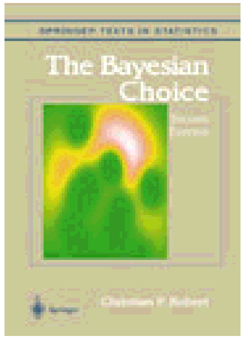
Bayesian Statistics

Christian P. Robert
Université Paris Dauphine

Based on THE BAYESIAN CHOICE



Springer-Verlag 2001

1 Introduction

Parametric model:

Observations $x_1, \ldots, x_n$ generated from a probability distribution $f_i(x_i|\theta_i, x_1, \ldots, x_{i-1}) = f_i(x_i|\theta_i, x_{1:i-1})$

$$x = (x_1, \ldots, x_n) \sim f(x|\theta), \qquad \theta = (\theta_1, \ldots, \theta_n)$$

Associated likelihood

$$\ell(\theta|x) = f(x|\theta)$$

[inverted density]

1.1 The Bayesian framework

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

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

[Thomas Bayes, 1764]

Actualization principle

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*

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

Definition 1 A Bayesian statistical model is made of a parametric statistical model,

$$(\mathcal{X}, f(x|\theta)) \, ,$$

and a prior distribution on the parameters,

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

Justifications

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

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.

Given X, what inference can we make on p?

Modern translation:

Derive the posterior distribution of p given X, when

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

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

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

then

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

i.e.

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

[Beta distribution]

1.2 Prior and posterior distributions

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

(a) the *joint distribution* of (θ, x),

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

(b) the *marginal distribution* of x,

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

(c) the *posterior distribution* of θ,

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

(d) 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 central to Bayesian inference

- Operates conditional upon the observations
- Incorporates the requirement of the Likelihood Principle
- 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

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

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

and

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

Natural confidence region

$$
\begin{aligned}
C &= \{\theta; \pi(\theta|x) > k\} \\
&= \left\{\theta; \left|\theta - \frac{10}{11}x\right| > k'\right\}
\end{aligned}
$$

Highest posterior density (HPD) region

1.3 Improper prior distributions

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

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

Improper prior distribution

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. Generally more acceptable to non-Bayesians, with frequentist justifications, such as:
   (i) *minimaxity*
   (ii) *admissibility*
   (iii) *invariance*
6. Improper priors prefered to vague proper priors such as a $\mathcal{N}(0, 100^2)$ distribution
7. Penalization factor in

$$\min_d \int \mathrm{L}(\theta, d)\pi(\theta) f(x|\theta)\, dx\, d\theta$$

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

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

Example 2 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 ω]

Warning

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

[Lindley, 1990]

Example 3 Consider a $\theta \sim \mathcal{N}(0, \tau^2)$ prior. Then

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

for any (a, b)

Example 4 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 $x \neq 0, n$.

2 From Prior Information to Prior Distributions

The most critical and most criticized point of Bayesian analysis !

Because...

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!

Rather...

The prior is a tool summarizing available information as well as uncertainty related with this information,

And...

Ungrounded prior distributions produce unjustified posterior inference

2.1 Subjective determination

Example 5 Capture-recapture experiment on migrations between zones

Prior information on capture and survival probabilities, p_t and q_{it}

	Time	2	3	4	5	6
p_t	Mean	0.3	0.4	0.5	0.2	0.2
	95% cred. int.	[0.1,0.5]	[0.2,0.6]	[0.3,0.7]	[0.05,0.4]	[0.05,0.4]

	Site	A		B	
	Time	t=1,3,5	t=2,4	t=1,3,5	t=2,4
q_{it}	Mean	0.7	0.65	0.7	0.7
	95% cred. int.	[0.4,0.95]	[0.35,0.9]	[0.4,0.95]	[0.4,0.95]

Corresponding prior modeling

Time	2	3	4	5	6
Dist.	$\mathcal{B}e(6,14)$	$\mathcal{B}e(8,12)$	$\mathcal{B}e(12,12)$	$\mathcal{B}e(3.5,14)$	$\mathcal{B}e(3.5,14)$

