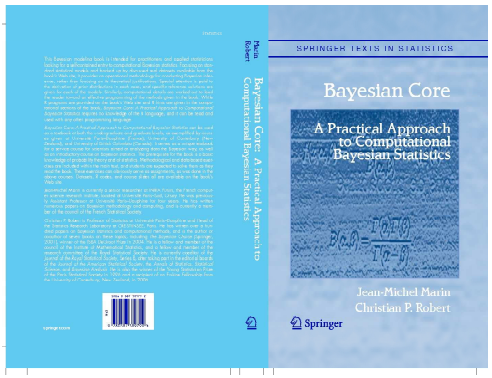
Bayesian Core: A Practical Approach to Computational Bayesian Statistics

Jean-Michel Marin & Christian P. Robert

October 24, 2007



Outline

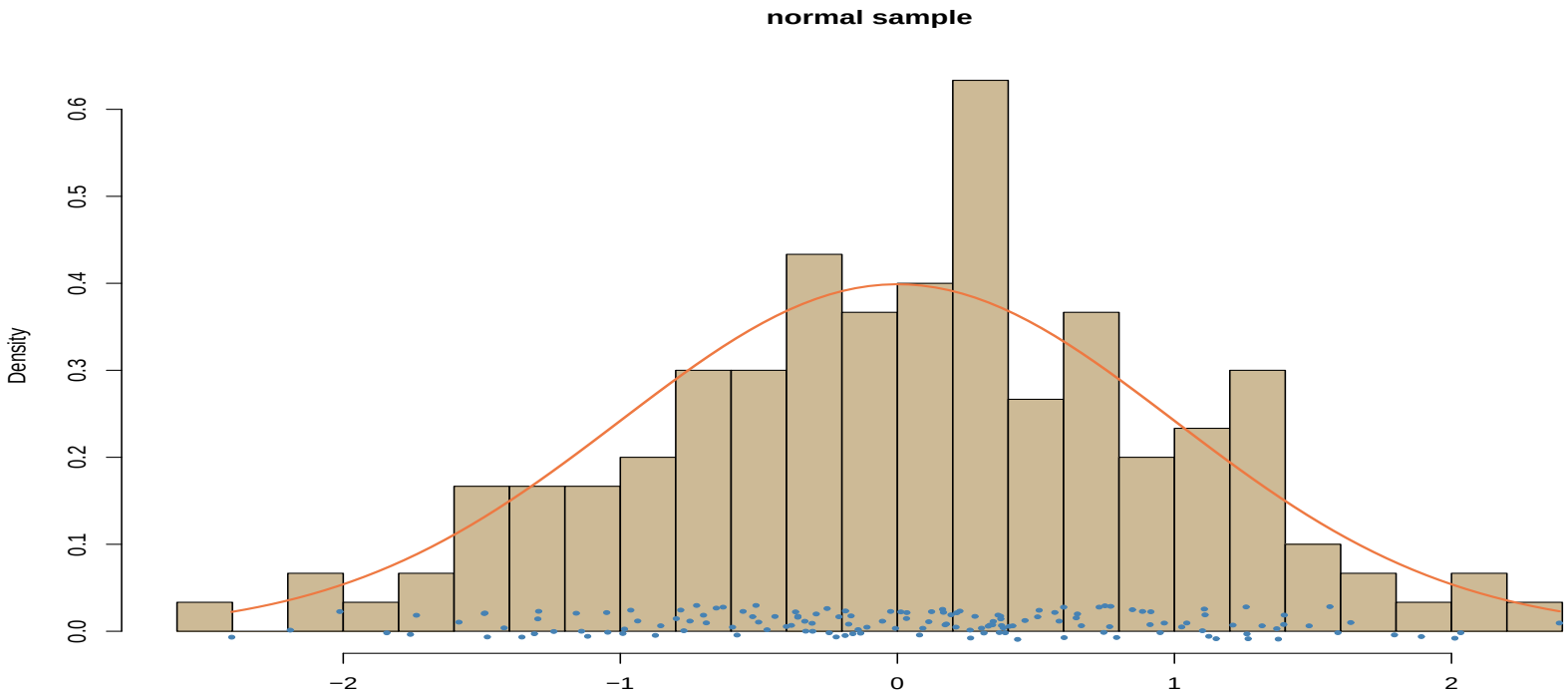
The normal model

Normal model

Sample

$$x_1, \ldots, x_n$$

from a normal $\mathcal{N}(\mu, \sigma^2)$ distribution



Inference on (μ, σ) based on this sample

- Estimation of [transforms of] (μ, σ)

Inference on (μ, σ) based on this sample

- Estimation of [transforms of] (μ, σ)
- Confidence region [interval] on (μ, σ)

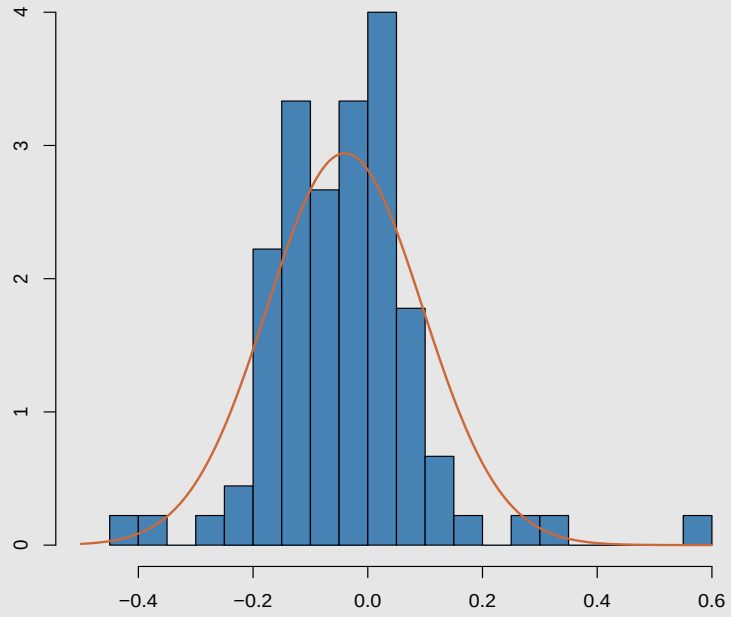
Inference on (μ, σ) based on this sample

- Estimation of [transforms of] (μ, σ)
- Confidence region [interval] on (μ, σ)
- Test on (μ, σ) and comparison with other samples

Datasets

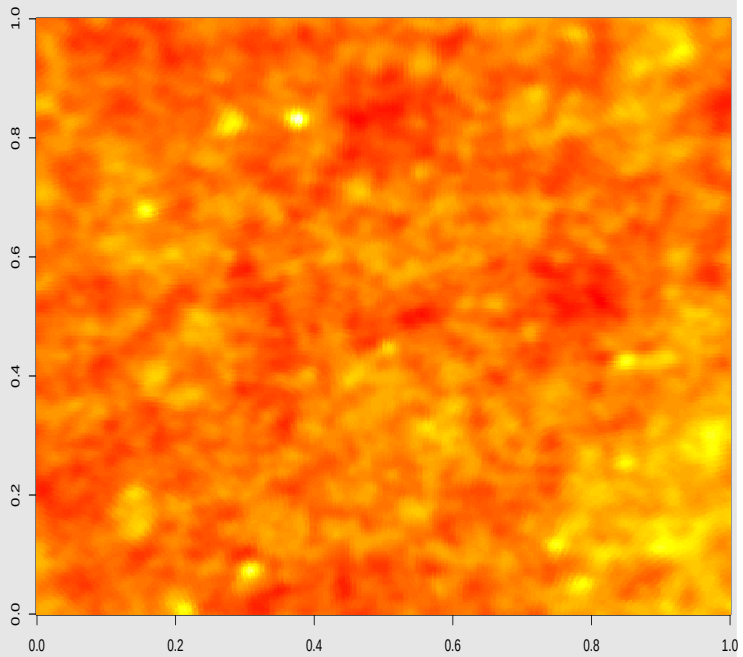
Larcenies =normaldata

Relative changes in reported larcenies between 1991 and 1995 (relative to 1991) for the 90 most populous US counties *(Source: FBI)*



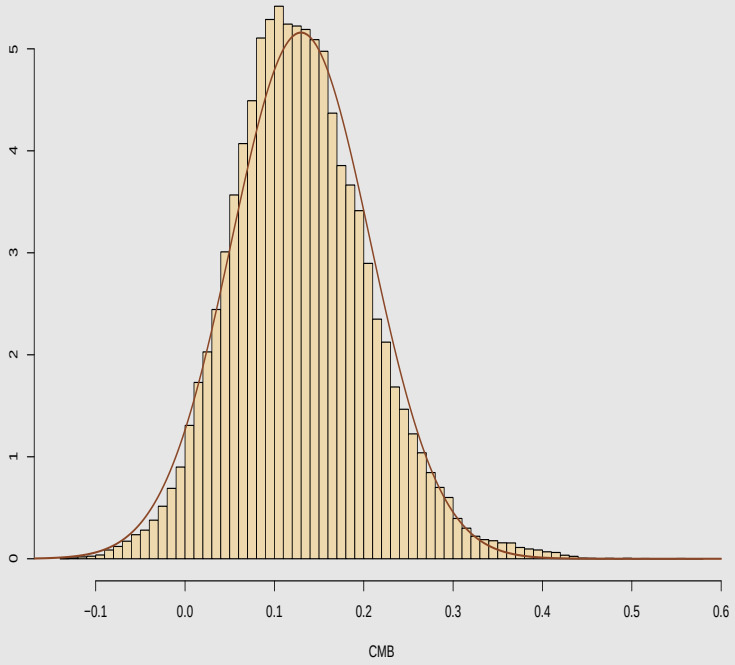
Cosmological background =CMBdata

Spectral representation of the "cosmological microwave background" (CMB), i.e. electromagnetic radiation from photons back to $300,000$ years after the Big Bang, expressed as difference in apparent temperature from the mean temperature



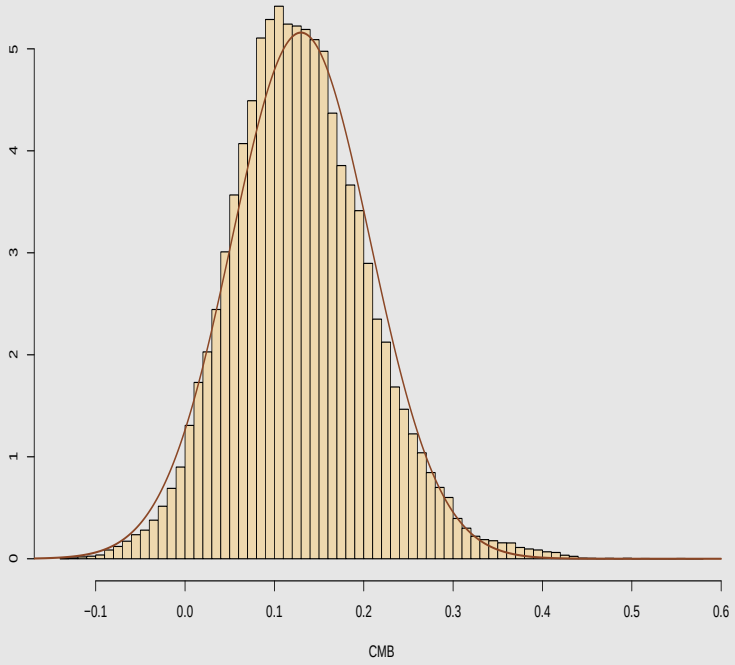
Cosmological background =CMBdata

Normal estimation



Cosmological background =CMBdata

Normal estimation



The Bayesian toolbox

Bayes theorem = Inversion of probabilities

The Bayesian toolbox

Bayes theorem = Inversion of probabilities

If A and E are events such that $P(E) \neq 0$, $P(A|E)$ and $P(E|A)$ are related by

$$\begin{aligned} P(A|E) &= \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)} \\ &= \frac{P(E|A)P(A)}{P(E)} \end{aligned}$$

Who's Bayes?

