Convergence mode for prior distributions

Pierre Druilhet

Laboratoire de Mathématiques UMR CNRS 6620
Université Blaise Pascal, Clermont-Ferrand (France)

Bayesian week
CIRM 2016

Introduction

The use on improper priors has been largely debated in the literature

- Against improper prior
    - it is not a probability,
    - many undesirable feature appears : marginalization paradox, non-conglomerability, Jeffreys-Lindley paradox,...
- For improper prior
    - necessary to obtain complete class of admissible estimator.
    - appear in the construction of "non-informative" priors.
    - corresponds to limiting posterior distributions.
    - replacing an improper prior by a proper prior approximation may provide a false impression of safety.

A BASIC EXAMPLE

- $X|\theta \sim \mathcal{N}(\theta, 1)$,
  $\theta \sim \mathcal{N}(0, n)$

$$\mathbb{E}_{\mathcal{N}(0,n)}(\theta|x) = \frac{n}{n+1}x \xrightarrow[n\to+\infty]{} x = \mathbb{E}_{\lambda}(\theta|x)$$

  where λ is the Laplace prior.

- So, we would like to write

$$\mathcal{N}(0, n) \xrightarrow[n\to+\infty]{} \lambda \quad \text{(Laplace prior)}.$$

- **Question** : does this apparent convergent depends on the likelihood $f(x|\theta)$ or is it intrinsic ?

A basic example

- $X|\theta \sim \mathcal{N}(\theta, 1)$,
  $\theta \sim \mathcal{N}(0, n)$

$$\mathbb{E}_{\mathcal{N}(0,n)}(\theta|x) = \frac{n}{n+1}x \xrightarrow[n\to+\infty]{} x = \mathbb{E}_{\lambda}(\theta|x)$$

  where λ is the Laplace prior.

- So, we would like to write

$$\mathcal{N}(0, n) \xrightarrow[n\to+\infty]{} \lambda \quad \text{(Laplace prior).}$$

- **Question** : does this apparent convergent depends on the likelihood $f(x|\theta)$ or is it intrinsic ?

A basic example

- $X|\theta \sim \mathcal{N}(\theta, 1)$,
  $\theta \sim \mathcal{N}(0, n)$

$$\mathbb{E}_{\mathcal{N}(0,n)}(\theta|x) = \frac{n}{n+1}x \xrightarrow[n\to+\infty]{} x = \mathbb{E}_{\lambda}(\theta|x)$$

  where λ is the Laplace prior.

- So, we would like to write

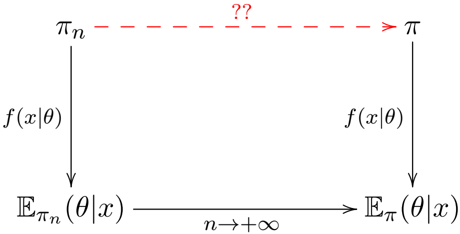
$$\mathcal{N}(0, n) \xrightarrow[n\to+\infty]{} \lambda \quad \text{(Laplace prior)}.$$

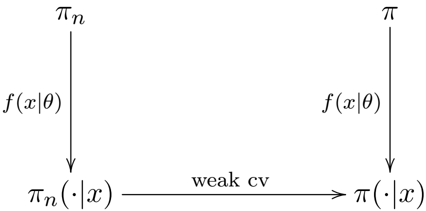
- **Question** : does this apparent convergent depends on the likelihood $f(x|\theta)$ or is it intrinsic ?