Site	A		B	
Time	t=1,3,5	t=2,4	t=1,3,5	t=2,4
Dist.	$\mathcal{B}e(6.0,2.5)$	$\mathcal{B}e(6.5,3.5)$	$\mathcal{B}e(6.0,2.5)$	$\mathcal{B}e(6.0,2.5)$

Strategies for prior determination

- 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

- Select a **maximum entropy** prior when prior characteristics are known:

$$\mathbb{E}^\pi[g_k(\theta)] = \omega_k \qquad (k = 1, \ldots, K)$$

with solution, in the discrete case

$$\pi^*(\theta_i) = \frac{\exp\left\{\sum_1^K \lambda_k g_k(\theta_i)\right\}}{\sum_j \exp\left\{\sum_1^K \lambda_k g_k(\theta_j)\right\}},$$

and, in the continuous case,

$$\pi^*(\theta) = \frac{\exp\left\{\sum_1^K \lambda_k g_k(\theta)\right\} \pi_0(\theta)}{\int \exp\left\{\sum_1^K \lambda_k g_k(\eta)\right\} \pi_0(d\eta)},$$

the λ_k's being Lagrange multipliers and π_0 a reference measure [Caveat]

- **Parametric approximations**

  Restrict choice of π to a *parameterised* density

$$\pi(\theta|\lambda)$$

and determine the corresponding (hyper-)parameters

$$\lambda$$

through the *moments* or *quantiles* of π

Example 6 For the normal model $x \sim \mathcal{N}(\theta, 1)$, ranges of the posterior moments for fixed prior moments $\mu_1 = 0$ and μ_2.

		Minimum	Maximum	Maximum
μ_2	x	mean	mean	variance
3	0	-1.05	1.05	3.00
3	1	-0.70	1.69	3.63
3	2	-0.50	2.85	5.78
1.5	0	-0.59	0.59	1.50
1.5	1	-0.37	1.05	1.97
1.5	2	-0.27	2.08	3.80

[Goutis, 1990]

2.2 Conjugate priors

Specific parametric family with analytical properties

Definition 2 *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 $\mathcal{F}$.*

[Raiffa and Schlaifer (1961)]

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 $\pi(\theta)$
- Exchangeability motivations
- Device of virtual past observations
- Linearity of some estimators
- Tractability and simplicity
- First approximations to adequate priors, backed up by robustness analysis

Exponential families

Definition 3 *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 $\Theta \subset \mathrm{I\!R}^k$, $\mathcal{X} \subset \mathrm{I\!R}^k$ and

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

the family is said to be natural.

Interesting analytical properties :

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

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

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

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Negative Binomial $\mathcal{N}eg(m,\theta)$	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

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

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. $f(x|\theta)$,

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

But...

Example 7 When $x \sim \mathcal{B}e(\alpha, \theta)$ with known α,

$$f(x|\theta) \propto \frac{\Gamma(\alpha+\theta)(1-x)^\theta}{\Gamma(\theta)} \,,$$

conjugate distribution not so easily manageable

$$\pi(\theta|x_0, \lambda) \propto \left(\frac{\Gamma(\alpha+\theta)}{\Gamma(\theta)}\right)^\lambda (1-x_0)^\theta$$

