{\pi}(\mathfrak{M}=1|\mathcal{D}_n)}{\mathbb{P}^{\pi}(\mathfrak{M}=2)/\mathbb{P}^{\pi}(\mathfrak{M}=1)}\,,$$

While this quantity is a simple one-to-one transform of the posterior probability, it can be used for Bayesian model choice without first resorting to a determination of the prior weights of both models.

$$B_{21}^{\pi}(\mathcal{D}_n) = \frac{\int_{\Theta_2} \ell_2(\theta_2|\mathcal{D}_n)\pi_2(\theta_2)\,\mathrm{d}\theta_2}{\int_{\Theta_1} \ell_1(\theta_1|\mathcal{D}_n)\pi_1(\theta_1)\,\mathrm{d}\theta_1} = \frac{m_2(\mathcal{D}_n)}{m_1(\mathcal{D}_n)}\,,$$

The evidence brought by the data $\mathcal{D}_n$ can be calibrated using for instance Jeffreys' scale of evidence:

- if $\log_{10}(B_{21}^{\pi})$ is between 0 and 0.5, the evidence against model $\mathfrak{M}_1$ is *weak*,
- if it is between 0.5 and 1, it is *substantial*,
- if it is between 1 and 2, it is *strong*, and
- if it is above 2, it is *decisive*.

While this scale is purely arbitrary, it provides a reference for model assessment in a generic setting.

The Bayes factor contains an automated penalization for complexity.

The Ban on Improper Priors

Looking at the expression of the Bayes factor,

$$B_{21}^{\pi}(\mathcal{D}_n) = \frac{\int_{\Theta_2} \ell_2(\theta_2|\mathcal{D}_n)\pi_2(\theta_2)\,\mathrm{d}\theta_2}{\int_{\Theta_1} \ell_1(\theta_1|\mathcal{D}_n)\pi_1(\theta_1)\,\mathrm{d}\theta_1}\,,$$

it is clear that, when either π_1 or π_2 are improper, it is impossible to normalise the improper measures in a unique manner.

Therefore, the Bayes factor becomes completely arbitrary since it can be multiplied by one or two arbitrary constants.

Since improper priors are an essential part of the Bayesian approach, there are many proposals found in the literature to overcome this ban.

Most of those proposals rely on a device that transforms the improper prior into a proper probability distribution by exploiting a fraction of the data $\mathcal{D}_n$.

The variety of available solutions is due to the many possibilities of removing the dependence on the choice of the portion of the data used in the first step.

The resulting procedures are called *pseudo-Bayes factors*, although some may actually correspond to true Bayes factors.

There is a major exception to this ban on improper priors that we can exploit.

If both models under comparison have parameters that have similar enough meanings to share the same prior distribution, as for instance a measurement error σ^2, then the normalisation issue vanishes.

Note that we are not assuming that parameters are *common* to both models and thus that we do not contradict the earlier warning about different parameters to different models.

When comparing two id normal samples, $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, with respective distributions $\mathcal{N}(\mu_x, \sigma^2)$ and $\mathcal{N}(\mu_y, \sigma^2)$, we can examine whether or not the two means are identical, i.e. $\mu_x = \mu_y$ (corresponding to model M_1).

To take advantage of the structure of this model, we can assume that σ^2 is a measurement error with a similar meaning under both models and thus that the same prior $\pi_\sigma(\sigma^2)$ can be used under both models. This means that the Bayes factor

$$B_{21}^{\pi}(\mathcal{D}_n) = \frac{\int \ell_2(\mu_x, \mu_y, \sigma|\mathcal{D}_n)\pi(\mu_x, \mu_y)\pi_\sigma(\sigma^2)\,\mathrm{d}\sigma^2\,\mathrm{d}\mu_x\,\mathrm{d}\mu_y}{\int \ell_1(\mu, \sigma|\mathcal{D}_n)\pi_\mu(\mu)\pi_\sigma(\sigma^2)\,\mathrm{d}\sigma^2\,\mathrm{d}\mu}$$

does not depend on the normalizing constant used for $\pi_\sigma(\sigma^2)$ and thus that we can still use an improper prior such as $\pi_\sigma(\sigma^2) = 1/\sigma^2$ in that case.

Furthermore, we can rewrite μ_x and μ_y as $\mu_x = \mu - \xi$ and $\mu_y = \mu + \xi$, respectively, and use a prior of the form $\pi(\mu, \xi) = \pi_\mu(\mu)\pi_\xi(\xi)$ on the new parameterization so that, again, the same prior π_μ can be used under both models.

The same cancellation of the normalizing constant occurs for π_μ, which means a Jeffreys prior $\pi_\mu(\mu) = 1$ can be used.