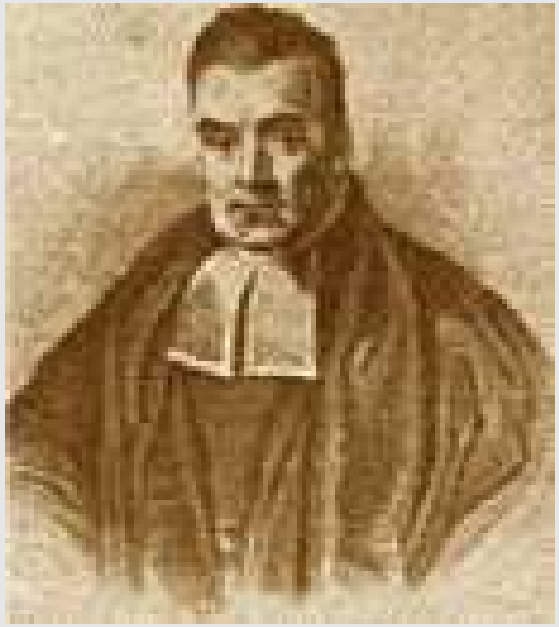
Reverend Thomas Bayes (ca. 1702–1761)

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.

Who's Bayes?

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*.

New perspective

- *Uncertainty* on the parameters θ of a model modeled through a *probability* distribution π on Θ, called *prior distribution*

New perspective

- *Uncertainty* on the parameters θ of a model modeled through a *probability* distribution π on Θ, called *prior distribution*
- *Inference* based on the distribution of θ conditional on x, $\pi(\theta|x)$, called *posterior distribution*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta}\,.$$

Bayesian model

A Bayesian statistical model is made of

1. a likelihood

$$f(x|\theta),$$

Bayesian model

A Bayesian statistical model is made of

1. a likelihood

$$f(x|\theta),$$

and of

2. a prior distribution on the parameters,

$$\pi(\theta)\,.$$

Justifications

- Semantic drift from unknown θ to random θ

Justifications

- Semantic drift from unknown θ to random θ
- Actualization of information/knowledge on θ by extracting information/knowledge on θ contained in the observation x

Justifications

- Semantic drift from unknown θ to random θ
- Actualization of information/knowledge on θ by extracting information/knowledge on θ contained in the observation x
- Allows incorporation of imperfect/imprecise information in the decision process

Justifications

- Semantic drift from unknown θ to random θ
- Actualization of information/knowledge on θ by extracting information/knowledge on θ contained in the observation x
- Allows incorporation of imperfect/imprecise information in the decision process
- Unique mathematical way to condition upon the observations (conditional perspective)

Example (Normal illustration $(\sigma^2 = 1)$)

Assume

$$\pi(\theta) = \exp\{-\theta\}\, \mathbb{I}_{\theta>0}$$

Example (Normal illustration $(\sigma^2 = 1)$)

Assume

$$\pi(\theta) = \exp\{-\theta\}\, \mathbb{I}_{\theta>0}$$

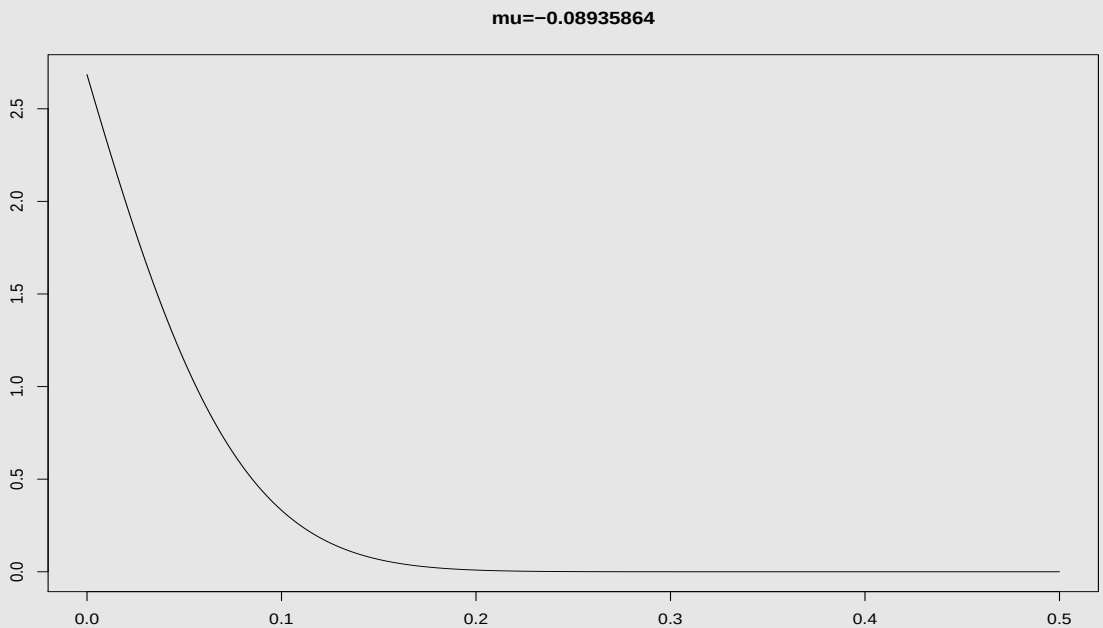
Then

$$\begin{aligned}
\pi(\theta|x_1,\ldots,x_n) &\propto \exp\{-\theta\}\, \exp\{-n(\theta-\overline{x})^2/2\}\, \mathbb{I}_{\theta>0} \\
&\propto \exp\left\{-n\theta^2/2 + \theta(n\overline{x}-1)\right\}\, \mathbb{I}_{\theta>0} \\
&\propto \exp\left\{-n(\theta-(\overline{x}-1/n))^2/2\right\}\, \mathbb{I}_{\theta>0}
\end{aligned}$$

Example (Normal illustration (2))

Truncated normal distribution

$$\mathcal{N}^+((\overline{x} - 1/n), 1/n)$$



Prior and posterior distributions

Given $f(x|\theta)$ and $\pi(\theta)$, several distributions of interest:

1. the *joint distribution* of (θ, x),

$$\varphi(\theta, x) = f(x|\theta)\pi(\theta)\,;$$

Prior and posterior distributions

Given $f(x|\theta)$ and $\pi(\theta)$, several distributions of interest:

1. the *joint distribution* of (θ, x),

$$\varphi(\theta, x) = f(x|\theta)\pi(\theta)\,;$$

2. the *marginal distribution* of x,

$$\begin{aligned} m(x) &= \int \varphi(\theta, x)\, d\theta \\ &= \int f(x|\theta)\pi(\theta)\, d\theta\,; \end{aligned}$$

3. the *posterior distribution* of θ,

$$\begin{aligned}\pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\,d\theta} \\ &= \frac{f(x|\theta)\pi(\theta)}{m(x)}\,;\end{aligned}$$

4. the *predictive distribution* of y, when $y \sim g(y|\theta, x)$,

$$g(y|x) = \int g(y|\theta, x)\pi(\theta|x)d\theta\,.$$

Posterior distribution center of to Bayesian inference

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations

Posterior distribution center of to Bayesian inference

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information/knowledge and information brought by x

Posterior distribution center of to Bayesian inference

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information/knowledge and information brought by x
- Avoids averaging over the unobserved values of x

Posterior distribution center of to Bayesian inference

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information/knowledge and information brought by x
- Avoids averaging over the unobserved values of x
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected

Posterior distribution center of to Bayesian inference

$$\pi(\theta|x) \propto f(x|\theta)\,\pi(\theta)$$

- Operates conditional upon the observations
- Integrate simultaneously prior information/knowledge and information brought by x
- Avoids averaging over the unobserved values of x
- Coherent updating of the information available on θ, independent of the order in which i.i.d. observations are collected
- Provides a complete inferential scope and an unique motor of inference

Example (Normal-normal case)

Consider $x|\theta \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(a, 10)$.

$$\begin{aligned}
\pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \propto \exp\left(-\frac{(x-\theta)^2}{2} - \frac{(\theta-a)^2}{20}\right) \\
&\propto \exp\left(-\frac{11\theta^2}{20} + \theta(x + a/10)\right) \\
&\propto \exp\left(-\frac{11}{20}\left\{\theta - ((10x + a)/11)\right\}^2\right)
\end{aligned}$$

Example (Normal-normal case)

Consider $x|\theta \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(a, 10)$.

$$\begin{aligned}
\pi(\theta|x) &\propto f(x|\theta)\pi(\theta) \propto \exp\left(-\frac{(x-\theta)^2}{2} - \frac{(\theta-a)^2}{20}\right) \\
&\propto \exp\left(-\frac{11\theta^2}{20} + \theta(x + a/10)\right) \\
&\propto \exp\left(-\frac{11}{20}\left\{\theta - ((10x+a)/11)\right\}^2\right)
\end{aligned}$$

and

$$\theta|x \sim \mathcal{N}\left((10x+a)/11, 10/11\right)$$

Prior selection

The prior distribution is the key to Bayesian inference

Prior selection

The prior distribution is the key to Bayesian inference

But...

In practice, it seldom occurs that the available prior information is precise enough to lead to an exact determination of the prior distribution

Prior selection

The prior distribution is the key to Bayesian inference

But...

In practice, it seldom occurs that the available prior information is precise enough to lead to an exact determination of the prior distribution

There is no such thing as *the* prior distribution!

Strategies for prior determination

Ungrounded prior distributions produce unjustified posterior inference.
—Anonymous, ca. 2006

Strategies for prior determination

Ungrounded prior distributions produce unjustified posterior inference.
—Anonymous, ca. 2006

- Use a partition of Θ in sets (e.g., intervals), determine the probability of each set, and approach π by an *histogram*

Strategies for prior determination

Ungrounded prior distributions produce unjustified posterior inference.
—Anonymous, ca. 2006

- Use a partition of Θ in sets (e.g., intervals), determine the probability of each set, and approach π by an *histogram*
- Select significant elements of Θ, evaluate their respective likelihoods and deduce a likelihood curve proportional to π

Strategies for prior determination

Ungrounded prior distributions produce unjustified posterior inference.
—Anonymous, ca. 2006

- Use a partition of Θ in sets (e.g., intervals), determine the probability of each set, and approach π by an *histogram*
- Select significant elements of Θ, evaluate their respective likelihoods and deduce a likelihood curve proportional to π
- Use the *marginal distribution* of x,

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)\, d\theta$$

Strategies for prior determination

Ungrounded prior distributions produce unjustified posterior inference.
—Anonymous, ca. 2006

- Use a partition of Θ in sets (e.g., intervals), determine the probability of each set, and approach π by an *histogram*
- Select significant elements of Θ, evaluate their respective likelihoods and deduce a likelihood curve proportional to π
- Use the *marginal distribution* of x,

$$m(x) = \int_{\Theta} f(x|\theta)\pi(\theta)\, d\theta$$

- Empirical and *hierarchical* Bayes techniques

Conjugate priors

Specific parametric family with analytical properties

Conjugate prior

A family $\mathcal{F}$ of probability distributions on Θ is *conjugate* for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

Conjugate priors

Specific parametric family with analytical properties

Conjugate prior

A family $\mathcal{F}$ of probability distributions on Θ is *conjugate* for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

Only of interest when $\mathcal{F}$ is *parameterised* : switching from prior to posterior distribution is reduced to an updating of the corresponding parameters.