Example 8 Coin spun on its edge, proportion θ of *heads*

When spinning n times a given coin, number of heads

$$x \sim \mathcal{B}(n, \theta)$$

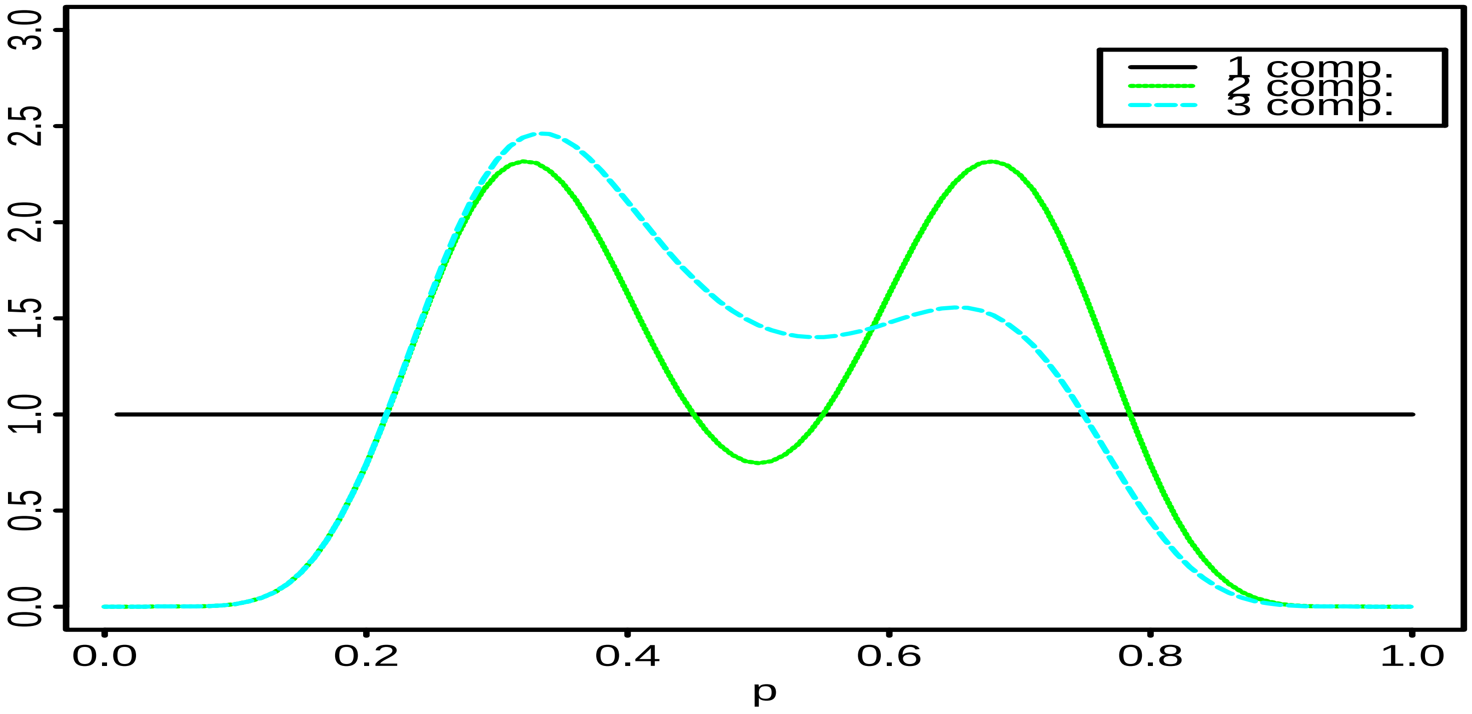
Flat prior, or mixture prior

$$\frac{1}{2}\left[\mathcal{B}e(10, 20) + \mathcal{B}e(20, 10)\right]$$

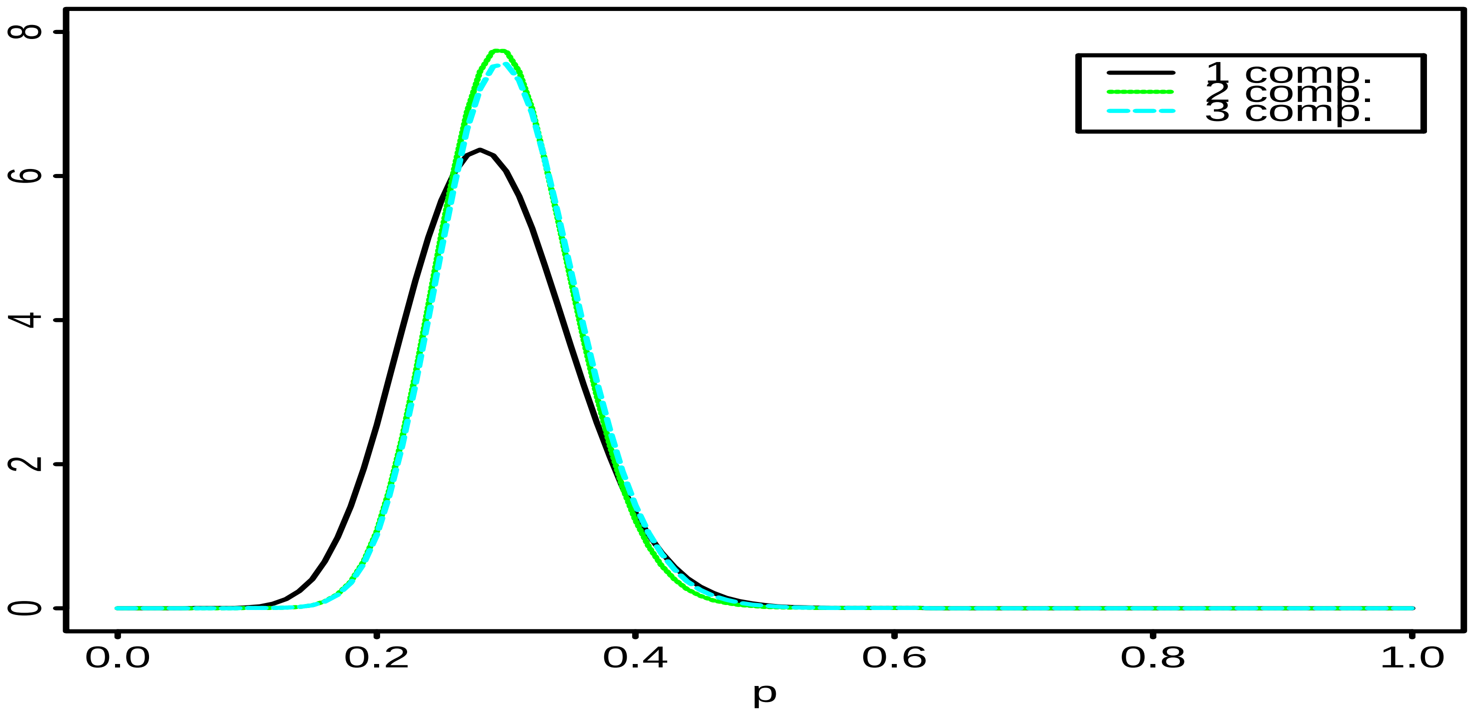
or

$$0.5\,\mathcal{B}e(10, 20) + 0.2\,\mathcal{B}e(15, 15) + 0.3\,\mathcal{B}e(20, 10).$$

Mixtures of natural conjugate distributions also make conjugate families



Three prior distributions for a spinning-coin experiment



Posterior distributions for 50 observations

2.3 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]

Re-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]

2.3.1 Laplace’s prior

Principle of *Insufficient Reason* (Laplace)

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

Extension to continuous spaces

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

- Lack of reparameterization invariance/coherence

$$\psi = e^{\theta} \qquad \pi_1(\psi) = \frac{1}{\psi} \neq \pi_2(\psi) = 1$$

- Problems of properness

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

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

2.3.2 Invariant priors

Principle: Agree with the natural symmetries of the problem

- Identify invariance structures as group action

$$\begin{aligned}
\mathcal{G} &: \quad x \longrightarrow g(x) \sim f(g(x)|\bar{g}(\theta)) \\
\bar{\mathcal{G}} &: \quad \theta \longrightarrow \bar{g}(\theta) \\
\mathcal{G}^* &: \quad L(d,\theta) = L(g^*(d), \bar{g}(\theta))
\end{aligned}$$

- Determine an invariant prior

$$\pi(\bar{g}(A)) = \pi(A)$$

Solution: Right Haar measure

But...

- Requires invariance to be part of the decision problem
- Missing in most discrete setups (Poisson)

2.3.3 The Jeffreys prior

Based on Fisher information

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

The Jeffreys prior distribution is

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

Pros & Cons

- Relates to information theory
- Agrees with most invariant priors
- Parameterization invariant
- Suffers from dimensionality curse
- Not coherent for Likelihood Principle

Example 9

$$x \sim \mathcal{N}_p(\theta, I_p), \qquad \eta = \|\theta\|^2, \qquad \pi(\eta) = \eta^{p/2-1}$$

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

Example 10 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

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

2.3.4 Reference priors

Generalizes Jeffreys priors by distinguishing between nuisance and interest parameters

