Chapter 4 :
Decision theory and Bayesian analysis

5. Decision theory and Bayesian analysis
   - Bayesian modelling
   - Conjugate priors
   - Improper prior distributions
   - Bayesian inference

A pedestrian example

paired and orphan socks

A drawer contains an unknown number of socks, some of which can be paired and some of which are orphans (single). One takes at random 11 socks without replacement from this drawer: no pair can be found among those. What can we infer about the total number of socks in the drawer?

- sounds like an impossible task
- one observation $x = 11$ and two unknowns, n_{socks} and n_{pairs}
- writing the likelihood is a challenge [exercise]

A pedestrian example

paired and orphan socks

A drawer contains an unknown number of socks, some of which can be paired and some of which are orphans (single). One takes at random 11 socks without replacement from this drawer: no pair can be found among those. What can we infer about the total number of socks in the drawer?

- sounds like an impossible task
- one observation $x = 11$ and two unknowns, n_{socks} and n_{pairs}
- writing the likelihood is a challenge [exercise]

A prioris on socks

Given parameters n_{socks} and n_{pairs}, set of socks

$$\mathcal{S} = \left\{ s_1, s_1, \ldots, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}+1}, \ldots, s_{n_{\text{socks}}} \right\}$$

and 11 socks picked at random from $\mathcal{S}$ give X unique socks.

Rassmus' reasoning

If you are a family of 3-4 persons then a guesstimate would be that you have something like 15 pairs of socks in store. It is also possible that you have much more than 30 socks. So as *a prior* for n_{socks} I'm going to use a negative binomial with mean 30 and standard deviation 15.
On $n_{\text{pairs}}/2n_{\text{socks}}$ I'm going to put a Beta *prior* distribution that puts most of the probability over the range 0.75 to 1.0,

[Rassmus Bååth's Research Blog, Oct 20th, 2014]

A prioris on socks

Given parameters n_{socks} and n_{pairs}, set of socks

$$\mathcal{S} = \left\{ s_1, s_1, \ldots, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}}, s_{n_{\text{pairs}}+1}, \ldots, s_{n_{\text{socks}}} \right\}$$

and 11 socks picked at random from $\mathcal{S}$ give X unique socks.

Rassmus' reasoning

If you are a family of 3-4 persons then a guesstimate would be that you have something like 15 pairs of socks in store. It is also possible that you have much more than 30 socks. So as *a prior* for n_{socks} I'm going to use a negative binomial with mean 30 and standard deviation 15.
On $n_{\text{pairs}}/2n_{\text{socks}}$ I'm going to put a Beta *prior* distribution that puts most of the probability over the range 0.75 to 1.0,

[Rassmus Bååth's Research Blog, Oct 20th, 2014]

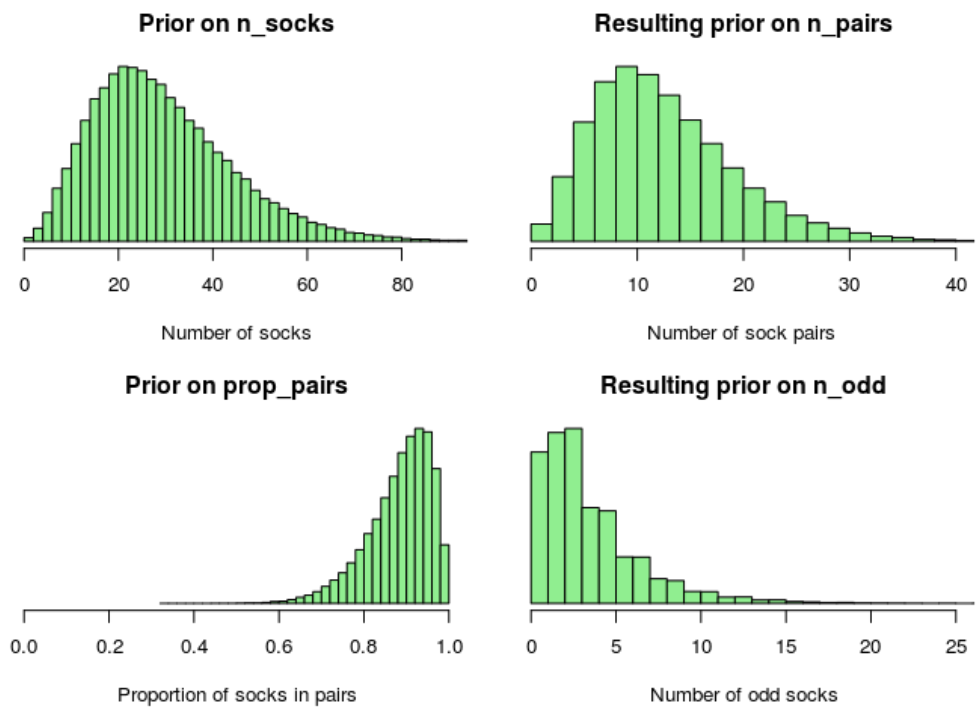
Simulating the experiment

Given a *prior* distribution on n_{socks} and n_{pairs},

$$n_{\text{socks}} \sim \mathcal{N}eg(30, 15) \quad n_{\text{pairs}} | n_{\text{socks}} \sim {}^{n_{\text{socks}}}/_{2}\mathcal{B}e(15, 2)$$