Justifications

- Limited/finite information conveyed by x
- Preservation of the structure of $\pi(\theta)$

Justifications

- Limited/finite information conveyed by x
- Preservation of the structure of $\pi(\theta)$
- Exchangeability motivations
- Device of virtual past observations

Justifications

- Limited/finite information conveyed by x
- Preservation of the structure of $\pi(\theta)$
- Exchangeability motivations
- Device of virtual past observations
- Linearity of some estimators
- But mostly...

Justifications

- Limited/finite information conveyed by x
- Preservation of the structure of $\pi(\theta)$
- Exchangeability motivations
- Device of virtual past observations
- Linearity of some estimators
- But mostly... tractability and simplicity

Justifications

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

Exponential families

Sampling models of interest

Exponential family

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\}$$

is called an *exponential family of dimension* k. When $\Theta \subset \mathbb{R}^k$, $\mathcal{X} \subset \mathbb{R}^k$ and

$$f(x|\theta) = h(x)\exp\{\theta\cdot x - \Psi(\theta)\},$$

the family is said to be *natural*.

Analytical properties of exponential families

- Sufficient statistics (Pitman–Koopman Lemma)

Analytical properties of exponential families

- Sufficient statistics (Pitman–Koopman Lemma)
- Common enough structure (normal, Poisson, &tc...)

Analytical properties of exponential families

- Sufficient statistics (Pitman–Koopman Lemma)
- Common enough structure (normal, Poisson, &tc...)
- Analyticity ($\mathbb{E}[x] = \nabla\Psi(\theta)$, ...)

Analytical properties of exponential families

- Sufficient statistics (Pitman–Koopman Lemma)
- Common enough structure (normal, Poisson, &tc...)
- Analyticity ($\mathbb{E}[x] = \nabla\Psi(\theta)$, ...)
- Allow for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta\cdot\mu - \lambda\Psi(\theta)} \qquad \lambda > 0$$

Standard exponential families

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
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 $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + x, \beta + n - x)$

More...

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

Linearity of the posterior mean

If

$$\theta \sim \pi_{\lambda,\mu}(\theta) \propto e^{\theta\cdot\mu-\lambda\Psi(\theta)}$$

with $\mu \in \mathcal{X}$, then

$$\mathbb{E}^{\pi}[\nabla\Psi(\theta)] = \frac{\mu}{\lambda}.$$

where $\nabla\Psi(\theta) = (\partial\Psi(\theta)/\partial\theta_1, \ldots, \partial\Psi(\theta)/\partial\theta_p)$

Linearity of the posterior mean

If

$$\theta \sim \pi_{\lambda,\mu}(\theta) \propto e^{\theta\cdot\mu-\lambda\Psi(\theta)}$$

with $\mu \in \mathcal{X}$, then

$$\mathbb{E}^{\pi}[\nabla\Psi(\theta)] = \frac{\mu}{\lambda}.$$

where $\nabla\Psi(\theta) = (\partial\Psi(\theta)/\partial\theta_1, \ldots, \partial\Psi(\theta)/\partial\theta_p)$
Therefore, if $x_1, \ldots, x_n$ are i.i.d. $f(x|\theta)$,

$$\mathbb{E}^{\pi}[\nabla\Psi(\theta)|x_1, \ldots, x_n] = \frac{\mu + n\bar{x}}{\lambda + n}.$$

Example (Normal-normal)

In the normal $\mathcal{N}(\theta, \sigma^2)$ case, conjugate also normal $\mathcal{N}(\mu, \tau^2)$ and

$$\mathbb{E}^{\pi}[\nabla\Psi(\theta)|x] = \mathbb{E}^{\pi}[\theta|x] = \rho(\sigma^2\mu + \tau^2 x)$$

where

$$\rho^{-1} = \sigma^2 + \tau^2$$

Example (Full normal)

In the normal $\mathcal{N}(\mu, \sigma^2)$ case, when both μ and σ are unknown, there still is a conjugate prior on $\theta = (\mu, \sigma^2)$, of the form

$$(\sigma^2)^{-\lambda_\sigma} \exp - \left\{\lambda_\mu(\mu - \xi)^2 + \alpha\right\} / 2\sigma^2$$

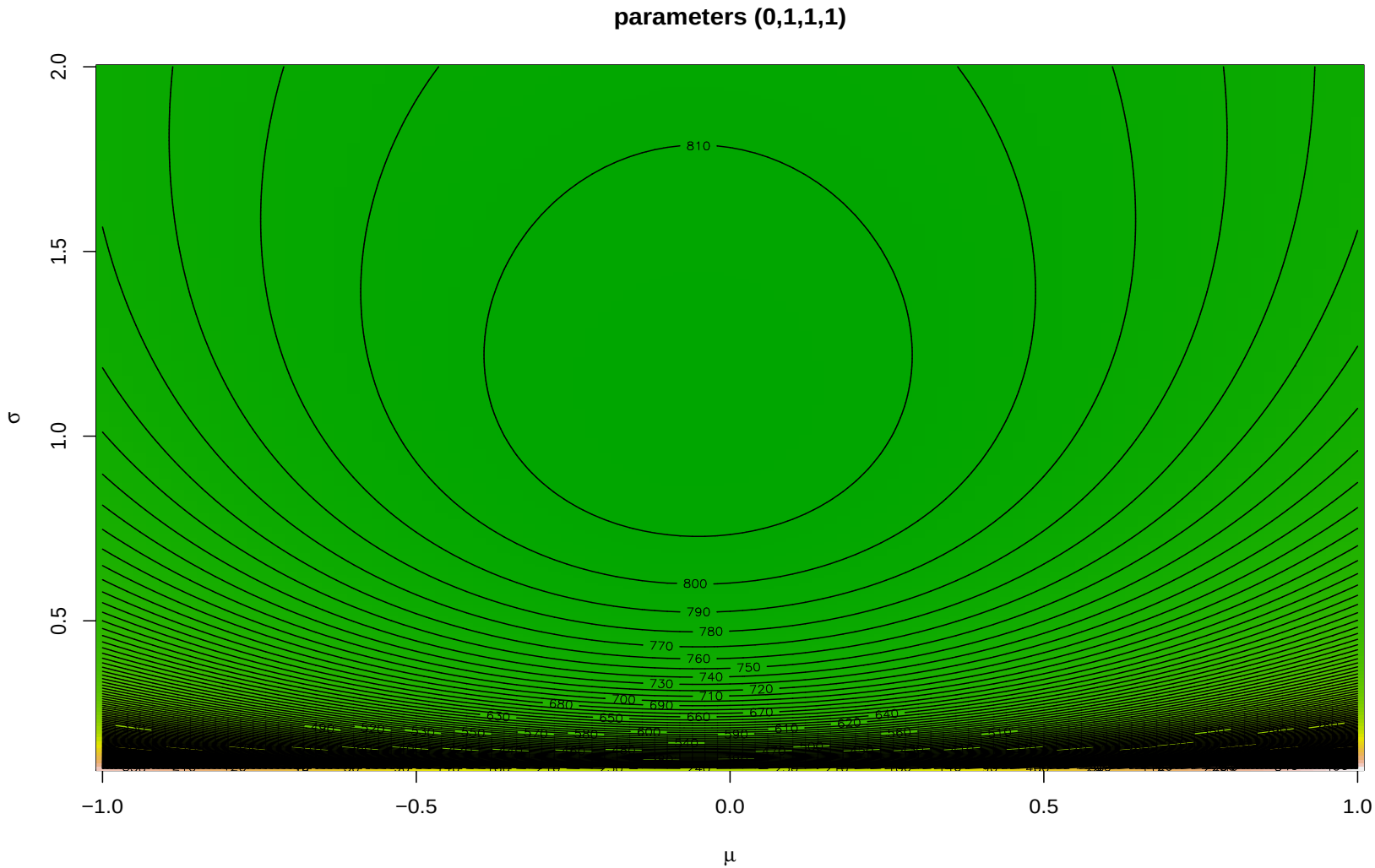
Example (Full normal)

In the normal $\mathcal{N}(\mu, \sigma^2)$ case, when both μ and σ are unknown, there still is a conjugate prior on $\theta = (\mu, \sigma^2)$, of the form

$$(\sigma^2)^{-\lambda_\sigma} \exp - \left\{\lambda_\mu(\mu - \xi)^2 + \alpha\right\} / 2\sigma^2$$

since

$$\begin{aligned}
\pi(\mu, \sigma^2 | x_1, \ldots, x_n) &\propto (\sigma^2)^{-\lambda_\sigma} \exp - \left\{\lambda_\mu(\mu - \xi)^2 + \alpha\right\} / 2\sigma^2 \\
&\quad \times (\sigma^2)^{-n/2} \exp - \left\{n(\mu - \overline{x})^2 + s_x^2\right\} / 2\sigma^2 \\
&\propto (\sigma^2)^{-\lambda_\sigma - n/2} \exp - \Bigg\{ (\lambda_\mu + n)(\mu - \xi_x)^2 \\
&\quad + \alpha + s_x^2 + \frac{n\lambda_\mu(\overline{x} - \xi)^2}{n + \lambda_\mu} \Bigg\} / 2\sigma^2
\end{aligned}$$


parameters (0,1,1,1)
σ
μ
2.0
1.5
1.0
0.5
−1.0
−0.5
0.0
0.5
1.0
810
800
790
780
770
760
750
740
730
720
710
700
690
680
670
660
650
640
630
620
610

Improper prior distribution

Extension from a prior distribution to a prior σ-finite measure π such that

$$\int_\Theta \pi(\theta)\, d\theta = +\infty$$

Improper prior distribution

Extension from a prior distribution to a prior σ-finite measure π such that

$$\int_\Theta \pi(\theta)\, d\theta = +\infty$$

Formal extension: π cannot be interpreted as a probability any longer

Justifications

1. Often only way to derive a prior in noninformative/automatic settings

Justifications

1. Often only way to derive a prior in noninformative/automatic settings
2. Performances of associated estimators usually good

Justifications

1. Often only way to derive a prior in noninformative/automatic settings
2. Performances of associated estimators usually good
3. Often occur as limits of proper distributions

Justifications

1. Often only way to derive a prior in noninformative/automatic settings
2. Performances of associated estimators usually good
3. Often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior

Justifications

1. Often only way to derive a prior in noninformative/automatic settings
2. Performances of associated estimators usually good
3. Often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Improper priors (infinitely!) preferable to vague proper priors such as a $\mathcal{N}(0, 100^2)$ distribution [e.g., BUGS]

Validation

Extension of the posterior distribution $\pi(\theta|x)$ associated with an improper prior π given by Bayes's formula

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta},$$

when

$$\int_\Theta f(x|\theta)\pi(\theta)\,d\theta < \infty$$

Example (Normal+improper)

If $x \sim \mathcal{N}(\theta, 1)$ and $\pi(\theta) = \varpi$, constant, the pseudo marginal distribution is

$$m(x) = \varpi \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(x-\theta)^2/2\right\} d\theta = \varpi$$

and the posterior distribution of θ is

$$\pi(\theta \mid x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\theta)^2}{2}\right\},$$

i.e., corresponds to $\mathcal{N}(x, 1)$.