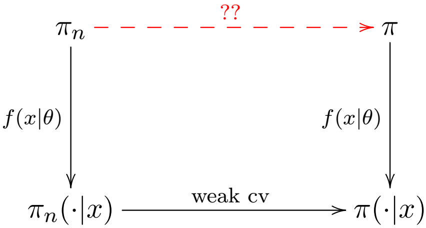
More generally

$$
\begin{array}{ccc}
\pi_n & & \pi \\
\Big\downarrow {\scriptstyle f(x|\theta)} & & \Big\downarrow {\scriptstyle f(x|\theta)} \\
\mathbb{E}_{\pi_n}(\theta|x) & \xrightarrow[n\to+\infty]{} & \mathbb{E}_{\pi}(\theta|x)
\end{array}
$$

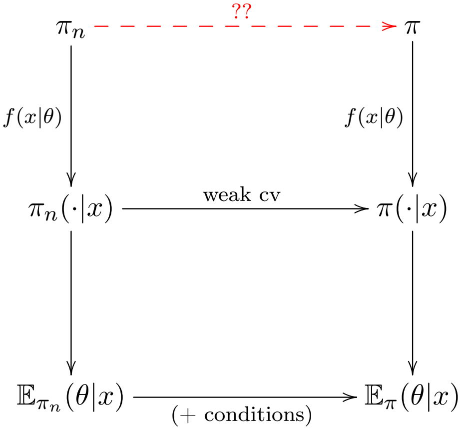
More generally




π_n
π
$f(x|\theta)$
$f(x|\theta)$
$\pi_n(\cdot|x)$
weak cv
$\pi(\cdot|x)$



π_n
??
π
$f(x|\theta)$
$f(x|\theta)$
$\pi_n(\cdot|x)$
weak cv
$\pi(\cdot|x)$


π_n
??
π
$f(x|\theta)$
$f(x|\theta)$
$\pi_n(\cdot|x)$
weak cv
$\pi(\cdot|x)$
$\mathbb{E}_{\pi_n}(\theta|x)$
(+ conditions)
$\mathbb{E}_{\pi}(\theta|x)$

Notation

Let Π be a measure on Θ and h a real-valued function:

$$\Pi(h) = \int_{\Theta} h(\theta)\pi(\theta)\, d\theta$$

Weak convergence of probability measures

$$\Pi_n \xrightarrow[n\to+\infty]{weakly} \Pi \iff \Pi_n(h) \xrightarrow[n\to+\infty]{} \Pi(h), \quad \forall h \in \mathcal{C}_b.$$

where C_b ={ continuous real-valued bounded functions on Θ}

Vague convergence of Radon measures

If Π_n and Π are non-null Radon measures (finite on compact sets),

$$\Pi_n \xrightarrow[n\to+\infty]{vaguely} \Pi \iff \Pi_n(h) \xrightarrow[n\to+\infty]{} \Pi(h), \quad \forall h \in \mathcal{C}_K.$$

where C_K is the set of real-valued continuous function with compact support.

If Π_n are probability measures that converges vaguely to Π, then

$$\Pi(\Theta) \leq \limsup_n \Pi_n(\Theta) = 1.$$

Thus a sequence of probability measures cannot converge to an improper prior.

We have to find another convergence mode !

Two approaches

- Using projective space
- Using finitely additive probabilities (FAP)

Projective space of measure (PSM)

The priors Π and $\alpha\,\Pi$ (with $\alpha > 0$) give the same posterior. So, we identify them:

$$\Pi \sim \Pi' \Longleftrightarrow \Pi' = \alpha\,\Pi \quad \text{for some } \alpha > 0.$$

$$\Downarrow$$

The quotient space associated to this equivalence relation is a **projective space**. The equivalent class associated to Π is denote by

$$\overline{\Pi} = \{\alpha\Pi; \alpha > 0\}\,.$$

For example, the Laplace prior

$$\pi(\theta) \propto 1$$

is an object defined in the PSM and one representative is $\pi(\theta) = 1$.

Projective space of measure (PSM)

The priors Π and $\alpha\,\Pi$ (with $\alpha > 0$) give the same posterior. So, we identify them:

$$\Pi \sim \Pi' \Longleftrightarrow \Pi' = \alpha\,\Pi \quad \text{for some } \alpha > 0.$$

$$\Downarrow$$

The quotient space associated to this equivalence relation is a **projective space**. The equivalent class associated to Π is denote by

$$\overline{\Pi} = \{\alpha\Pi; \alpha > 0\}\,.$$

For example, the Laplace prior

$$\pi(\theta) \propto 1$$

is an object defined in the PSM and one representative is $\pi(\theta) = 1$.