Principle: maximize the information brought by the data

$$\mathbb{E}^n \left[\int \pi(\theta|x_n) \log(\pi(\theta|x_n)/\pi(\theta)) d\theta \right]$$

and consider the limit of the π_n

Outcome: most usually, Jeffreys prior

Nuisance parameters:

For $\theta = (\lambda, \omega)$,

$$\pi(\lambda|\omega) = \pi_J(\lambda|\omega) \qquad \text{with fixed } \omega$$

Jeffreys' prior conditional on ω, and

$$\pi(\omega) = \pi_J(\omega)$$

for the marginal model

$$f(x|\omega) \propto \int f(x|\theta)\pi_J(\lambda|\omega)d\lambda$$

- Depends on ordering
- Problems of definition

Example 11 Neyman–Scott problem

Observation of x_{ij} iid $\mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, 2$.

The usual Jeffreys prior for this model is

$$\pi(\mu_1, \ldots, \mu_n, \sigma) = \sigma^{-n-1}$$

which is inconsistent because

$$\mathbb{E}[\sigma^2 | x_{11}, \ldots, x_{n2}] = s^2/(2n-2),$$

where

$$s^2 = \sum_{i=1}^{n} \frac{(x_{i1} - x_{i2})^2}{2},$$

Associated reference prior with $\theta_1 = \sigma$ and $\theta_2 = (\mu_1, \ldots, \mu_n)$ gives

$$\begin{aligned} \pi(\theta_2|\theta_1) &\propto 1\,, \\ \pi(\sigma) &\propto 1/\sigma \end{aligned}$$

Therefore,

$$\mathbb{E}[\sigma^2|x_{11}, \ldots, x_{n2}] = s^2/(n-2)$$

2.3.5 Matching priors

Frequency-validated priors:

Some posterior probabilities

$$\pi(g(\theta) \in C_x | x) = 1 - \alpha$$

must coincide with the corresponding frequentist coverage

$$P_\theta(C_x \ni g(\theta)) = \int \mathbb{I}_{C_x}(g(\theta))\, f(x|\theta)\, dx\,,$$

...asymptotically

For instance, Welch and Peers' identity

$$P_\theta(\theta \leq k_\alpha(x)) = 1 - \alpha + O(n^{-1/2})$$

and for Jeffreys' prior,

$$P_\theta(\theta \leq k_\alpha(x)) = 1 - \alpha + O(n^{-1})$$

In general, choice of a matching prior dictated by the cancelation of a first order term in an **Edgeworth expansion**, like

$$[I''(\theta)]^{-1/2} I'(\theta) \nabla \log \pi(\theta) + \nabla^t \{ I'(\theta) [I''(\theta)]^{-1/2} \} = 0\,.$$

Example 12 **Linear calibration model**

$$y_i = \alpha+\beta x_i+\varepsilon_i, \qquad y_{0j} = \alpha+\beta x_0+\varepsilon_{0j}\,, \qquad (i = 1,\ldots,n,\ j = 1,\ldots,k)$$

with $\theta = (x_0, \alpha, \beta, \sigma^2)$ and x_0 quantity of interest

One-sided differential equation:

$$
\begin{aligned}
&|\beta|^{-1} s^{-1/2} \frac{\partial}{\partial x_0}\{e(x_0)\pi(\theta)\} - e^{-1/2}(x_0)\operatorname{sgn}(\beta) n^{-1} s^{1/2} \frac{\partial \pi(\theta)}{\partial x_0} \\
&\quad - e^{-1/2}(x_0)(x_0 - \bar{x}) s^{-1/2} \frac{\partial}{\partial \beta}\{\operatorname{sgn}(\beta)\pi(\theta)\} = 0
\end{aligned}
$$

with

$$
s = \Sigma(x_i - \bar{x})^2, \quad e(x_0) = [(n+k)s + nk(x_0 - \bar{x})^2]/nk \, .
$$

Solutions

$$\pi(x_0, \alpha, \beta, \sigma^2) \propto e(x_0)^{(d-1)/2} |\beta|^d g(\sigma^2)\,,$$

where g arbitrary.