[independent of ϖ]

Meaningless as probability distribution

The mistake is to think of them [the non-informative priors] as representing ignorance
—Lindley, 1990—

Meaningless as probability distribution

The mistake is to think of them [the non-informative priors] as representing ignorance
—Lindley, 1990—

Example

Consider a $\theta \sim \mathcal{N}(0, \tau^2)$ prior. Then

$$P^{\pi}\left(\theta \in [a, b]\right) \longrightarrow 0$$

when $\tau \longrightarrow \infty$ for any (a, b)

Noninformative prior distributions

What if all we know is that we know "nothing" ?!

Noninformative prior distributions

What if all we know is that we know "nothing" ?!

In the absence of prior information, prior distributions solely derived from the sample distribution $f(x|\theta)$

Noninformative prior distributions

What if all we know is that we know "nothing" ?!

In the absence of prior information, prior distributions solely derived from the sample distribution $f(x|\theta)$

> *Noninformative priors cannot be expected to represent exactly total ignorance about the problem at hand, but should rather be taken as reference or default priors, upon which everyone could fall back when the prior information is missing.*
> *—Kass and Wasserman, 1996—*

Laplace's prior

Principle of *Insufficient Reason* (Laplace)

$$\Theta = \{\theta_1, \cdots, \theta_p\} \qquad \pi(\theta_i) = 1/p$$

Laplace's prior

Principle of *Insufficient Reason* (Laplace)

$$\Theta = \{\theta_1, \cdots, \theta_p\} \qquad \pi(\theta_i) = 1/p$$

Extension to continuous spaces

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

[Lebesgue measure]

Who's Laplace?

Pierre Simon de Laplace (1749–1827)

French mathematician and astronomer born in Beaumont en Auge (Normandie) who formalised mathematical astronomy in *Mécanique Céleste*. Survived the French revolution, the Napoleon Empire (as a comte!), and the Bourbon restauration (as a marquis!!).

Who's Laplace?

Pierre Simon de Laplace (1749–1827)

French mathematician and astronomer born in Beaumont en Auge (Normandie) who formalised mathematical astronomy in *Mécanique Céleste*. Survived the French revolution, the Napoleon Empire (as a comte!), and the Bourbon restauration (as a marquis!!).

In *Essai Philosophique sur les Probabilités*, Laplace set out a mathematical system of inductive reasoning based on probability, precursor to Bayesian Statistics.

Laplace's problem

- Lack of reparameterization invariance/coherence

$$\pi(\theta) \propto 1, \quad \text{and} \quad \psi = e^{\theta} \quad \pi(\psi) = \frac{1}{\psi} \neq 1 \quad (!!)$$

Laplace's problem

- Lack of reparameterization invariance/coherence

$$\pi(\theta) \propto 1, \quad \text{and} \quad \psi = e^\theta \quad \pi(\psi) = \frac{1}{\psi} \neq 1 \quad (!!)$$

- Problems of properness

$$x \sim \mathcal{N}(\mu, \sigma^2), \qquad \pi(\mu, \sigma) = 1$$

$$\begin{array}{rrcl} & \pi(\mu, \sigma|x) & \propto & e^{-(x-\mu)^2/2\sigma^2}\sigma^{-1} \\ \Rightarrow & \pi(\sigma|x) & \propto & 1 \qquad (!!!) \end{array}$$

Jeffreys' prior

Based on Fisher information

$$I^F(\theta) = \mathbb{E}_\theta \left[\frac{\partial \log \ell}{\partial \theta^t} \frac{\partial \log \ell}{\partial \theta} \right]$$

Ron Fisher (1890–1962)

Jeffreys' prior

Based on Fisher information

$$I^F(\theta) = \mathbb{E}_\theta \left[\frac{\partial \log \ell}{\partial \theta^t} \frac{\partial \log \ell}{\partial \theta} \right]$$

Ron Fisher (1890–1962)

the Jeffreys prior distribution is

$$\pi^J(\theta) \propto |I^F(\theta)|^{1/2}$$

Who's Jeffreys?

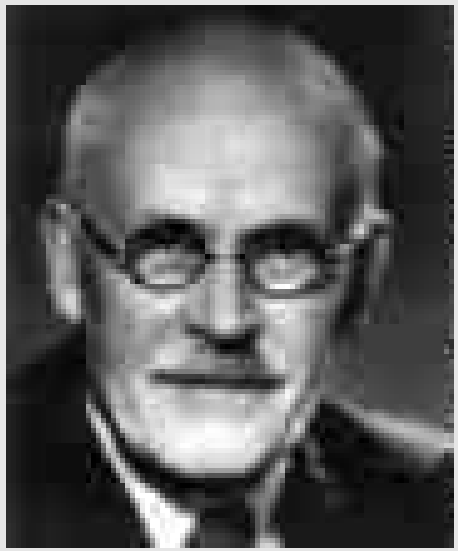
Sir Harold Jeffreys (1891–1989)

English mathematician, statistician, geophysicist, and astronomer. Founder of English Geophysics & originator of the theory that the Earth core is liquid.

Who's Jeffreys?

Sir Harold Jeffreys (1891–1989)

English mathematician, statistician, geophysicist, and astronomer. Founder of English Geophysics & originator of the theory that the Earth core is liquid.
Formalised Bayesian methods for the analysis of geophysical data and ended up writing *Theory of Probability*



Pros & Cons

- Relates to information theory

Pros & Cons

- Relates to information theory
- Agrees with most invariant priors

Pros & Cons

- Relates to information theory
- Agrees with most invariant priors
- Parameterization invariant

Pros & Cons

- Relates to information theory
- Agrees with most invariant priors
- Parameterization invariant
- Suffers from dimensionality curse

Evaluating estimators

Purpose of most inferential studies: to provide the statistician/client with a *decision* $d \in \mathcal{D}$

Evaluating estimators

Purpose of most inferential studies: to provide the statistician/client with a *decision* $d \in \mathcal{D}$
Requires an evaluation criterion/loss function for decisions and estimators

$$\mathrm{L}(\theta, d)$$

Evaluating estimators

Purpose of most inferential studies: to provide the statistician/client with a *decision* $d \in \mathcal{D}$
Requires an evaluation criterion/loss function for decisions and estimators

$$\mathrm{L}(\theta, d)$$

There exists an axiomatic derivation of the existence of a loss function.

[DeGroot, 1970]

Loss functions

Decision procedure δ^π usually called estimator
(while its *value* $\delta^\pi(x)$ is called estimate of θ)

Loss functions

Decision procedure δ^π usually called estimator
(while its *value* $\delta^\pi(x)$ is called estimate of θ)

Impossible to uniformly minimize (in d) the loss function

$$\mathrm{L}(\theta, d)$$

when θ is unknown

Bayesian estimation

Principle Integrate over the space Θ to get the posterior expected loss

$$\begin{aligned} &= \mathbb{E}^{\pi}[\mathrm{L}(\theta, d)|x] \\ &= \int_{\Theta} \mathrm{L}(\theta, d)\pi(\theta|x)\, d\theta, \end{aligned}$$

and minimise in d

Bayes estimates

Bayes estimator

A *Bayes estimate* associated with a prior distribution π and a loss function L is

$$\arg\min_d \mathbb{E}^\pi [\mathrm{L}(\theta, d)|x] \,.$$

The quadratic loss

Historically, first loss function (Legendre, Gauss, Laplace)

$$\mathrm{L}(\theta, d) = (\theta - d)^2$$

The quadratic loss

Historically, first loss function (Legendre, Gauss, Laplace)

$$\mathrm{L}(\theta, d) = (\theta - d)^2$$

The Bayes estimate $\delta^\pi(x)$ associated with the prior π and with the quadratic loss is the posterior expectation

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta|x] = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)\,d\theta}{\int_\Theta f(x|\theta)\pi(\theta)\,d\theta}.$$

The absolute error loss

Alternatives to the quadratic loss:

$$\mathrm{L}(\theta, d) = \mid \theta - d \mid,$$

or

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d, \\ k_1(d - \theta) & \text{otherwise.} \end{cases}$$

The absolute error loss

Alternatives to the quadratic loss:

$$\mathrm{L}(\theta, d) = \mid \theta - d \mid,$$

or

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d, \\ k_1(d - \theta) & \text{otherwise.} \end{cases}$$

Associated Bayes estimate is $(k_2/(k_1 + k_2))$ fractile of $\pi(\theta|x)$

MAP estimator

With no loss function, consider using the maximum a posteriori (MAP) estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

MAP estimator

With no loss function, consider using the maximum a posteriori (MAP) estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

- Penalized likelihood estimator

MAP estimator

With no loss function, consider using the maximum a posteriori (MAP) estimator

$$\arg\max_{\theta} \ell(\theta|x)\pi(\theta)$$

- Penalized likelihood estimator
- Further appeal in restricted parameter spaces

Example (Binomial probability)

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors:

$$\pi^J(\theta) = \frac{1}{B(1/2, 1/2)} \theta^{-1/2}(1-\theta)^{-1/2},$$

$$\pi_1(\theta) = 1 \quad \text{and} \quad \pi_2(\theta) = \theta^{-1}(1-\theta)^{-1}.$$

Example (Binomial probability)

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.
Possible priors:

$$\pi^J(\theta) = \frac{1}{B(1/2, 1/2)} \theta^{-1/2}(1-\theta)^{-1/2},$$

$$\pi_1(\theta) = 1 \quad \text{and} \quad \pi_2(\theta) = \theta^{-1}(1-\theta)^{-1}.$$

Corresponding MAP estimators:

$$\begin{aligned} \delta^{\pi^J}(x) &= \max\left(\frac{x-1/2}{n-1}, 0\right), \\ \delta^{\pi_1}(x) &= x/n, \\ \delta^{\pi_2}(x) &= \max\left(\frac{x-1}{n-2}, 0\right). \end{aligned}$$

Not always appropriate

Example (Fixed MAP)

Consider

$$f(x|\theta) = \frac{1}{\pi}\left[1+(x-\theta)^2\right]^{-1},$$

and $\pi(\theta) = \frac{1}{2}e^{-|\theta|}$.