Q-VAGUE CONVERGENCE

Let Π_n be a sequence of Radon measures, the convergence, called q-vague convergence (Bioche and D. 2016), derived from the quotient space topology is:

$$\Pi_n \xrightarrow[n\to+\infty]{q-vaguely} \Pi \iff \overline{\Pi}_n \xrightarrow[n\to+\infty]{\text{quotient topology}} \overline{\Pi}$$

More tractable, staying in the initial space, we have:

$$\Pi_n \xrightarrow[n\to+\infty]{q-vaguely} \Pi \iff a_n\Pi_n \xrightarrow[n\to+\infty]{vaguely} \Pi \text{ for some } a_n > 0.$$

Q-VAGUE CONVERGENCE

Let Π_n be a sequence of Radon measures, the convergence, called q-vague convergence (Bioche and D. 2016), derived from the quotient space topology is:

$$\Pi_n \xrightarrow[n\to+\infty]{q-vaguely} \Pi \iff \overline{\Pi}_n \xrightarrow[n\to+\infty]{\text{quotient topology}} \overline{\Pi}$$

More tractable, staying in the initial space, we have:

$$\Pi_n \xrightarrow[n\to+\infty]{q-vaguely} \Pi \iff a_n\Pi_n \xrightarrow[n\to+\infty]{vaguely} \Pi \text{ for some } a_n > 0.$$

Proposition

If Π_n and Π are probabilities,

$$\Pi_n \xrightarrow[n\to+\infty]{q-vaguely} \Pi \iff \Pi_n \xrightarrow[n\to+\infty]{weakly} \Pi$$

So, the q-vague converge is an extension of the usual weak convergence of probability measures.

Proposition

The limit is unique (up to within a scalar factor) : the projective space is an Hausdorff space.

Proposition

Any improper prior is limit of a sequence of probability measures.
Any probability measure is limit of a sequence of improper measures.

Examples

- $\mathcal{N}(0,n) \xrightarrow[n\to+\infty]{q-vaguely} \lambda$,
- $\mathcal{U}_{[-n,n]} \xrightarrow[n\to+\infty]{q-vaguely} \lambda$,
- $\mathcal{N}(n,n) \xrightarrow[n\to+\infty]{q-vaguely} e^{\theta}d\theta$,
- $\mathcal{N}(n,\sqrt{n})$ does not converge,
- $\mathcal{P}(n)$ does not converge.

Example (2)

A gamma distribution $\gamma(a_n, b_n)$ converge to $1/\theta \mathbb{1}_{\theta>0}$ when a_n and b_n go to 0.

- For $\gamma(\frac{1}{n}, \frac{1}{n})$, $\mathbb{E}_n(\theta) = 1$ and $\lim_n \mathrm{var}_n = +\infty$.
- For $\gamma(\frac{1}{n}, \frac{1}{\sqrt{n}})$, $\lim_n \mathbb{E}_n(\theta) = 0$ and $\lim_n \mathrm{var}_n(\theta) = 1$.
- For $\gamma(\frac{1}{n}, \frac{1}{n^{\frac{1}{3}}})$, $\lim_n \mathbb{E}_n(\theta) = 0$ and $\lim_n \mathrm{var}_n(\theta) = 0$.
- For $\gamma(\frac{1}{n}, \frac{1}{n^2})$, $\lim_n \mathbb{E}_n(\theta) = +\infty$ and $\lim_n \mathrm{var}_n(\theta) = +\infty$.
- For $\gamma(\frac{1}{n}, \frac{1}{n^{\frac{2}{3}}})$, $\lim_n \mathbb{E}_n(\theta) = 0$ and $\lim_n \mathrm{var}_n(\theta) = +\infty$.