However, we need a proper and well-defined prior on ξ, for instance $\xi \sim \mathcal{N}(0, \tau^2)$.

$$B_{21}^{\pi}(\mathcal{D}_n) = \frac{\displaystyle\int e^{-n\left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\,\sigma^{-2n-2}e^{-\xi^2/2\tau^2}/\tau\sqrt{2\pi}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu\,\mathrm{d}\xi}{\displaystyle\int e^{-n\left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s_{xy}^2\right]/2\sigma^2}\,\sigma^{-2n-2}\,\mathrm{d}\sigma^2\,\mathrm{d}\mu}$$

$$= \frac{\displaystyle\int \left[(\mu-\xi-\bar{x})^2+(\mu+\xi-\bar{y})^2+s_{xy}^2\right]^{-n} e^{-\xi^2/2\tau^2}/\tau\sqrt{2\pi}\,\mathrm{d}\mu\,\mathrm{d}\xi}{\displaystyle\int \left[(\mu-\bar{x})^2+(\mu-\bar{y})^2+s_{xy}^2\right]^{-n}\,\mathrm{d}\mu},$$

where s_{xy}^2 denotes the average

$$s_{xy}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\,.$$

While the denominator can be completely integrated out, the numerator cannot. A numerical approximation to B_{21}^{π} is thus necessary.

Comparing models $M_1 : \mathcal{N}(0,\sigma^2)$ under the prior $\pi_1(\sigma^2) = 1/\sigma^2$ and $M_2 : \mathcal{N}(\mu,\sigma^2)$ under the prior made of $\pi_2(\sigma^2) = 1/\sigma^2$ and $\pi_2(\mu|\sigma^2)$ equal to the normal $\mathcal{N}(0,\sigma^2)$ density, the Bayes factor is

$$
\begin{aligned}
B_{21}^{\pi}(\mathcal{D}_n) &= \frac{\displaystyle\int e^{-[n(\bar{x}-\mu)^2+s^2]/2\sigma^2}\, e^{-\mu^2/2\sigma^2}\, \sigma^{-n-1-2}\, \frac{\mathrm{d}\mu \mathrm{d}\sigma^2}{\sqrt{2\pi}}}{\displaystyle\int e^{-[n\bar{x}^2+s^2]/2\sigma^2}\, \sigma^{-n-2}\, \mathrm{d}\sigma^2} \\
&= \frac{\displaystyle\int e^{-(n+1)[\mu-n\bar{x}/(n+1)]^2}\, e^{-[n\bar{x}^2/(n+1)+s^2]/2\sigma^2}\, \sigma^{-n-3}\, \frac{\mathrm{d}\mu \mathrm{d}\sigma^2}{\sqrt{2\pi}}}{\left[\dfrac{n\bar{x}^2+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)}
\end{aligned}
$$

$$
\begin{aligned}
B_{21}^{\pi}(\mathcal{D}_n) &= \frac{\displaystyle\int (n+1)^{-1/2}\, e^{-[n\bar{x}^2/(n+1)+s^2]/2\sigma^2}\, \sigma^{-n-2}\, \mathrm{d}\sigma^2}{\left[\dfrac{n\bar{x}^2+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)} \\
&= \frac{(n+1)^{-1/2}\, \left[\dfrac{n\bar{x}^2/(n+1)+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)}{\left[\dfrac{n\bar{x}^2+s^2}{2}\right]^{-n/2} \Big/ \Gamma(n/2)} \\
&= (n+1)^{-1/2}\, \left[\frac{n\bar{x}^2+s^2}{n\bar{x}^2/(n+1)+s^2}\right]^{n/2},
\end{aligned}
$$

taking once again advantage of the normalising constant of the gamma distribution. It therefore increases to infinity with $\bar{x}^2/s^2$, starting from $1/\sqrt{n+1}$ when $\bar{x}=0$.

Bayesian Model Averaging

The posterior probabilities in the model space can be used to average over the decisions coming from different models.

Suppose that we are interested in the prediction of z and that, for model k, the predictive distribution of z is $g_k(z|\mathcal{D}_n)$.

The average predictive of z is

$$\sum_{k=1}^{J} \mathbb{P}^{\pi}(\mathfrak{M} = k|\mathcal{D}_n) g_k(z|\mathcal{D}_n)\,.$$

Difficulties with the Bayesian model choice paradigm

Prior difficulties:

- When we have prior informations, how to choose the prior distributions on the parameters of each model in a compatible way? What about the prior distribution in the models's space?

- When we do not have any prior information, **we can not use improper prior distribution**. Indeed, in that case, the models's posterior probabilities are only defined up to some arbitrary constants. How to choose the various prior distributions?

Computational difficulties:

- How to approximate the evidences?
- When the number of models in consideration is huge, how to explore the models's space?