Not always appropriate

Example (Fixed MAP)

Consider

$$f(x|\theta) = \frac{1}{\pi}\left[1+(x-\theta)^2\right]^{-1},$$

and $\pi(\theta) = \frac{1}{2}e^{-|\theta|}$. Then the MAP estimate of θ is always

$$\delta^\pi(x) = 0$$

Credible regions

Natural confidence region: Highest posterior density (HPD) region

$$C_\alpha^\pi = \{\theta; \pi(\theta|x) > k_\alpha\}$$

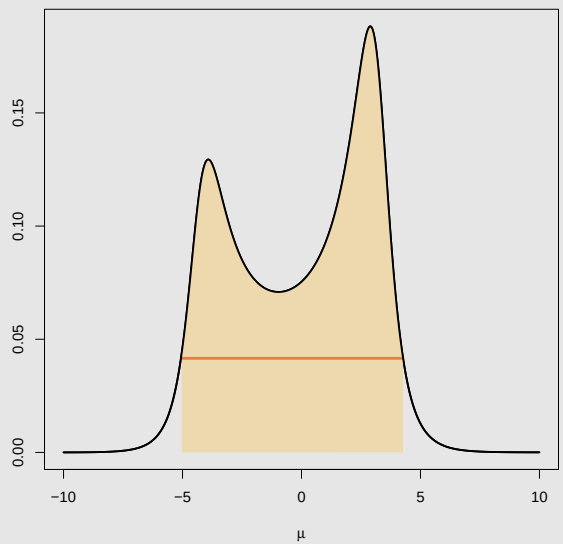
Credible regions

Natural confidence region: Highest posterior density (HPD) region

$$C_\alpha^\pi = \{\theta; \pi(\theta|x) > k_\alpha\}$$

Optimality

The HPD regions give the highest probabilities of containing θ for a given volume



Example

If the posterior distribution of θ is $\mathcal{N}(\mu(x), \omega^{-2})$ with $\omega^2 = \tau^{-2} + \sigma^{-2}$ and $\mu(x) = \tau^2 x/(\tau^2 + \sigma^2)$, then

$$C_\alpha^\pi = \left[\mu(x) - k_\alpha \omega^{-1}, \mu(x) + k_\alpha \omega^{-1}\right],$$

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

Example

If the posterior distribution of θ is $\mathcal{N}(\mu(x), \omega^{-2})$ with $\omega^2 = \tau^{-2} + \sigma^{-2}$ and $\mu(x) = \tau^2 x/(\tau^2 + \sigma^2)$, then

$$C_\alpha^\pi = \left[\mu(x) - k_\alpha \omega^{-1}, \mu(x) + k_\alpha \omega^{-1}\right],$$

where k_α is the $\alpha/2$-quantile of $\mathcal{N}(0,1)$.
If τ goes to $+\infty$,

$$C_\alpha^\pi = \left[x - k_\alpha \sigma, x + k_\alpha \sigma\right],$$

the "usual" (classical) confidence interval

Full normal

Under [almost!] *Jeffreys prior* prior

$$\pi(\mu, \sigma^2) = 1/\sigma^2,$$

posterior distribution of (μ, σ)

$$\begin{aligned} \mu|\sigma, \bar{x}, s_x^2 &\sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right), \\ \sigma^2|\bar{x}, s_x^2 &\sim \mathcal{IG}\left(\frac{n-1}{2}, \frac{s_x^2}{2}\right). \end{aligned}$$

Full normal

Under [almost!] *Jeffreys prior* prior

$$\pi(\mu, \sigma^2) = 1/\sigma^2,$$

posterior distribution of (μ, σ)

$$\begin{aligned} \mu|\sigma, \bar{x}, s_x^2 &\sim \mathcal{N}\left(\bar{x}, \frac{\sigma^2}{n}\right), \\ \sigma^2|\bar{x}, s_x^2 &\sim \mathcal{IG}\left(\frac{n-1}{2}, \frac{s_x^2}{2}\right). \end{aligned}$$

Then

$$\begin{aligned} \pi(\mu|\bar{x}, s_x^2) &\propto \int \omega^{1/2} \exp -\omega \frac{n(\bar{x}-\mu)^2}{2} \omega^{(n-3)/2} \exp\{-\omega s_x^2/2\} \, \mathrm{d}\omega \\ &\propto \left[s_x^2 + n(\bar{x}-\mu)^2\right]^{-n/2} \end{aligned}$$

[$\mathcal{T}_{n-1}$ distribution]

Normal credible interval

Derived credible interval on μ

$$[\bar{x} - t_{\alpha/2,n-1} s_x / \sqrt{n(n-1)}, \bar{x} + t_{\alpha/2,n-1} s_x / \sqrt{n(n-1)}]$$

Normal credible interval

Derived credible interval on μ

$$[\bar{x} - t_{\alpha/2,n-1} s_x / \sqrt{n(n-1)}, \bar{x} + t_{\alpha/2,n-1} s_x / \sqrt{n(n-1)}]$$

normaldata

Corresponding 95% confidence region for μ

$$[-0.070, -0.013,]$$

Since 0 does not belong to this interval, reporting a significant decrease in the number of larcenies between 1991 and 1995 is acceptable

Testing hypotheses

Deciding about validity of assumptions or restrictions on the parameter θ from the data, represented as

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0$$

Testing hypotheses

Deciding about validity of assumptions or restrictions on the parameter θ from the data, represented as

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0$$

Binary outcome of the decision process: *accept* [coded by 1] or *reject* [coded by 0]

$$\mathcal{D} = \{0, 1\}$$

Testing hypotheses

Deciding about validity of assumptions or restrictions on the parameter θ from the data, represented as

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0$$

Binary outcome of the decision process: *accept* [coded by 1] or *reject* [coded by 0]

$$\mathcal{D} = \{0, 1\}$$

Bayesian solution formally very close from a likelihood ratio test statistic, but numerical values often strongly differ from classical solutions

The $0-1$ loss

Rudimentary loss function

$$L(\theta, d) = \begin{cases} 1-d & \text{if } \theta \in \Theta_0 \\ d & \text{otherwise,} \end{cases}$$

Associated Bayes estimate

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } P^\pi(\theta \in \Theta_0 | x) > \dfrac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

The $0-1$ loss

Rudimentary loss function

$$L(\theta, d) = \begin{cases} 1-d & \text{if } \theta \in \Theta_0 \\ d & \text{otherwise,} \end{cases}$$

Associated Bayes estimate

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } P^\pi(\theta \in \Theta_0 | x) > \dfrac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Intuitive structure

Extension

Weighted $0-1$ (or a_0-a_1) loss

$$\mathrm{L}(\theta,d)=\begin{cases}0 & \text{if } d=\mathbb{I}_{\Theta_0}(\theta),\\ a_0 & \text{if } \theta\in\Theta_0 \text{ and } d=0,\\ a_1 & \text{if } \theta\notin\Theta_0 \text{ and } d=1,\end{cases}$$

Extension

Weighted $0-1$ (or a_0-a_1) loss

$$\mathrm{L}(\theta,d)=\begin{cases}0 & \text{if } d=\mathbb{I}_{\Theta_0}(\theta),\\ a_0 & \text{if } \theta\in\Theta_0 \text{ and } d=0,\\ a_1 & \text{if } \theta\notin\Theta_0 \text{ and } d=1,\end{cases}$$

Associated Bayes estimator

$$\delta^\pi(x)=\begin{cases}1 & \text{if } P^\pi(\theta\in\Theta_0|x)>\dfrac{a_1}{a_0+a_1},\\ 0 & \text{otherwise.}\end{cases}$$

Example (Normal-normal)

For $x \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu, \tau^2)$, $\pi(\theta|x)$ is $\mathcal{N}(\mu(x), \omega^2)$ with

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

Example (Normal-normal)

For $x \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu, \tau^2)$, $\pi(\theta|x)$ is $\mathcal{N}(\mu(x), \omega^2)$ with

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

To test $H_0 : \theta < 0$, we compute

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

Example (Normal-normal (2))

If z_{a_0,a_1} is the $a_1/(a_0+a_1)$ quantile, i.e.,

$$\Phi(z_{a_0,a_1}) = a_1/(a_0+a_1)\,,$$

H_0 is accepted when

$$-\mu(x) > z_{a_0,a_1}\omega,$$

the upper acceptance bound then being

$$x \leq -\frac{\sigma^2}{\tau^2}\mu - (1+\frac{\sigma^2}{\tau^2})\omega z_{a_0,a_1}\,.$$

Bayes factor

Bayesian testing procedure depends on $P^\pi(\theta \in \Theta_0|x)$ or alternatively on the **Bayes factor**

$$B_{10}^\pi = \frac{\{P^\pi(\theta \in \Theta_1|x)/P^\pi(\theta \in \Theta_0|x)\}}{\{P^\pi(\theta \in \Theta_1)/P^\pi(\theta \in \Theta_0)\}}$$

in the absence of loss function parameters a_0 and a_1

Associated reparameterisations

Corresponding models $\mathcal{M}_1$ vs. $\mathcal{M}_0$ compared via

$$B_{10}^{\pi} = \frac{P^{\pi}(\mathcal{M}_1|x)}{P^{\pi}(\mathcal{M}_0|x)} \Big/ \frac{P^{\pi}(\mathcal{M}_1)}{P^{\pi}(\mathcal{M}_0)}$$

Associated reparameterisations

Corresponding models $\mathcal{M}_1$ vs. $\mathcal{M}_0$ compared via

$$B_{10}^{\pi} = \frac{P^{\pi}(\mathcal{M}_1|x)}{P^{\pi}(\mathcal{M}_0|x)} \Big/ \frac{P^{\pi}(\mathcal{M}_1)}{P^{\pi}(\mathcal{M}_0)}$$

If we rewrite the prior as

$$\pi(\theta) = \mathsf{Pr}(\theta \in \Theta_1) \times \pi_1(\theta) + \mathsf{Pr}(\theta \in \Theta_0) \times \pi_0(\theta)$$

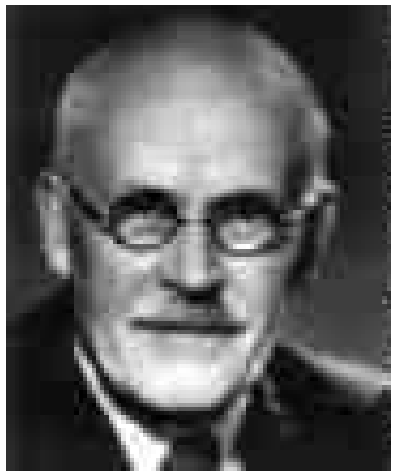
then

$$B_{10}^{\pi} = \int f(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1 \Big/ \int f(x|\theta_0)\pi_0(\theta_0)\mathrm{d}\theta_0 = m_1(x)/m_0(x)$$

[Akin to likelihood ratio]

Jeffreys' scale

1. if $\log_{10}(B_{10}^{\pi})$ varies between 0 and 0.5, the evidence against H_0 is *poor*,
2. if it is between 0.5 and 1, it is is *substantial*,
3. if it is between 1 and 2, it is *strong*, and
4. if it is above 2 it is *decisive*.



Point null difficulties

If π absolutely continuous,

$$P^{\pi}(\theta = \theta_0) = 0 \ldots$$

Point null difficulties

If π absolutely continuous,

$$P^{\pi}(\theta = \theta_0) = 0 \ldots$$

How can we test $H_0 : \theta = \theta_0$?!