Example (3)

For the beta distribution, there are two limits in the litterature:

$$\beta\left(\frac{1}{n},\frac{1}{n}\right) \xrightarrow[n\to+\infty]{weakly} \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$$

and

$$\beta\left(\frac{1}{n},\frac{1}{n}\right) \xrightarrow[n\to+\infty]{} \frac{1}{\theta(1-\theta)} \quad \text{(Haldane prior)}$$

Example (3)

For the beta distribution, there are two limits in the litterature:

$$\text{if } \Theta = [0,1], \quad \beta\left(\frac{1}{n}, \frac{1}{n}\right) \xrightarrow[n\to+\infty]{q-vaguely} \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$$

and

$$\beta\left(\frac{1}{n}, \frac{1}{n}\right) \xrightarrow[n\to+\infty]{} \frac{1}{\theta(1-\theta)} \quad \text{(Haldane prior)}$$

Example (3)

For the beta distribution, there are two limits in the litterature:

$$\text{if } \Theta = [0,1], \quad \beta\left(\frac{1}{n}, \frac{1}{n}\right) \xrightarrow[n\to+\infty]{q-vaguely} \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$$

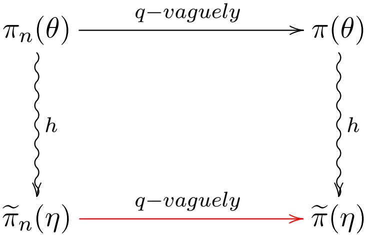
and

$$\text{if } \Theta =]0,1[, \quad \beta\left(\frac{1}{n}, \frac{1}{n}\right) \xrightarrow[n\to+\infty]{q-vaguely} \frac{1}{\theta(1-\theta)} \quad \text{(Haldane prior)}$$

Re-parameterizaton

The $q - vague$ convergence is transmitted by continous re-parameterization

Let $\eta = h(\theta)$, with h continuous



Exemple

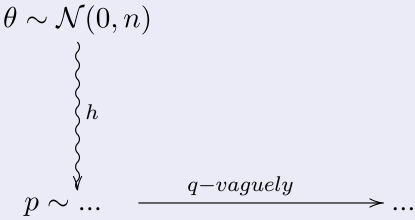
$X|\theta \sim B(N, p)$

$\theta = \log\left(\frac{p}{1-p}\right)$

$\theta \sim \mathcal{N}(0, n)$,

$p = \frac{e^\theta}{1+e^\theta} = h(\theta)$

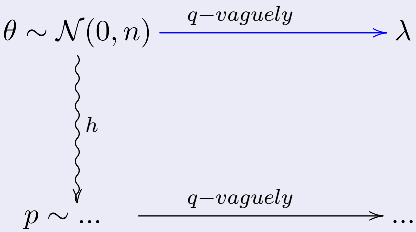
When n goes to $+\infty$, what is the limiting prior for p ?



Exemple

$X|\theta \sim B(N,p)$
$\theta = \log\left(\frac{p}{1-p}\right)$
$\theta \sim \mathcal{N}(0,n)$,
$p = \frac{e^\theta}{1+e^\theta} = h(\theta)$

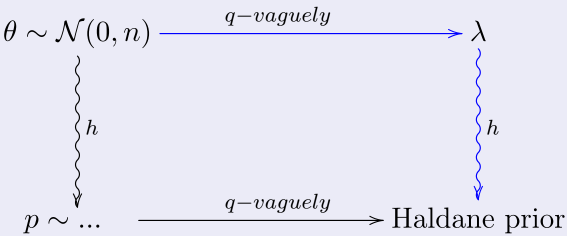
When n goes to $+\infty$, what is the limiting prior for p ?



Exemple

$X|\theta \sim B(N,p)$
$\theta = \log\left(\frac{p}{1-p}\right)$
$\theta \sim \mathcal{N}(0,n)$,
$p = \frac{e^\theta}{1+e^\theta} = h(\theta)$

When n goes to $+\infty$, what is the limiting prior for p ?



Approximation of Jeffreys' prior by conjugate priors.

Let $X|\theta \sim f(x|\theta) = \exp\{\theta \cdot t(x) - \phi(\theta)\}\, h(x)$ (exponential model).

We want to approximate the Jeffreys prior $\pi^J(\theta) \propto |I_X(\theta)|^{1/2}$ by conjugate priors.