Reference priors

Partition	Prior
$(x_0, \alpha, \beta, \sigma^2)$	$\lvert\beta\rvert(\sigma^2)^{-5/2}$
$x_0, \alpha, \beta, \sigma^2$	$e(x_0)^{-1/2}(\sigma^2)^{-1}$
$x_0, \alpha, (\sigma^2, \beta)$	$e(x_0)^{-1/2}(\sigma^2)^{-3/2}$
$x_0, (\alpha, \beta), \sigma^2$	$e(x_0)^{-1/2}(\sigma^2)^{-1}$
$x_0, (\alpha, \beta, \sigma^2)$	$e(x_0)^{-1/2}(\sigma^2)^{-2}$

2.3.6 Other approaches

- Rissanen’s transmission information theory and minimum length priors
- Testing priors
- stochastic complexity

3 Decision-Theoretic Foundations of Statistical Inference

3.1 Evaluating estimators

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

Requires an evaluation criterion for decisions and estimators

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

[loss function]

Bayesian Decision Theory

Three spaces/factors:

(1) On $\mathcal{X}$, distribution for the observation, $f(x|\theta)$;

(2) On Θ, prior distribution for the parameter, $\pi(\theta)$;

(3) On $\Theta \times \mathcal{D}$, loss function associated with the decisions, $\mathrm{L}(\theta, \delta)$;

Foundations

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

[DeGroot, 1970]

3.2 Loss functions

Decision procedure δ usually called estimator

(while its *value* $\delta(x)$ called estimate of θ)

Impossible to uniformly minimize (in d) the loss function

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

when θ is unknown

Frequentist Principle

Average loss (or frequentist risk)

$$\begin{aligned} R(\theta,\delta) &= \mathbb{E}_\theta[\mathrm{L}(\theta,\delta(x))] \\ &= \int_{\mathcal{X}} \mathrm{L}(\theta,\delta(x))\, f(x|\theta)\, dx \end{aligned}$$

Principle Select the best estimator based on the risk function

Difficulties with frequentist paradigm

(1) Error averaged over the different values of x proportionally to the density $f(x|\theta)$: not so appealing for a client, who wants optimal results for **her** data x!

(2) Assumption of repeatability of experiments not always grounded.

(3) $R(\theta, \delta)$ is a function of θ: there is no total ordering on the set of procedures.

Bayesian principle

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

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

Bayesian principle (contd.)

Alternative Integrate over the space Θ and compute *integrated risk*

$$\begin{aligned} r(\pi, \delta) &= \mathbb{E}^{\pi}[R(\theta, \delta)] \\ &= \int_{\Theta} \int_{\mathcal{X}} \mathrm{L}(\theta, \delta(x))\, f(x|\theta)\, dx\, \pi(\theta)\, d\theta \end{aligned}$$

which induces a total ordering on estimators.

Therefore,

$$\boxed{\textbf{existence of an optimal decision}}$$

Bayes estimator

An estimator minimizing

$$r(\pi, \delta)$$

can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ which minimizes

$$\rho(\pi, \delta | x)$$

since

$$r(\pi, \delta) = \int_{\mathcal{X}} \rho(\pi, \delta(x)|x) m(x)\, dx.$$

Both approaches give the same estimator

Definition 4 *A* **Bayes estimator** *associated with a prior distribution* π *and a loss function* L *is*

$$\arg\min_{\delta} r(\pi, \delta)$$

The value $r(\pi) = r(\pi, \delta^{\pi})$ *is called the Bayes risk*

Infinite Bayes risk

Above result valid for both proper and improper priors when

$$r(\pi) < \infty$$

Otherwise, **generalized Bayes estimator** defined pointwise:

$$\delta^\pi(x) = \arg\min_d \rho(\pi, d|x)$$

if $\rho(\pi, d|x)$ is well-defined for every x.