possible to

1. generate new values of n_{socks} and n_{pairs},
2. generate a new observation of X, number of unique socks out of 11.
3. accept the pair $(n_{\text{socks}}, n_{\text{pairs}})$ if the realisation of X is equal to 11



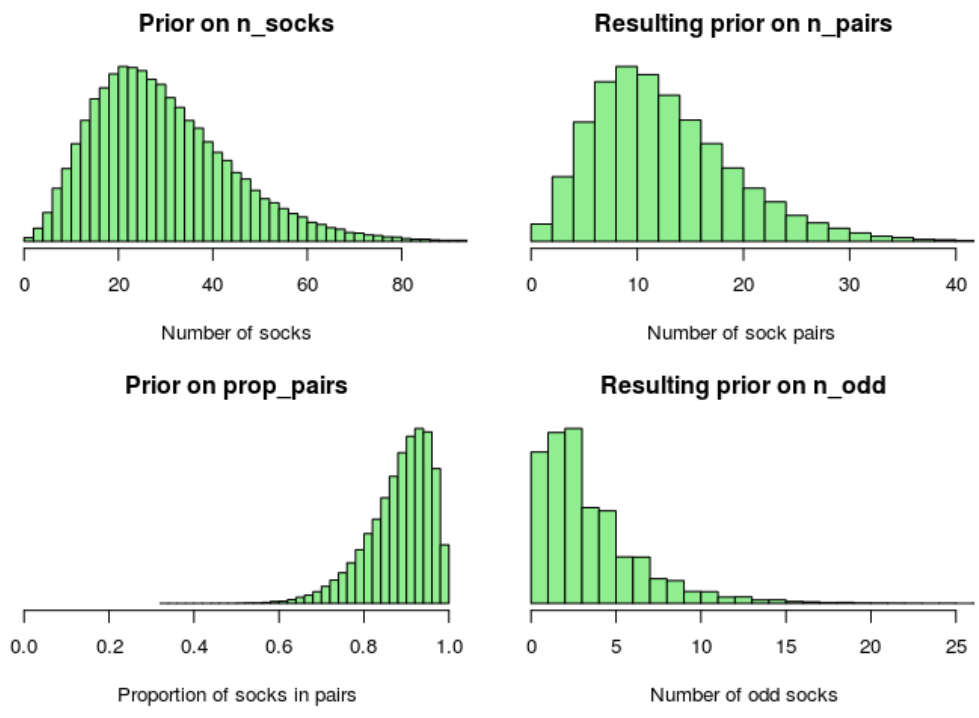
Simulating the experiment

Given a *prior* distribution on n_{socks} and n_{pairs},

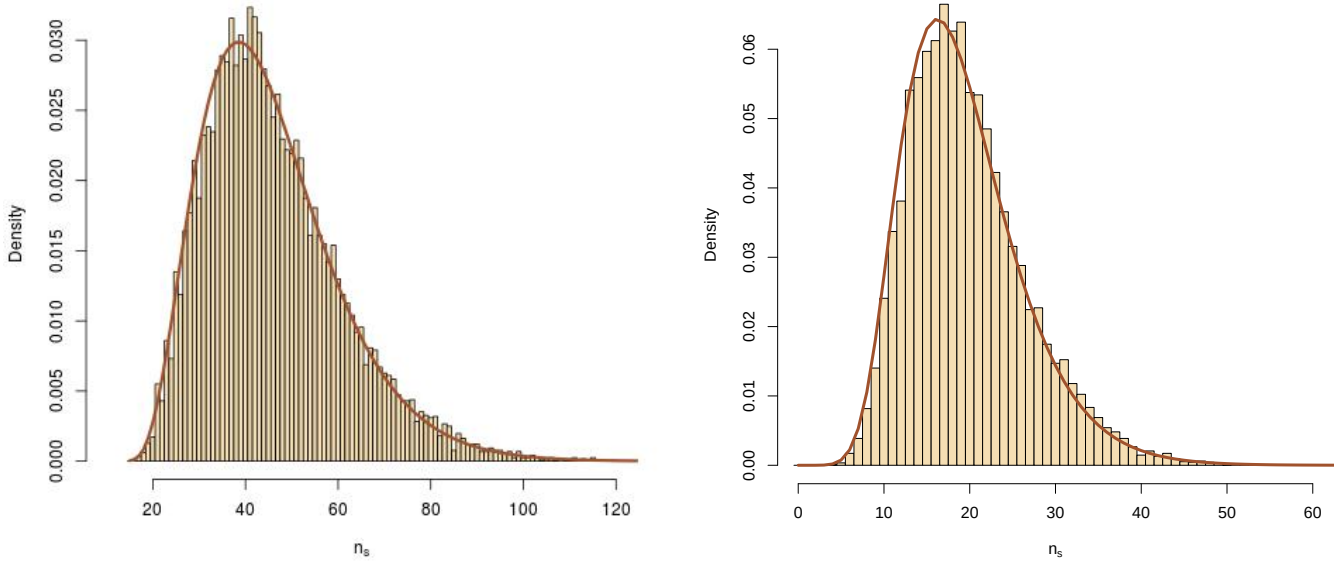
$$n_{\text{socks}} \sim \mathcal{N}eg(30, 15) \quad n_{\text{pairs}} | n_{\text{socks}} \sim {}^{n_{\text{socks}}}/_{2}\mathcal{B}e(15, 2)$$