New prior for new hypothesis

Testing point null difficulties requires a modification of the prior distribution so that

$$\pi(\Theta_0) > 0 \quad \text{and} \quad \pi(\Theta_1) > 0$$

(hidden information) or

$$\pi(\theta) = P^\pi(\theta \in \Theta_0) \times \pi_0(\theta) + P^\pi(\theta \in \Theta_1) \times \pi_1(\theta)$$

New prior for new hypothesis

Testing point null difficulties requires a modification of the prior distribution so that

$$\pi(\Theta_0) > 0 \quad \text{and} \quad \pi(\Theta_1) > 0$$

(hidden information) or

$$\pi(\theta) = P^\pi(\theta \in \Theta_0) \times \pi_0(\theta) + P^\pi(\theta \in \Theta_1) \times \pi_1(\theta)$$

[E.g., when $\Theta_0 = \{\theta_0\}$, π_0 is Dirac mass at θ_0]

Posteriors with Dirac masses

If $H_0 : \theta = \theta_0 \ (= \Theta_1)$,

$$\rho = P^\pi(\theta = \theta_0) \quad \text{and} \quad \pi(\theta) = \rho \mathbb{I}_{\theta_0}(\theta) + (1-\rho)\pi_1(\theta)$$

then

$$\begin{aligned} \pi(\Theta_0|x) &= \frac{f(x|\theta_0)\rho}{\int f(x|\theta)\pi(\theta)\, d\theta} \\ &= \frac{f(x|\theta_0)\rho}{f(x|\theta_0)\rho + (1-\rho)m_1(x)} \end{aligned}$$

with

$$m_1(x) = \int_{\Theta_1} f(x|\theta)\pi_1(\theta)\, d\theta.$$

Example (Normal-normal)

For $x \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(0, \tau^2)$, to test of $H_0 : \theta = 0$ requires a modification of the prior, with

$$\pi_1(\theta) \propto e^{-\theta^2/2\tau^2} \mathbb{I}_{\theta \neq 0}$$

and $\pi_0(\theta)$ the Dirac mass in 0

Example (Normal-normal)

For $x \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(0, \tau^2)$, to test of $H_0 : \theta = 0$ requires a modification of the prior, with

$$\pi_1(\theta) \propto e^{-\theta^2/2\tau^2} \mathbb{I}_{\theta \neq 0}$$

and $\pi_0(\theta)$ the Dirac mass in 0
Then

$$\begin{aligned} \frac{m_1(x)}{f(x|0)} &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}} \frac{e^{-x^2/2(\sigma^2+\tau^2)}}{e^{-x^2/2\sigma^2}} \\ &= \sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left\{ \frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)} \right\}, \end{aligned}$$

Example (cont'd)

and

$$\pi(\theta = 0|x) = \left[1 + \frac{1-\rho}{\rho}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right]^{-1}.$$

For $z = x/\sigma$ and $\rho = 1/2$:

z	0	0.68	1.28	1.96
$\pi(\theta = 0\|z, \tau = \sigma)$	0.586	0.557	0.484	0.351
$\pi(\theta = 0\|z, \tau = 3.3\sigma)$	0.768	0.729	0.612	0.366

Banning improper priors

Impossibility of using improper priors for testing!

Banning improper priors

Impossibility of using improper priors for testing!

Reason: When using the representation

$$\pi(\theta) = P^{\pi}(\theta \in \Theta_1) \times \pi_1(\theta) + P^{\pi}(\theta \in \Theta_0) \times \pi_0(\theta)$$

π_1 and π_0 must be normalised

Example (Normal point null)

When $x \sim \mathcal{N}(\theta, 1)$ and $H_0 : \theta = 0$, for the improper prior $\pi(\theta) = 1$, the prior is transformed as

$$\pi(\theta) = \frac{1}{2}\mathbb{I}_0(\theta) + \frac{1}{2} \cdot \mathbb{I}_{\theta \neq 0},$$

and

$$\begin{aligned} \pi(\theta = 0|x) &= \frac{e^{-x^2/2}}{e^{-x^2/2} + \int_{-\infty}^{+\infty} e^{-(x-\theta)^2/2}\, d\theta} \\ &= \frac{1}{1 + \sqrt{2\pi} e^{x^2/2}}. \end{aligned}$$

Example (Normal point null (2))

Consequence: H_0 is bounded from above by

$$\pi(\theta = 0|x) \leq 1/(1 + \sqrt{2\pi}) = 0.285$$

x	0.0	1.0	1.65	1.96	2.58
$\pi(\theta = 0\|x)$	0.285	0.195	0.089	0.055	0.014

Regular tests: Agreement with the classical p-value (but...)

Example (Normal one-sided)

For $x \sim \mathcal{N}(\theta, 1)$, $\pi(\theta) = 1$, and $H_0 : \theta \leq 0$ to test versus $H_1 : \theta > 0$

$$\pi(\theta \leq 0|x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-(x-\theta)^2/2}\, d\theta = \Phi(-x).$$

The generalized Bayes answer is also the *p-value*

Example (Normal one-sided)

For $x \sim \mathcal{N}(\theta, 1)$, $\pi(\theta) = 1$, and $H_0 : \theta \leq 0$ to test versus $H_1 : \theta > 0$

$$\pi(\theta \leq 0|x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-(x-\theta)^2/2}\, d\theta = \Phi(-x).$$

The generalized Bayes answer is also the *p-value*

normaldata

If $\pi(\mu, \sigma^2) = 1/\sigma^2$,

$$\pi(\mu \geq 0|x) = 0.0021$$

since $\mu|x \sim \mathcal{T}_{89}(-0.0144, 0.000206)$.

Jeffreys–Lindley paradox

Limiting arguments not valid in testing settings: Under a conjugate prior

$$\pi(\theta = 0|x) = \left\{1 + \frac{1-\rho}{\rho}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left[\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right]\right\}^{-1},$$

which converges to 1 when τ goes to $+\infty$, **for every** x

Jeffreys–Lindley paradox

Limiting arguments not valid in testing settings: Under a conjugate prior

$$\pi(\theta = 0|x) = \left\{1 + \frac{1-\rho}{\rho}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left[\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right]\right\}^{-1},$$

which converges to 1 when τ goes to $+\infty$, **for every** x

Difference with the "noninformative" answer

$$[1 + \sqrt{2\pi}\exp(x^2/2)]^{-1}$$

[↯ Invalid answer]

Normalisation difficulties

If g_0 and g_1 are σ-finite measures on the subspaces Θ_0 and Θ_1, the choice of the normalizing constants influences the Bayes factor:
If g_i replaced by $c_i g_i$ $(i = 0, 1)$, Bayes factor multiplied by c_0/c_1

Normalisation difficulties

If g_0 and g_1 are σ-finite measures on the subspaces Θ_0 and Θ_1, the choice of the normalizing constants influences the Bayes factor:
If g_i replaced by $c_i g_i$ $(i = 0, 1)$, Bayes factor multiplied by c_0/c_1

Example

If the Jeffreys prior is uniform and $g_0 = c_0$, $g_1 = c_1$,

$$\begin{aligned}\pi(\theta \in \Theta_0|x) &= \frac{\rho c_0 \int_{\Theta_0} f(x|\theta)\, d\theta}{\rho c_0 \int_{\Theta_0} f(x|\theta)\, d\theta + (1-\rho) c_1 \int_{\Theta_1} f(x|\theta)\, d\theta} \\ &= \frac{\rho \int_{\Theta_0} f(x|\theta)\, d\theta}{\rho \int_{\Theta_0} f(x|\theta)\, d\theta + (1-\rho)[\mathbf{c_1/c_0}] \int_{\Theta_1} f(x|\theta)\, d\theta}\end{aligned}$$

Monte Carlo integration

Generic problem of evaluating an integral

$$\mathfrak{I} = \mathbb{E}_f[h(X)] = \int_{\mathcal{X}} h(x)\, f(x)\, dx$$

where $\mathcal{X}$ is uni- or multidimensional, f is a closed form, partly closed form, or implicit density, and h is a function

Monte Carlo Principle

Use a sample $(x_1, \ldots, x_m)$ from the density f to approximate the integral $\mathfrak{I}$ by the empirical average

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(x_j)$$

Monte Carlo Principle

Use a sample $(x_1, \ldots, x_m)$ from the density f to approximate the integral $\mathfrak{I}$ by the empirical average

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h(x_j)$$

Convergence of the average

$$\overline{h}_m \longrightarrow \mathbb{E}_f[h(X)]$$

by the **Strong Law of Large Numbers**

Bayes factor approximation

For the normal case

$$\begin{aligned} x_1, \ldots, x_n &\sim \mathcal{N}(\mu + \xi, \sigma^2) \\ y_1, \ldots, y_n &\sim \mathcal{N}(\mu - \xi, \sigma^2) \\ \text{and} \qquad & H_0 : \xi = 0 \end{aligned}$$

under prior

$$\pi(\mu, \sigma^2) = 1/\sigma^2 \quad \text{and} \quad \xi \sim \mathcal{N}(0, 1)$$

$$B_{01}^{\pi} = \frac{\left[(\bar{x} - \bar{y})^2 + S^2\right]^{-n+1/2}}{\int \left[(2\xi - \bar{x} - \bar{y})^2 + S^2\right]^{-n+1/2} e^{-\xi^2/2} \, \mathrm{d}\xi / \sqrt{2\pi}}$$

Example

CMBdata

Simulate $\xi_1, \ldots, \xi_{1000} \sim \mathcal{N}(0,1)$ and approximate B_{01}^{π} with

$$\widehat{B_{01}^{\pi}} = \frac{\left[(\bar{x}-\bar{y})^2 + S^2\right]^{-n+1/2}}{\frac{1}{1000}\sum_{i=1}^{1000}\left[(2\xi_i - \bar{x} - \bar{y})^2 + S^2\right]^{-n+1/2}} = 89.9$$

when $\bar{x} = 0.0888\,, \quad \bar{y} = 0.1078\,, \quad S^2 = 0.00875$

Precision evaluation

Estimate the variance with

$$v_m = \frac{1}{m}\frac{1}{m-1}\sum_{j=1}^{m}\left[h(x_j) - \overline{h}_m\right]^2,$$

and for m large,

$$\left\{\overline{h}_m - \mathbb{E}_f[h(X)]\right\} / \sqrt{v_m} \approx \mathcal{N}(0,1).$$

Precision evaluation

Estimate the variance with

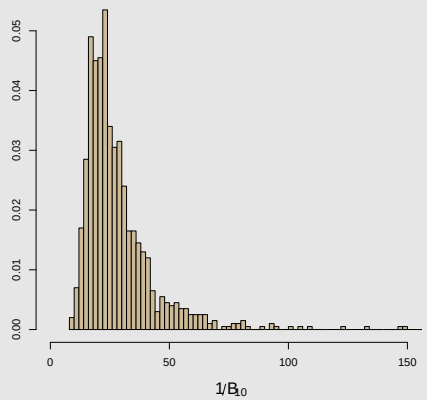
$$v_m = \frac{1}{m}\frac{1}{m-1}\sum_{j=1}^{m}[h(x_j) - \overline{h}_m]^2,$$

and for m large,

$$\left\{\overline{h}_m - \mathbb{E}_f[h(X)]\right\} / \sqrt{v_m} \approx \mathcal{N}(0,1).$$