We define the invariant family of conjugate priors (see D. and Pommeret, 2012):

$$\pi^J_{\alpha,\beta}(\theta) \propto \exp\{\alpha.\theta - \beta\phi(\theta)\}\, |I_\theta(\theta)|^{\frac{1}{2}},$$

We have the convergence result:

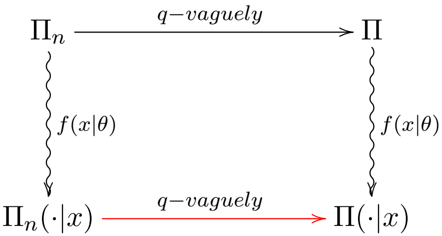
$$\Pi^J_{\alpha_n,\beta_n} \xrightarrow[\substack{\alpha_n \to 0 \\ \beta_n \to 0}]{q-vaguely} \Pi^J$$

Note that the standard conjugate prior $\pi_{\alpha,\beta}(\theta) \propto \exp\{\alpha.\theta - \beta\phi(\theta)\}$ approximates the Laplace prior (and is not invariant by re-parameterization).

From prior to posterior convergence

If the likelihood $f(x|\theta)$ is continuous in θ, the q-vague convergence is transmitted to the posterior.

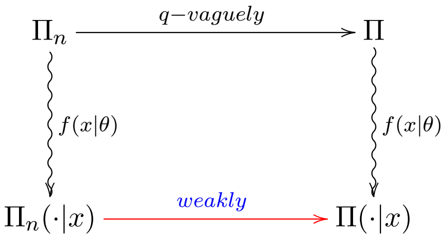
$$\Pi_n \xrightarrow[n \to +\infty]{q-vaguely} \Pi \implies \Pi_n(\cdot|x) \xrightarrow[n \to +\infty]{q-vaguely} \Pi(\cdot|x)$$



FROM PRIOR TO POSTERIOR CONVERGENCE

If the likelihood $f(x|\theta)$ is continuous in θ, the q-vague convergence is transmitted to the posterior.

$$\Pi_n \xrightarrow[n\to+\infty]{q-vaguely} \Pi \implies \Pi_n(\cdot|x) \xrightarrow[n\to+\infty]{q-vaguely} \Pi(\cdot|x)$$



if Π_n are probabilities and $\Pi(\cdot|x)$ is proper

Proposition

Let Π_n be probability measures which converges to an improper prior Π, then Π_n tends to concentrate outside any compact set, i.e.
for any compact K,

$$\Pi_n(K) \xrightarrow[n\to+\infty]{} 0$$

Application: This property also holds for prior predictive probabilities :
If

$$p(x) = \int f(x|\theta)\pi(\theta)\, d\theta$$

is defined, then it is improper and therefore the sequence of probabilities

$$p_n(x) = \int f(x|\theta)\pi_n(\theta)\, d\theta$$

that converges q-vaguely to an improper prior and therefore (under mild assumptions).

Therefore, x generated by $p_n(x)$ will be almost always very far from the data, even in dimension 1 and ABC methods may be inefficient.

Finitely additive probability(FAP)

Bayesian inference with improper prior may be legitimated by the use of FAP (see De Finetti, 1975, Heath and Sudderth 1978, Schervish, Seidenfeld and Kadane, 1984).

The idea is to remove the assumption of countably additivity from Kolmogorov axiomatic, which is

$$\Pi(\bigcup_{n\in\mathbb{N}} \uparrow A_n) = \lim_n \Pi(A_n).$$

A FAP is usually defined implicitly from the posterior distribution related to an improper prior.

There is **no uniqueness** and the proof of existence needs the axiom of choice.