possible to

1. generate new values of n_{socks} and n_{pairs},
2. generate a new observation of X, number of unique socks out of 11.
3. accept the pair $(n_{\text{socks}}, n_{\text{pairs}})$ if the realisation of X is equal to 11



Meaning

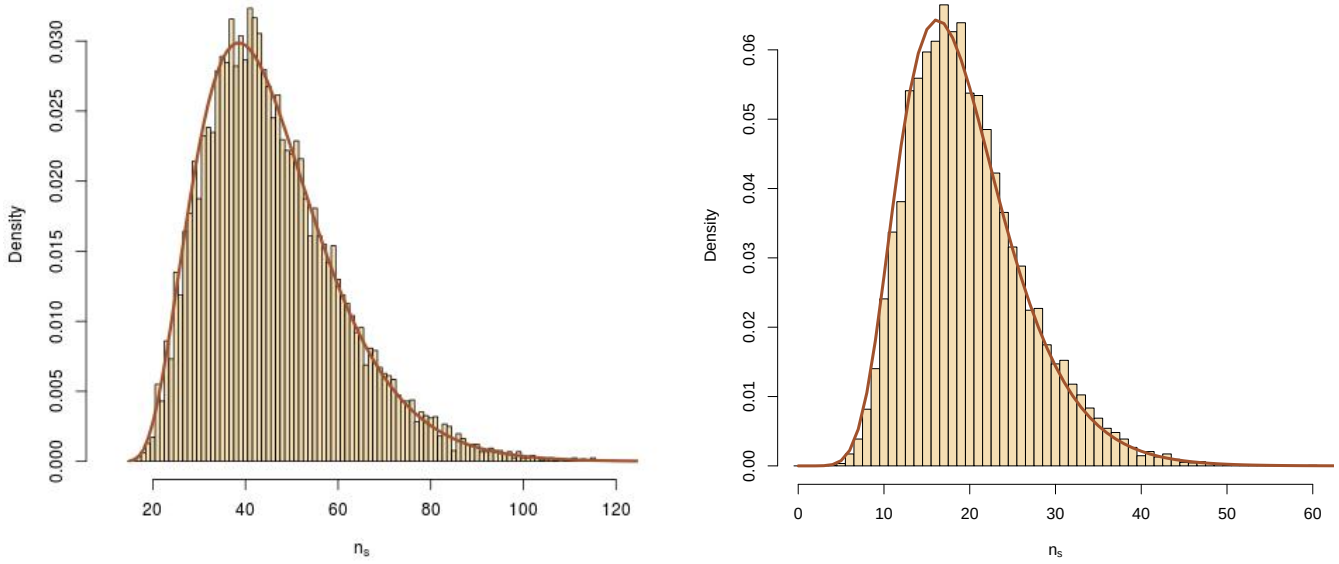


The outcome of this simulation method returns a distribution on the pair $(n_{\text{socks}}, n_{\text{pairs}})$ that is the conditional distribution of the pair given the observation $X = 11$

Proof: Generations from $\pi(n_{\text{socks}}, n_{\text{pairs}})$ are accepted with probability

$$\mathbb{P}\{X = 11 | (n_{\text{socks}}, n_{\text{pairs}})\}$$

Meaning



The outcome of this simulation method returns a distribution on the pair $(n_{\text{socks}}, n_{\text{pairs}})$ that is the conditional distribution of the pair given the observation $X = 11$

Proof: Hence accepted values distributed from

$$\pi(n_{\text{socks}}, n_{\text{pairs}}) \times \mathbb{P}\{X = 11 | (n_{\text{socks}}, n_{\text{pairs}})\} = \pi(n_{\text{socks}}, n_{\text{pairs}} | X = 11)$$