Note

Construction of a convergence test and of confidence bounds on the approximation of $\mathbb{E}_f[h(X)]$



Example (Cauchy-normal)

For estimating a normal mean, a *robust* prior is a Cauchy prior

$$x \sim \mathcal{N}(\theta, 1), \quad \theta \sim \mathcal{C}(0, 1).$$

Under squared error loss, posterior mean

$$\delta^\pi(x) = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{\theta}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}{\displaystyle\int_{-\infty}^{\infty} \frac{1}{1+\theta^2} e^{-(x-\theta)^2/2} d\theta}$$

Example (Cauchy-normal (2))

Form of δ^π suggests simulating iid variables $\theta_1, \cdots, \theta_m \sim \mathcal{N}(x, 1)$ and calculate

$$\hat{\delta}_m^\pi(x) = \frac{\sum_{i=1}^m \dfrac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^m \dfrac{1}{1+\theta_i^2}} \,.$$

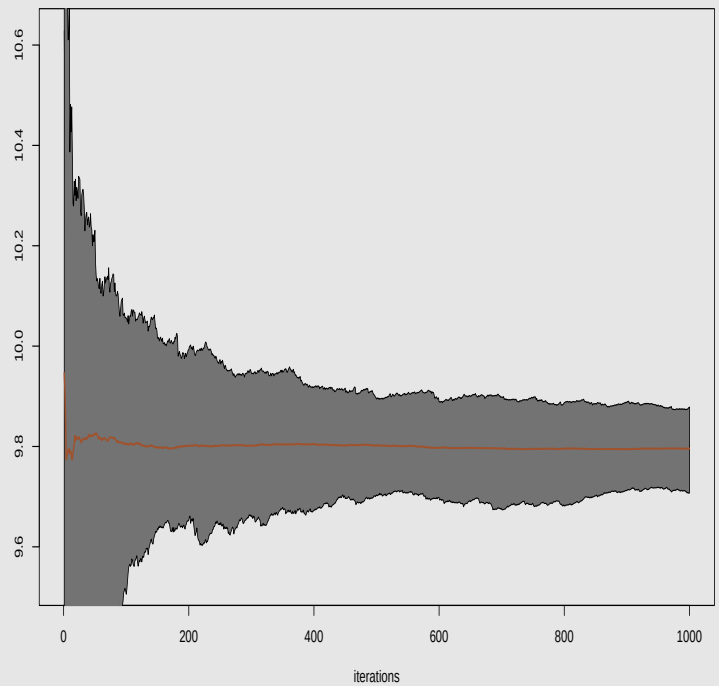
Example (Cauchy-normal (2))

Form of δ^π suggests simulating iid variables $\theta_1, \cdots, \theta_m \sim \mathcal{N}(x, 1)$ and calculate

$$\hat{\delta}_m^\pi(x) = \frac{\sum_{i=1}^m \frac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^m \frac{1}{1+\theta_i^2}} \,.$$

LLN implies

$$\hat{\delta}_m^\pi(x) \longrightarrow \delta^\pi(x) \text{ as } m \longrightarrow \infty.$$



Importance sampling

Simulation from f (the true density) is not necessarily **optimal**

Importance sampling

Simulation from f (the true density) is not necessarily **optimal**

Alternative to direct sampling from f is **importance sampling**, based on the alternative representation

$$\mathbb{E}_f[h(x)] = \int_{\mathcal{X}} \left[h(x)\,\frac{f(x)}{g(x)}\right]\, g(x)\,dx = \mathbb{E}_g\left[h(x)\frac{f(x)}{g(x)}\right]$$

which allows us to use **other** distributions than f

Importance sampling (cont'd)

Importance sampling algorithm

Evaluation of

$$\mathbb{E}_f[h(x)] = \int_{\mathcal{X}} h(x)\, f(x)\, dx$$

by

1. Generate a sample $x_1, \ldots, x_m$ from a distribution g
2. Use the approximation

$$\frac{1}{m} \sum_{j=1}^{m} \frac{f(x_j)}{g(x_j)}\, h(x_j)$$

Justification

Convergence of the estimator

$$\frac{1}{m}\sum_{j=1}^{m}\frac{f(x_j)}{g(x_j)}\,h(x_j)\longrightarrow\mathbb{E}_f[h(x)]$$

1. converges for any choice of the distribution g as long as $\text{supp}(g)\supset\text{supp}(f)$

Justification

Convergence of the estimator

$$\frac{1}{m}\sum_{j=1}^{m}\frac{f(x_j)}{g(x_j)}\,h(x_j)\longrightarrow\mathbb{E}_f[h(x)]$$

1. converges for any choice of the distribution g as long as $\text{supp}(g)\supset\text{supp}(f)$
2. Instrumental distribution g chosen from distributions easy to simulate

Justification

Convergence of the estimator

$$\frac{1}{m}\sum_{j=1}^{m}\frac{f(x_j)}{g(x_j)}\,h(x_j)\longrightarrow\mathbb{E}_f[h(x)]$$

1. converges for any choice of the distribution g as long as $\text{supp}(g)\supset\text{supp}(f)$
2. Instrumental distribution g chosen from distributions easy to simulate
3. Same sample (generated from g) can be used repeatedly, not only for different functions h, but also for different densities f

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

$$\mathbb{E}_f\left[h^2(x)\frac{f(x)}{g(x)}\right] = \int_{\mathcal{X}} h^2(x)\,\frac{f^2(x)}{g(x)}\,dx < \infty\,.$$

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

$$\mathbb{E}_f\left[h^2(x)\frac{f(x)}{g(x)}\right] = \int_{\mathcal{X}} h^2(x)\,\frac{f^2(x)}{g(x)}\,dx < \infty\,.$$

2. Instrumental distributions with tails lighter than those of f (that is, with $\sup f/g = \infty$) not appropriate, because weights $f(x_j)/g(x_j)$ vary widely, giving too much importance to a few values x_j.

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

$$\mathbb{E}_f\left[h^2(x)\frac{f(x)}{g(x)}\right] = \int_{\mathcal{X}} h^2(x)\,\frac{f^2(x)}{g(x)}\,dx < \infty\,.$$

2. Instrumental distributions with tails lighter than those of f (that is, with $\sup f/g = \infty$) not appropriate, because weights $f(x_j)/g(x_j)$ vary widely, giving too much importance to a few values x_j.
3. If $\sup f/g = M < \infty$, the accept-reject algorithm can be used as well to simulate f directly.

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

$$\mathbb{E}_f\left[h^2(x)\frac{f(x)}{g(x)}\right] = \int_{\mathcal{X}} h^2(x)\,\frac{f^2(x)}{g(x)}\,dx < \infty\,.$$

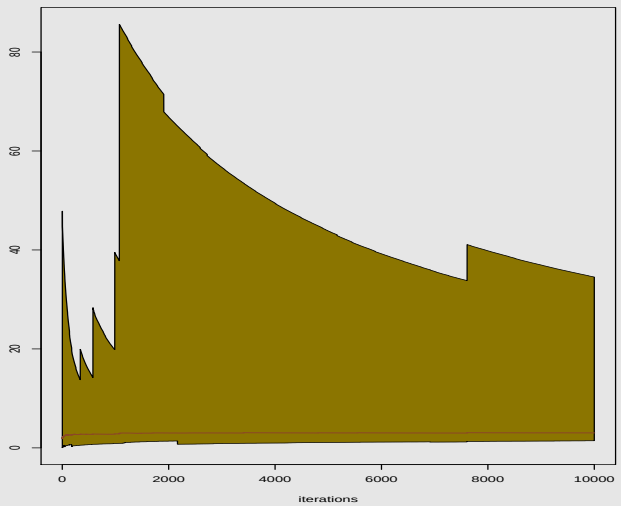
2. Instrumental distributions with tails lighter than those of f (that is, with $\sup f/g = \infty$) not appropriate, because weights $f(x_j)/g(x_j)$ vary widely, giving too much importance to a few values x_j.
3. If $\sup f/g = M < \infty$, the accept-reject algorithm can be used as well to simulate f directly.
4. IS suffers from curse of dimensionality

Example (Cauchy target)

Case of Cauchy distribution $\mathcal{C}(0,1)$ when importance function is Gaussian $\mathcal{N}(0,1)$.

Density ratio

$$\frac{p^\star(x)}{p_0(x)} = \sqrt{2\pi}\,\frac{\exp x^2/2}{\pi\,(1+x^2)}$$



very badly behaved: e.g.,

$$\int_{-\infty}^{\infty} \varrho(x)^2 p_0(x) dx = \infty$$

Poor performances of the associated importance sampling estimator

Practical alternative

$$\sum_{j=1}^{m} h(x_j)\, f(x_j)/g(x_j) \Big/ \sum_{j=1}^{m} f(x_j)/g(x_j)$$

where f and g are known up to constants.

1. Also converges to $\mathfrak{I}$ by the Strong Law of Large Numbers.
2. Biased, but the bias is quite small: may beat the unbiased estimator in squared error loss.

Example (Student's t distribution)

$x \sim \mathcal{T}(\nu, \theta, \sigma^2)$, with density

$$f_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} .$$

Without loss of generality, take $\theta = 0$, $\sigma = 1$.

Example (Student's t distribution)

$x \sim \mathcal{T}(\nu, \theta, \sigma^2)$, with density

$$f_\nu(x) = \frac{\Gamma((\nu+1)/2)}{\sigma\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{(x-\theta)^2}{\nu\sigma^2}\right)^{-(\nu+1)/2} .$$

Without loss of generality, take $\theta = 0$, $\sigma = 1$.
Integral of interest

$$\mathfrak{I} = \int \sqrt{\left|\frac{x}{1-x}\right|}\, f_\nu(x)\, \mathrm{d}x$$

Example (Student's t distribution (2))

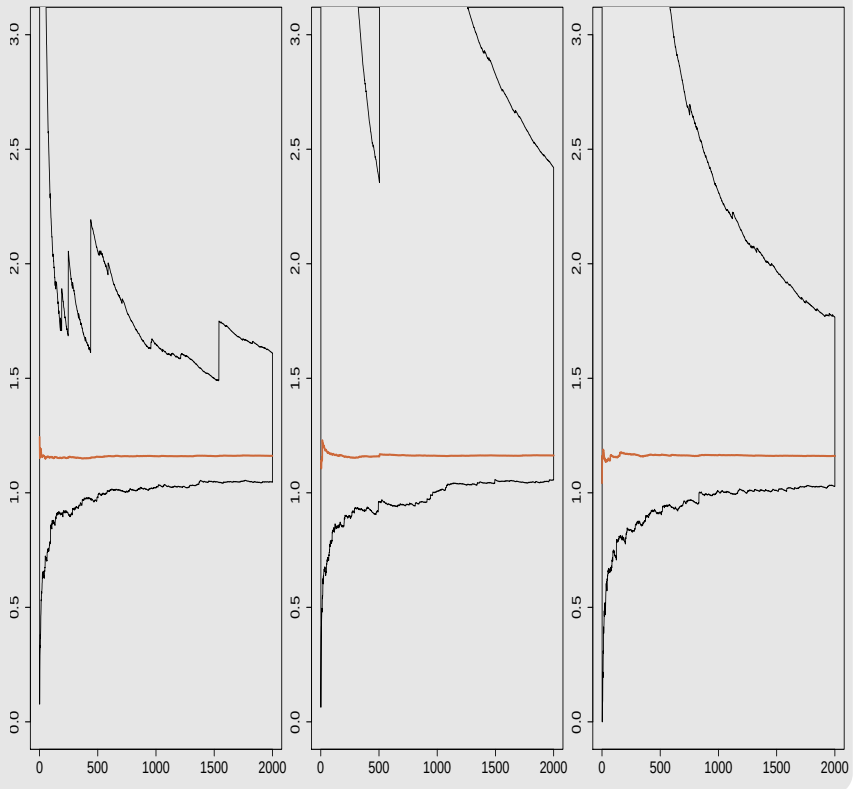
Choices of h:

1. Student $\mathcal{T}(\nu, 0, 1)$
2. Cauchy $\mathcal{C}(0, 1)$
3. Normal $\mathcal{N}(0, \nu/(\nu-2))$

Note: The ratio

$$\frac{f^2(x)}{h(x)} \propto \frac{e^{x^2(\nu-2)/2\nu}}{[1+x^2/\nu]^{(\nu+1)}}$$

does not have a finite integral



Explanation

Example (Student's t distribution (3))

Phenomenon due to the fact that h has a singularity at $x = 1$:

$$\int \frac{|x|}{|1-x|} f_\nu(x)\, \mathrm{d}x = \infty$$

Explanation

Example (Student's t distribution (3))

Phenomenon due to the fact that h has a singularity at $x = 1$:

$$\int \frac{|x|}{|1-x|} f_\nu(x)\,\mathrm{d}x = \infty$$

Consequence: the three estimators have infinite variance

Alternative

Example (Student's t distribution (4))

Choose a better behaved h:

Alternative

Example (Student's t distribution (4))

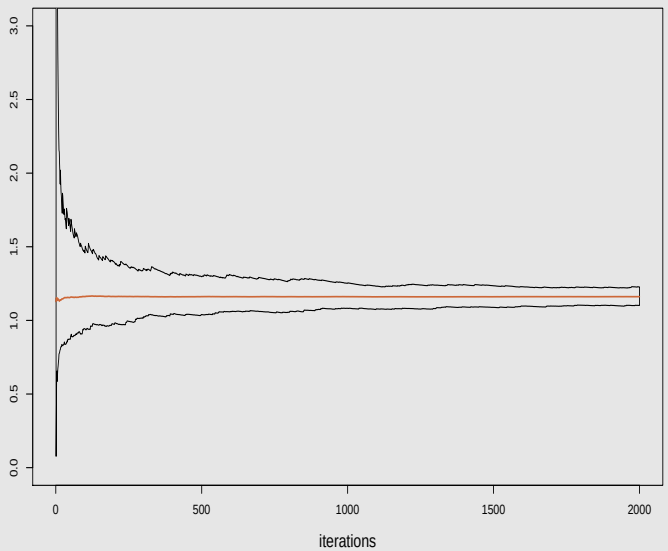
Choose a better behaved h: folded Gamma distribution, x symmetric around 1 with

$$|x-1| \sim \mathcal{G}a(\alpha, 1)$$

Then $h_1(x)f^2(x)/h(x)$ proportional to

$$\sqrt{x}\, f^2(x)\, |1-x|^{1-\alpha-1}\, \exp|1-x|$$

integrable around $x = 1$ when $\alpha < 1$.



Choice of importance function (termin'd)

The importance function may be π

Choice of importance function (termin'd)

The importance function may be π

- often inefficient if data informative
- impossible if π is improper

Choice of importance function (termin'd)

The importance function may be π

- often inefficient if data informative
- impossible if π is improper

Defensive sampling:

$$h(\theta) = \rho\pi(\theta) + (1-\rho)\pi(\theta|x) \qquad \rho \ll 1$$

Example (Cauchy/Normal)

Consider

$$x_1, \ldots, x_n \sim \mathcal{C}(\theta, 1) \qquad \text{and} \qquad \theta \sim \mathcal{N}(\mu, \sigma^2),$$

with known hyperparameters μ and σ^2.

Example (Cauchy/Normal)

Consider

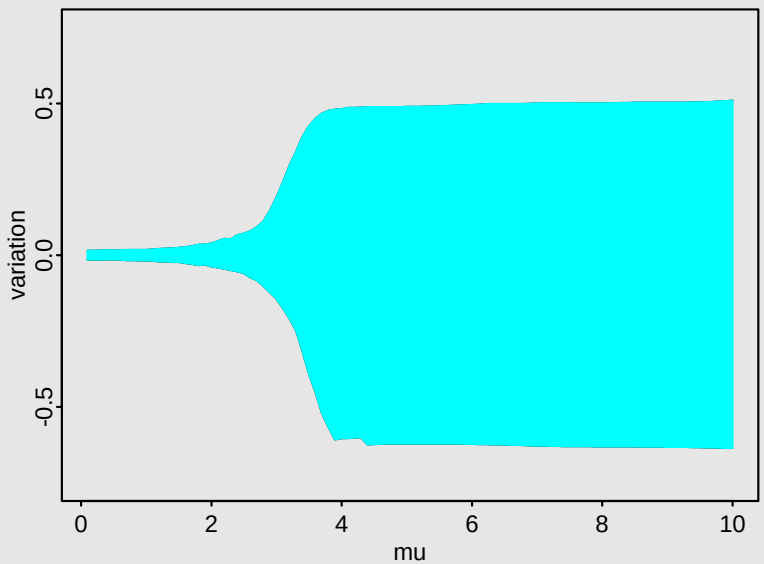
$$x_1, \ldots, x_n \sim \mathcal{C}(\theta, 1) \qquad \text{and} \qquad \theta \sim \mathcal{N}(\mu, \sigma^2),$$

with known hyperparameters μ and σ^2.
Since $\pi(\theta)$ is normal $\mathcal{N}(\mu, \sigma^2)$, possible to simulate a normal sample $\theta_1, \ldots, \theta_M$ and to approximate the Bayes estimator by

$$\hat{\delta}^{\pi}(x_1, \ldots, x_n) = \frac{\sum_{t=1}^{M} \theta_t \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}{\sum_{t=1}^{M} \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}.$$

Example (Cauchy/Normal (2))

Poor when the x_i's are all far from μ



90% range of variation for $n = 10$ observations from $\mathcal{C}(0, 1)$ distribution and $M = 1000$ simulations of θ as μ varies

Bridge sampling

Bayes factor

$$B_{12}^{\pi} = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2}$$

Bridge sampling

Bayes factor

$$B_{12}^{\pi} = \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2}$$

If

$$\begin{aligned} \pi_1(\theta_1|x) &\propto \tilde{\pi}_1(\theta_1|x) \\ \pi_2(\theta_2|x) &\propto \tilde{\pi}_2(\theta_2|x) \end{aligned}$$

then

$$B_{12}^{\pi} \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{\pi}_1(\theta_i|x)}{\tilde{\pi}_2(\theta_i|x)} \qquad \theta_i \sim \pi_2(\theta|x)$$

Prediction

If $x \sim f(x|\theta)$ and $z \sim g(z|x,\theta)$, the *predictive* of z is

$$g^{\pi}(z|x) = \int_{\Theta} g(z|x,\theta)\pi(\theta|x)\, d\theta.$$

Normal prediction

For $\mathscr{D}_n = (x_1, \ldots, x_n) \sim \mathscr{N}(\mu, \sigma^2)$ and

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-\lambda_\sigma - 3/2} \exp - \left\{ -\lambda_\mu (\mu - \xi)^2 + \alpha \right\} / 2\sigma^2 \,,$$

corresponding posterior

$$\mathscr{N}\left(\frac{\lambda_\mu \xi + n\overline{x}_n}{\lambda_\mu + n}, \frac{\sigma^2}{\lambda_\mu + n} \right) \times \mathscr{IG}\left(\lambda_\sigma + n/2, \left[\alpha + s_x^2 + \frac{n\lambda_\mu}{\lambda_\mu + n} (\overline{x} - \xi)^2 \right] / 2 \right) \,,$$

Normal prediction

For $\mathscr{D}_n = (x_1, \ldots, x_n) \sim \mathscr{N}(\mu, \sigma^2)$ and

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-\lambda_\sigma - 3/2} \exp - \left\{-\lambda_\mu(\mu - \xi)^2 + \alpha\right\} / 2\sigma^2 \,,$$

corresponding posterior

$$\mathscr{N}\left(\frac{\lambda_\mu \xi + n\overline{x}_n}{\lambda_\mu + n}, \frac{\sigma^2}{\lambda_\mu + n}\right) \times \mathscr{IG}\left(\lambda_\sigma + n/2, \left[\alpha + s_x^2 + \frac{n\lambda_\mu}{\lambda_\mu + n}(\overline{x} - \xi)^2\right] / 2\right) \,,$$

Notation

$$\mathscr{N}\left(\xi(\mathscr{D}_n), \sigma^2/\lambda_\mu(\mathscr{D}_n)\right) \times \mathscr{IG}\left(\lambda_\sigma(\mathscr{D}_n), \alpha(\mathscr{D}_n)/2\right)$$

Normal prediction (cont'd)

Predictive on x_{n+1}

$$
\begin{aligned}
f^\pi(x_{n+1}|\mathscr{D}_n) &\propto \int (\sigma^2)^{-\lambda_\sigma-2-n/2} \exp -(x_{n+1}-\mu)^2/2\sigma^2 \\
&\qquad \times \exp -\left\{\lambda_\mu(\mathscr{D}_n)(\mu-\xi(\mathscr{D}_n))^2 + \alpha(\mathscr{D}_n)\right\}/2\sigma^2 \,\mathrm{d}(\mu,\sigma^2) \\
&\propto \int (\sigma^2)^{-\lambda_\sigma-n/2-3/2} \exp -\left\{(\lambda_\mu(\mathscr{D}_n)+1)(x_{n+1}-\xi(\mathscr{D}_n))^2\right. \\
&\qquad \left./\lambda_\mu(\mathscr{D}_n) + \alpha(\mathscr{D}_n)\right\}/2\sigma^2 \,\mathrm{d}\sigma^2 \\
&\propto \left[\alpha(\mathscr{D}_n) + \frac{\lambda_\mu(\mathscr{D}_n)+1}{\lambda_\mu(\mathscr{D}_n)}(x_{n+1}-\xi(\mathscr{D}_n))^2\right]^{-(2\lambda_\sigma+n+1)/2}
\end{aligned}
$$

Student's t distribution with mean $\xi(\mathscr{D}_n)$ and $2\lambda_\sigma + n$ degrees of freedom.

normaldata

Noninformative case $\lambda_\mu = \lambda_\sigma = \alpha = 0$

$$f^\pi(x_{n+1}|\mathscr{D}_n) \propto \left[s_x^2 + \frac{n}{n+1}(x_{n+1} - \overline{x}_n)^2\right]^{-(n+1)/2} .$$

Predictive distribution on a 91st county is Student's t

$$\mathscr{T}(90, -0.0413, 0.136)$$