We try to define a limiting FAP directly from the prior sequence :

Let Π_n be a sequence of probabilities. A FAP Π is a limit if

$$\liminf_n \Pi_n(f) \leq \Pi(f) \leq \limsup_n \Pi_n(f), \quad \forall f \in \mathcal{C}_b.$$

Proposition

If Π_n converge q-vaguely to an improper prior and if a FAP limit is Π then, Π is a purely FAP (PFAP) .

i.e. let K_m be an increasing sequence that converges to θ, then

$$\Pi(\bigcup_{n\in\mathbb{N}} \uparrow K_n) = 1 \neq 0 \lim_n \Pi(K_n).$$

For example a PFAP on $\mathbb{N}$ satisfies $\Pi(\{i\}) = 0$ whereas $\Pi(\mathbb{N}) = 1$

The Jeffreys-Lindley paradox

We now explore the Jeffreys-Lindley paradox

We must distinguish between two Bayesian paradigms

P1 The prior may be considered as the way to draw the parameter (subjective approach).

- If we change the prior, we change the marginal model and the way to generate x. Therefore it is not relevant to consider the behavior of the posterior for x fixed.
- We can applied probability rules.
- It is relevant to consider the joint distribution (X, θ) or the prior predictive probability $p(x)$.
- Improper priors have to be banished

P2 The prior is way to make inference (objective approach).

- If we change the prior, it is relevant to consider the behavior of the posterior for x fixed.
- We cannot applied all the probability rules.
  "One should not interpret any non-subjective prior as a probability distribution" (Bernardo, 1997)
- It is not relevant to consider the joint distribution (X, θ) or the prior predictive probability $p(x)$
- Improper priors are part of the paradigm, at least has limit of proper prior.
- PSM seems more appropriate.

Let $X|\theta \sim \mathcal{N}(\theta, 1)$
and the point null hypothesis $H_0 : \theta = 0$ tested against $H_1 : \theta \neq 0$.
Consider the prior

$$\Pi = \frac{1}{2}\delta_0 + \frac{1}{2}\lambda$$

then

$$\Pi(\theta = 0|x) \leqslant \left[1 + \sqrt{2\pi}\right]^{-1} \approx 0.285$$

whatever the data are.
Moreover, the is no reason to choose λ instead of $\alpha\lambda$ with $\alpha > 0$. The posterior distribution depends on α !
So, let replace λ by an approximation

$$\Pi_n = \frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n)$$

We have now

$$\Pi_n(\theta = 0|x) \xrightarrow[n \to +\infty]{} 1$$

whatever the data are.
What is wrong ?

- The convergence is for x fixed, so we are in the second paradigm.
- The prior $\Pi = \frac{1}{2}\delta_0 + \frac{1}{2}\lambda$ involves an improper prior.

$\rightarrow$ you are in a PSM.

A convex combination of priors is not compatible with the PSM !

$\rightarrow$ You cannot consider it as a combination of separate prior on H_0 and H_1.
This is also true for $\frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n)$.

- In fact we have $\frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n)$ is not an approximation of $\Pi = \frac{1}{2}\delta_0 + \frac{1}{2}\lambda$, but

$$\frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n) \xrightarrow[n \to +\infty]{\text{q-vaguely}} \delta_0$$

So $\Pi_n(\theta = 0|x) \xrightarrow[n \to +\infty]{} 1$ is not surprising (but does not hold systematically).

- The convergence is for x fixed, so we are in the second paradigm.
- The prior $\Pi = \frac{1}{2}\delta_0 + \frac{1}{2}\lambda$ involves an improper prior.

$\rightarrow$ you are in a PSM.

A convex combination of priors is not compatible with the PSM !

$\rightarrow$ You cannot consider it as a combination of separate prior on H_0 and H_1.
This is also true for $\frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n)$.

- In fact we have $\frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n)$ is not an approximation of $\Pi = \frac{1}{2}\delta_0 + \frac{1}{2}\lambda$, but

$$\frac{\mathbf{1}}{\mathbf{2}}\delta_{\mathbf{0}} + \frac{\mathbf{1}}{\mathbf{2}}\mathcal{N}(\mathbf{0}, \mathbf{n}) \xrightarrow[\mathbf{n}\to+\infty]{\mathbf{q-vaguely}} \delta_{\mathbf{0}}$$

So $\Pi_n(\theta = 0|x) \xrightarrow[n\to+\infty]{} 1$ is not surprising (but does not hold systematically).