General principle

Bayesian principle Given a probability distribution on the parameter θ called prior

$$\pi(\theta)$$

and an observation x of $X \sim f(x|\theta)$, Bayesian inference relies on the conditional distribution of θ given $X = x$

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

called posterior distribution

[Bayes' theorem]

Thomas Bayes
(FRS, 1701?-1761)

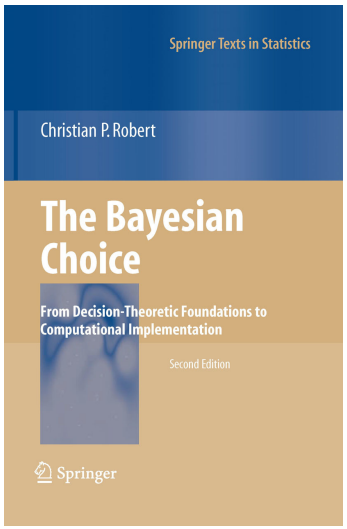
Bayesian inference

Posterior distribution

$$\pi(\theta|x)$$

as distribution on θ the parameter conditional on x the observation used for all aspects of inference

- point estimation, e.g., $\mathbb{E}[h(\theta)|x]$;
- confidence intervals, e.g., $\{\theta;\ \pi(\theta|x) \geqslant \kappa\}$;
- tests of hypotheses, e.g., $\pi(\theta = 0|x)$; and
- prediction of future observations



Central tool... central to Bayesian inference

Posterior defined up to a constant as

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

- Operates conditional upon the observation(s) $X = x$
- Integrate simultaneously prior information 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 [domino effect]
- Provides a complete inferential scope and a unique motor of inference

The thorny issue of the prior distribution

Compared with likelihood inference, based solely on

$$L(\theta|x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Bayesian inference introduces an extra measure $\pi(\theta)$ that is chosen *a priori*, hence subjectively by the statistician based on

- hypothetical range of θ
- guesstimates of θ with an associated (lack of) precision
- type of sampling distribution

Note There also exist reference solutions (see below)

The thorny issue of the prior distribution

Compared with likelihood inference, based solely on

$$L(\theta|x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

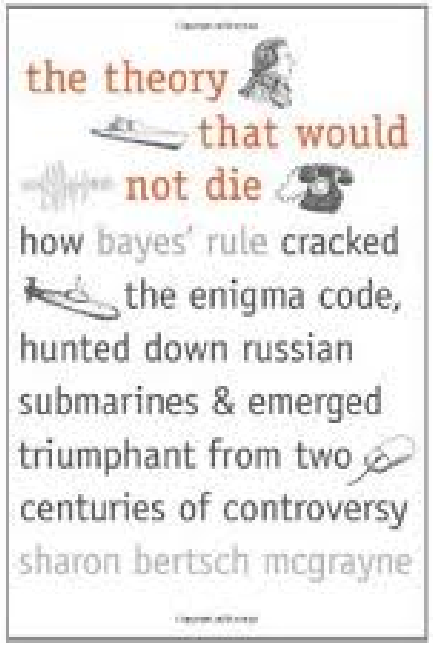
Bayesian inference introduces an extra measure $\pi(\theta)$ that is chosen *a priori*, hence subjectively by the statistician based on

- hypothetical range of θ
- guesstimates of θ with an associated (lack of) precision
- type of sampling distribution

Note There also exist reference solutions (see below)

Bayes' example

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p. Second ball O then rolled $\mathfrak{n}$ times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

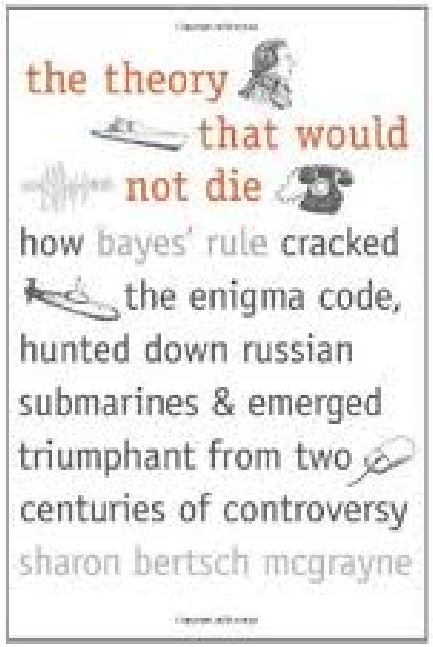


Bayes' example

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p. Second ball O then rolled n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

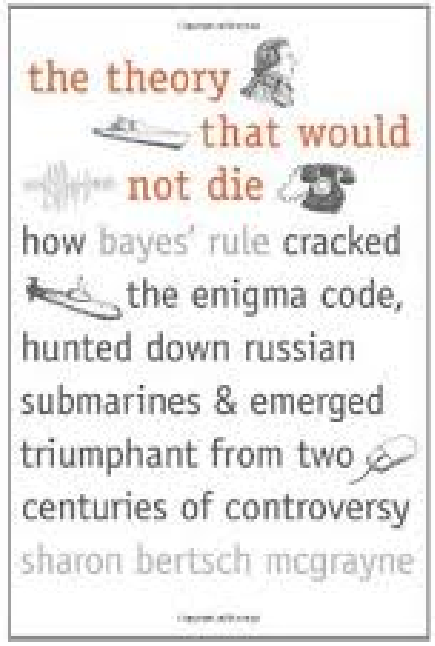
Thomas Bayes' question

Given X, what inference can we make on p?



Bayes' example

Billiard ball W rolled on a line of length one, with a uniform probability of stopping anywhere: W stops at p. Second ball O then rolled n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.



Modern translation:

Derive the posterior distribution of p given X, when

$$p \sim \mathcal{U}([0,1]) \text{ and } X \sim \mathcal{B}(n,p)$$

Resolution

Since

$$\begin{aligned} P(X = x|p) &= \binom{n}{x} p^x (1-p)^{n-x}, \\ P(a < p < b \text{ and } X = x) &= \int_a^b \binom{n}{x} p^x (1-p)^{n-x} dp \end{aligned}$$

and

$$P(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \, dp,$$

Resolution (2)

then

$$\begin{aligned} P(a < p < b | X = x) &= \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} \\ &= \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)}\ , \end{aligned}$$

i.e.

$$p|x \sim \mathcal{B}e(x+1, n-x+1)$$

[Beta distribution]

Resolution (2)

then

$$\begin{aligned} P(a < p < b | X = x) &= \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x}\, dp}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x}\, dp} \\ &= \frac{\int_a^b p^x (1-p)^{n-x}\, dp}{B(x+1, n-x+1)}\ , \end{aligned}$$

i.e.

$$p|x \sim \mathcal{B}e(x+1, n-x+1)$$

[Beta distribution]

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

Given a likelihood function $L(y|\theta)$, the family Π of priors π_0 on Θ is said to be conjugate if the posterior $\pi(\cdot|y)$ also belong to Π

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

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 $\mathcal{F}$.

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Conjugate priors

Easiest case is when prior distribution is within parametric family

Conjugacy

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 $\mathcal{F}$.

In this case, posterior inference is tractable and reduces to updating the hyperparameters* of the prior

Example In Thomas Bayes' example, the $\mathcal{B}e(a, b)$ prior is conjugate

*The hyperparameters are parameters of the priors; they are most often not treated as random variables

Exponential families and conjugacy

The family of exponential distributions

$$\begin{aligned} f(x|\theta) &= C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\} \\ &= h(x)\exp\{R(\theta)\cdot T(x) - \tau(\theta)\} \end{aligned}$$

allows for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta.\mu - \lambda\psi(\theta)}$$

Following Pitman-Koopman-Darmois' Lemma, only case [besides uniform distributions]

Exponential families and conjugacy

The family of exponential distributions

$$\begin{aligned} f(x|\theta) &= C(\theta)h(x)\exp\{R(\theta)\cdot T(x)\} \\ &= h(x)\exp\{R(\theta)\cdot T(x) - \tau(\theta)\} \end{aligned}$$

allows for conjugate priors

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)\, e^{\theta.\mu - \lambda\psi(\theta)}$$

Following Pitman-Koopman-Darmois' Lemma, only case [besides uniform distributions]

Illustration

Discrete/Multinomial & Dirichlet

If observations consist of positive counts $Y_1, \ldots, Y_d$ modelled by a Multinomial $\mathcal{M}(\theta_1, \ldots, \theta_p)$ distribution

$$L(y|\theta, n) = \frac{n!}{\prod_{i=1}^{d} y_i!} \prod_{i=1}^{d} \theta_i^{y_i}$$

conjugate family is the Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_d)$ distribution

$$\pi(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i}^{d} \theta_i^{\alpha_i - 1}$$

defined on the probability simplex $(\theta_i \geqslant 0, \sum_{i=1}^{d} \theta_i = 1)$, where Γ is the gamma function $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$

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

Standard exponential families [2]

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

Lemma If

$$\theta \sim \pi_{\lambda, x_0}(\theta) \propto e^{\theta \cdot x_0 - \lambda \psi(\theta)}$$

with $x_0 \in \mathcal{X}$, then

$$\mathbb{E}^{\pi}[\nabla \psi(\theta)] = \frac{x_0}{\lambda}.$$

Therefore, if $x_1, \ldots, x_n$ are i.i.d. $f(x|\theta)$,

$$\mathbb{E}^{\pi}[\nabla \psi(\theta) | x_1, \ldots, x_n] = \frac{x_0 + n\bar{x}}{\lambda + n}$$