Remark:

$$\frac{1}{2\sqrt{2\pi n}}\delta_0 + \frac{1}{2}\mathcal{N}(0,n) \xrightarrow[n\to+\infty]{} \frac{1}{2}\delta_0 + \frac{1}{2}\lambda$$

This kind of prior was also considered by C.P. Robert (1993, Stat. Sinica)

More generally

Proposition

Consider

- $\Theta = \{\theta_0\} \bigcup \Theta_1$
- $\Pi_n = \rho\delta_{\theta_0} + (1-\rho)\,\widetilde{\Pi}_n$ *where* $0 < \rho < 1$,
- $\widetilde{\Pi}_n$ *a probability on* Θ_1
- $\widetilde{\Pi}_n \xrightarrow{q-vaguely} \widetilde{\Pi}$ *improper*

Then

$$\Pi_n \xrightarrow{q-vaguely} \delta_{\theta_0}$$

.

Moreover, if $\theta \longmapsto f(x|\theta)$ *is continuous and belongs to* $\mathcal{C}_0(\Theta)$, *then*

$$\Pi_n(\theta = \theta_0|x) \xrightarrow{weakly} 1.$$

A NON-PARADOX

We consider now that we are in **paradigm 1** : when we change the prior, the way x is generated change.
Consider again

$$\Pi_n = \frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0, n)$$

Under the alternative hypothesis, the prior predictive probability

$$p_n^{(1)}(x) = \int_{H_1} f(x|\theta)\pi_n(\theta)d\theta.$$

is a $\mathcal{N}(0, 1+n)$ which converge q-vaguely to the Lebesgue measure.

Therefore $p_n^{(1)}(x)$ **tends to concentrate outside any given compact set**.

For any fixed $\alpha \approx 0$, define

$$I_\alpha(n) = \left[\,-u\sqrt{n}\,;\; u\sqrt{n}\,\right]$$

where u is the quantile of a $\mathcal{N}(0,1)$ of order $1-\alpha/2$. Then

$$\mathbb{P}_{H_1}(X \in I_\alpha(n)) = \alpha \approx 0$$

<u>For $x \notin I_\alpha(n)$</u>

$$\mathbb{P}(\theta = 0|x) \leq \left(1 + \sqrt{\frac{1}{1+n}} e^{u\,n/2}\right)^{-1} \xrightarrow{n\to+\infty} 0$$

and there is no paradox !

Another Jeffreys-Lindley paradox

(Dauxois, D. and Pommeret, 2006)
Consider a sample $X_1, ..., X_n$.
We have to choice a model between Poisson, Binomial(N,p) or Negative Binomial(N,p).
We assume that the mean m is known.
The distributions are characterized by their variance functions

$$V(m) = am^2 + m$$

where :

- $a = 0$ for the Poisson distribution.
- $a = -1/N$ for the Binomial Distribution.
- $a = 1/N$ for the negative Binomial Distribution.

So, we have to decide if "$a = 0$", "$a < 0$" or "$a > 0$".
We put a **truncated flat prior** for the parameter N for the Bin. and Neg. Bin. models.
We put probabilities p_0, p_- and p_+ on each model,

Another Jeffreys-Lindley paradox

(Dauxois, D. and Pommeret, 2006)
Consider a sample $X_1, ..., X_n$.
We have to choice a model between Poisson, Binomial(N,p) or Negative Binomial(N,p).
We assume that the mean m is known.
The distributions are characterized by their variance functions

$$V(m) = am^2 + m$$

where :

- $a = 0$ for the Poisson distribution.
- $a = -1/N$ for the Binomial Distribution.
- $a = 1/N$ for the negative Binomial Distribution.

So, we have to decide if "$a = 0$", "$a < 0$" or "$a > 0$".
We put a **truncated flat prior** for the parameter N for the Bin. and Neg. Bin. models.
We put probabilities p_0, p_- and p_+ on each model,

Thank you for your attention !