Improper distributions

Necessary extension from a prior probability distribution to a prior σ-finite positive measure π such that

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

Improper prior distribution

Note A σ-finite density with

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta < +\infty$$

can be renormalised into a probability density

Improper distributions

Necessary extension from a prior probability distribution to a prior σ-finite positive measure π such that

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

Improper prior distribution

Note A σ-finite density with

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta < +\infty$$

can be renormalised into a probability density

Improper distributions

Necessary extension from a prior probability distribution to a prior σ-finite positive measure π such that

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

Improper prior distribution

Note A σ-finite density with

$$\int_\Theta \pi(\theta)\,\mathrm{d}\theta < +\infty$$

can be renormalised into a probability density

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Justifications

Often automatic prior determination leads to improper prior distributions

1. Only way to derive a prior in noninformative settings
2. Performances of estimators derived from these generalized distributions usually good
3. Improper priors often occur as limits of proper distributions
4. More *robust* answer against possible *misspecifications* of the prior
5. Penalization factor

Validation

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

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

when

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

Validation

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

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

when

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

Normal illustration

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

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

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

[independent of ω]

Normal illustration

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

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

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

[independent of ω]

Normal illustration

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

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

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

[independent of ω]

Warning

The mistake is to think of them [non-informative priors] as representing ignorance

[Lindley, 1990]

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

$$\lim_{\tau \to \infty} \mathrm{P}^{\pi}(\theta \in [a, b]) = 0$$

for any (a, b)

Warning

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]

Normal illustration:

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

$$\lim_{\tau \to \infty} \mathrm{P}^{\pi}\left(\theta \in [a, b]\right) = 0$$

for any (a, b)

Haldane prior

Consider a binomial observation, $x \sim \mathcal{B}(n, p)$, and

$$\pi^*(p) \propto [p(1-p)]^{-1}$$

[Haldane, 1931]

The marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$.

[Not recommended!]

Haldane prior

Consider a binomial observation, $x \sim \mathcal{B}(n, p)$, and

$$\pi^*(p) \propto [p(1-p)]^{-1}$$

[Haldane, 1931]

The marginal distribution,

$$\begin{aligned} m(x) &= \int_0^1 [p(1-p)]^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= B(x, n-x), \end{aligned}$$

is only defined for $x \neq 0, n$.

[Not recommended!]

The Jeffreys prior

Based on Fisher information

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

Jeffreys prior density is

$$\pi^*(\theta) \propto |\mathfrak{I}(\theta)|^{1/2}$$

Pros & Cons

- relates to information theory
- agrees with most invariant priors
- parameterisation invariant

The Jeffreys prior

Based on Fisher information

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

Jeffreys prior density is

$$\pi^*(\theta) \propto |\mathfrak{I}(\theta)|^{1/2}$$

Pros & Cons

- relates to information theory
- agrees with most invariant priors
- parameterisation invariant

Example

If $x \sim \mathcal{N}_p(\theta, I_p)$, Jeffreys' prior is

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

and if $\eta = \|\theta\|^2$,

$$\pi(\eta) = \eta^{p/2-1}$$

and

$$\mathbb{E}^{\pi}[\eta|x] = \|x\|^2 + p$$

with bias $2p$

[Not recommended!]

Example

If $x \sim \mathcal{B}(n, \theta)$, Jeffreys' prior is

$$\mathcal{B}e(1/2, 1/2)$$

and, if $n \sim \mathcal{N}eg(x, \theta)$, Jeffreys' prior is

$$\begin{aligned} \pi_2(\theta) &= -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \right] \\ &= \mathbb{E}_\theta \left[\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} \right] = \frac{x}{\theta^2(1-\theta)}, \\ &\propto \theta^{-1}(1-\theta)^{-1/2} \end{aligned}$$

MAP estimator

When considering estimates of the parameter θ, one default solution is the maximum a posteriori (MAP) estimator

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

Motivations

- Most likely value of θ
- Penalized likelihood estimator
- Further appeal in restricted parameter spaces

MAP estimator

When considering estimates of the parameter θ, one default solution is the maximum a posteriori (MAP) estimator

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

Motivations

- Most likely value of θ
- Penalized likelihood estimator
- Further appeal in restricted parameter spaces

Illustration

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

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

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

Corresponding MAP estimators:

$$\begin{aligned}
\delta^*(x) &= \max\left(\frac{x-1/2}{n-1}, 0\right), \\
\delta_1(x) &= \frac{x}{n}, \\
\delta_2(x) &= \max\left(\frac{x-1}{n-2}, 0\right).
\end{aligned}$$

Illustration [opposite]

MAP not always appropriate:

When

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

and

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

then MAP estimator of θ is always

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

Prediction

Inference on new observations depending on the same parameter, conditional on the current data

If $x \sim f(x|\theta)$ [observed], $\theta \sim \pi(\theta)$, and $z \sim g(z|x,\theta)$ [unobserved], *predictive* of z is marginal conditional

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

time series illustration

Consider the AR(1) model

$$x_t = \rho x_{t-1} + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

predictive of x_T is then

$$x_T | x_{1:(T-1)} \sim \int \frac{\sigma^{-1}}{\sqrt{2\pi}} \exp\{-(x_T - \rho x_{T-1})^2 / 2\sigma^2\} \pi(\rho, \sigma | x_{1:(T-1)}) d\rho d\sigma ,$$

and $\pi(\rho, \sigma | x_{1:(T-1)})$ can be expressed in closed form

Posterior mean

Theorem The solution to

$$\arg\min_{\delta} \mathbb{E}^{\pi}\left[\|\theta-\delta\|^2 \middle| x\right]$$

is given by

$$\delta^{\pi}(x) = \mathbb{E}^{\pi}\left[\theta|x\right]$$

[Posterior mean = Bayes estimator under quadratic loss]

Posterior median

Theorem When $\theta \in \mathbb{R}$, the solution to

$$\arg\min_{\delta} \mathbb{E}^{\pi} \left[|\theta - \delta| \, | \, x \right]$$

is given by

$$\delta^{\pi}(x) = \text{median}^{\pi} \left(\theta | x \right)$$

[Posterior mean = Bayes estimator under absolute loss]

Obvious extension to

$$\arg\min_{\delta} \mathbb{E}^{\pi} \left[\sum_{i=1}^{p} |\theta_i - \delta| \, \middle| \, x \right]$$

Posterior median

Theorem When $\theta \in \mathbb{R}$, the solution to

$$\arg\min_{\delta} \mathbb{E}^{\pi} \left[|\theta - \delta| \,|\, x \right]$$

is given by

$$\delta^{\pi}(x) = \text{median}^{\pi} \left(\theta | x \right)$$

[Posterior mean = Bayes estimator under absolute loss]

Obvious extension to

$$\arg\min_{\delta} \mathbb{E}^{\pi} \left[\sum_{i=1}^{p} |\theta_i - \delta| \,\middle|\, x \right]$$

Inference with conjugate priors

For conjugate distributions, posterior expectations of the natural parameters may be expressed analytically, for one or several observations.

Distribution	Conjugate prior	Posterior mean
Normal	Normal	
$\mathcal{N}(\theta, \sigma^2)$	$\mathcal{N}(\mu, \tau^2)$	$\frac{\mu\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}$
Poisson	Gamma	
$\mathcal{P}(\theta)$	$\mathcal{G}(\alpha, \beta)$	$\frac{\alpha + x}{\beta + 1}$

Inference with conjugate priors

For conjugate distributions, posterior expectations of the natural parameters may be expressed analytically, for one or several observations.

Distribution	Conjugate prior	Posterior mean
Gamma	Gamma	
$\mathcal{G}(\nu, \theta)$	$\mathcal{G}(\alpha, \beta)$	$\frac{\alpha + \nu}{\beta + x}$
Binomial	Beta	
$\mathcal{B}(n, \theta)$	$\mathcal{B}e(\alpha, \beta)$	$\frac{\alpha + x}{\alpha + \beta + n}$
Negative binomial	Beta	
$\mathcal{N}eg(n, \theta)$	$\mathcal{B}e(\alpha, \beta)$	$\frac{\alpha + n}{\alpha + \beta + x + n}$
Multinomial	Dirichlet	
$\mathcal{M}_k(n; \theta_1, \ldots, \theta_k)$	$\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\frac{\alpha_i + x_i}{\left(\sum_j \alpha_j\right) + n}$
Normal	Gamma	
$\mathcal{N}(\mu, 1/\theta)$	$\mathcal{G}(\alpha/2, \beta/2)$	$\frac{\alpha + 1}{\beta + (\mu - x)^2}$

Illustration

Consider

$$x_1, \ ..., \ x_n \sim \mathcal{U}([0, \theta])$$

and $\theta \sim \mathcal{P}a(\theta_0, \alpha)$. Then

$$\theta | x_1, ..., x_n \sim \mathcal{P}a(\max(\theta_0, x_1, ..., x_n), \alpha + n)$$

and

$$\delta^{\pi}(x_1, ..., x_n) = \frac{\alpha + n}{\alpha + n - 1} \max(\theta_0, x_1, ..., x_n).$$

HPD region

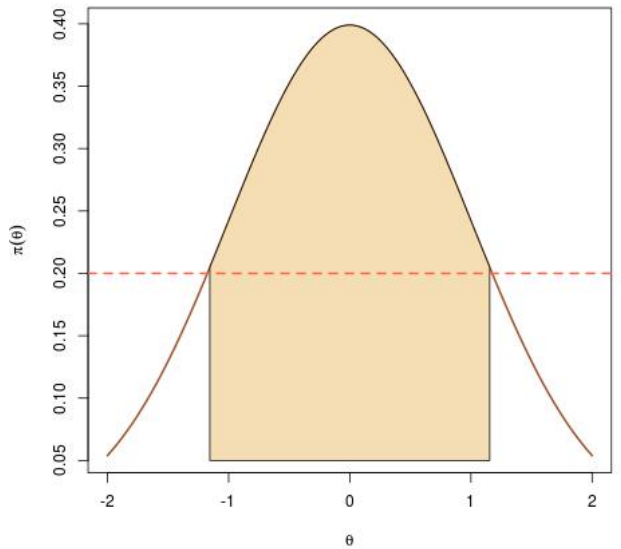
Natural confidence region based on $\pi(\cdot|x)$ is

$$\mathfrak{C}^\pi(x) = \{\theta; \pi(\theta|x) > k\}$$

with

$$\mathbb{P}^\pi(\theta \in \mathfrak{C}^\pi|x) = 1 - \alpha$$

Highest posterior density (HPD) region



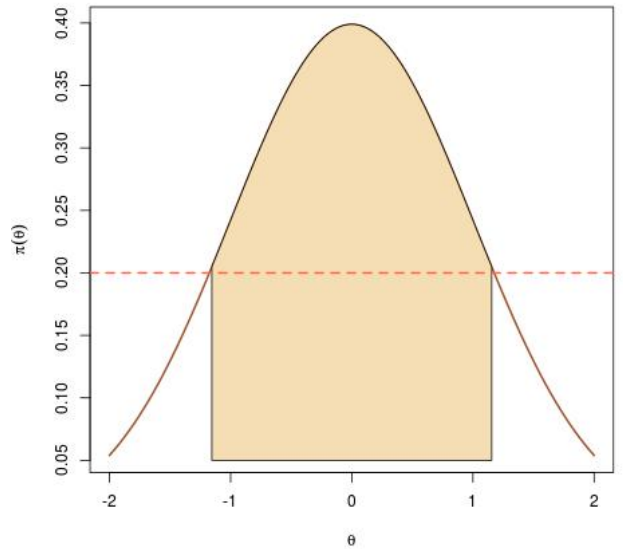
HPD region

Natural confidence region based on $\pi(\cdot|x)$ is

$$\mathfrak{C}^\pi(x) = \{\theta; \pi(\theta|x) > k\}$$

with

$$\mathbb{P}^\pi(\theta \in \mathfrak{C}^\pi|x) = 1 - \alpha$$



Highest posterior density (HPD) region

Example case $x \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(0, 10)$. Then

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

and

$$\begin{aligned}\mathfrak{C}^\pi(x) &= \left\{\theta; |\theta - {}^{10}/_{11}x| > k'\right\} \\ &= ({}^{10}/_{11}x - k', {}^{10}/_{11}x + k')\end{aligned}$$

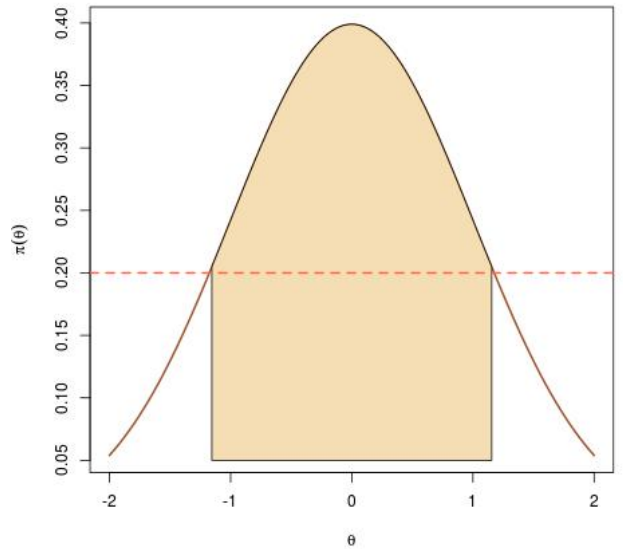
HPD region

Natural confidence region based on $\pi(\cdot|x)$ is

$$\mathfrak{C}^\pi(x) = \{\theta; \pi(\theta|x) > k\}$$

with

$$\mathbb{P}^\pi(\theta \in \mathfrak{C}^\pi|x) = 1 - \alpha$$



Highest posterior density (HPD) region

Warning Frequentist coverage is not $1 - \alpha$, hence name of credible rather than confidence region

Further validation of HPD regions as smallest-volume $1 - \alpha$-coverage regions