Warning: generalized Bayes $\neq$ improper Bayes

3.3 Minimaxity and admissibility

3.3.1 Minimaxity

Insurance against the worst case and total ordering on $\mathcal{D}^*$

Definition 5 *The minimax risk associated with a loss* L *is*

$$\bar{R} = \inf_{\delta \in \mathcal{D}^*} \sup_\theta R(\theta, \delta) = \inf_{\delta \in \mathcal{D}^*} \sup_\theta \mathbb{E}_\theta[L(\theta, \delta(x))],$$

and a minimax estimator is any estimator δ_0 *such that*

$$\sup_\theta R(\theta, \delta_0) = \bar{R}.$$

Criticisms

- Reasons in terms of the worst case
- Does not incorporate prior information
- Too conservative
- Difficult to exhibit

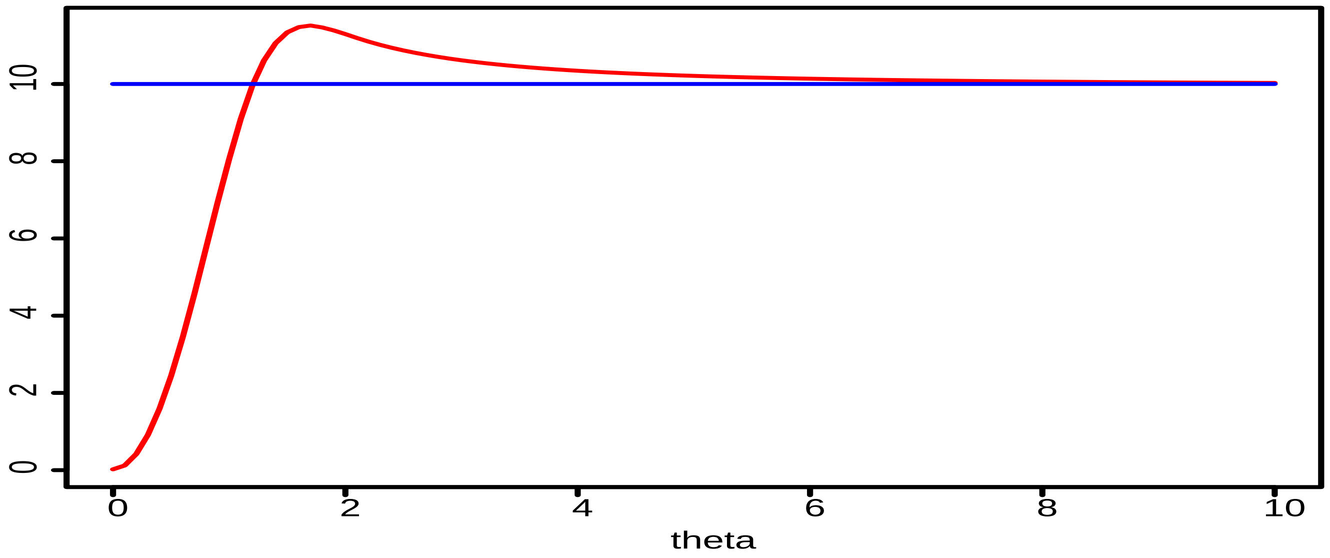
Example 13 Consider

$$\delta_2(x) = \begin{cases} \left(1 - \dfrac{2p-1}{||x||^2}\right) x & \text{if } ||x||^2 \geq 2p-1, \\ 0 & \text{otherwise,} \end{cases}$$

to estimate θ when $x \sim \mathcal{N}_p(\theta, I_p)$ under *quadratic loss*,

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

Comparison of δ_2 with $\delta_1(x) = x$, maximum likelihood estimator, for $p = 10$.



δ_2 cannot be minimax

3.3.2 Minimax vs. maximin

Existence:

If $\mathcal{D} \subset \mathbb{R}^k$ convex and compact, and if $\mathrm{L}(\theta, d)$ continuous and convex as a function of d for every $\theta \in \Theta$, there exists a nonrandomized minimax estimator.

Connection with Bayesian approach

The Bayes risks are always smaller than the minimax risk:

$$\underline{r} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) \leq \overline{r} = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

$\underline{r}$ maximin risk

least favourable prior

Definition 6 *The estimation problem has a value when $\underline{r} = \overline{r}$, i.e.*

$$\sup_{\pi} \inf_{\delta \in \mathcal{D}} r(\pi, \delta) = \inf_{\delta \in \mathcal{D}^*} \sup_{\theta} R(\theta, \delta).$$

When the problem has a value, some minimax estimators are Bayes estimators for the least favourable distributions.

Example 14 Consider $x \sim \mathcal{B}e(\theta)$ with

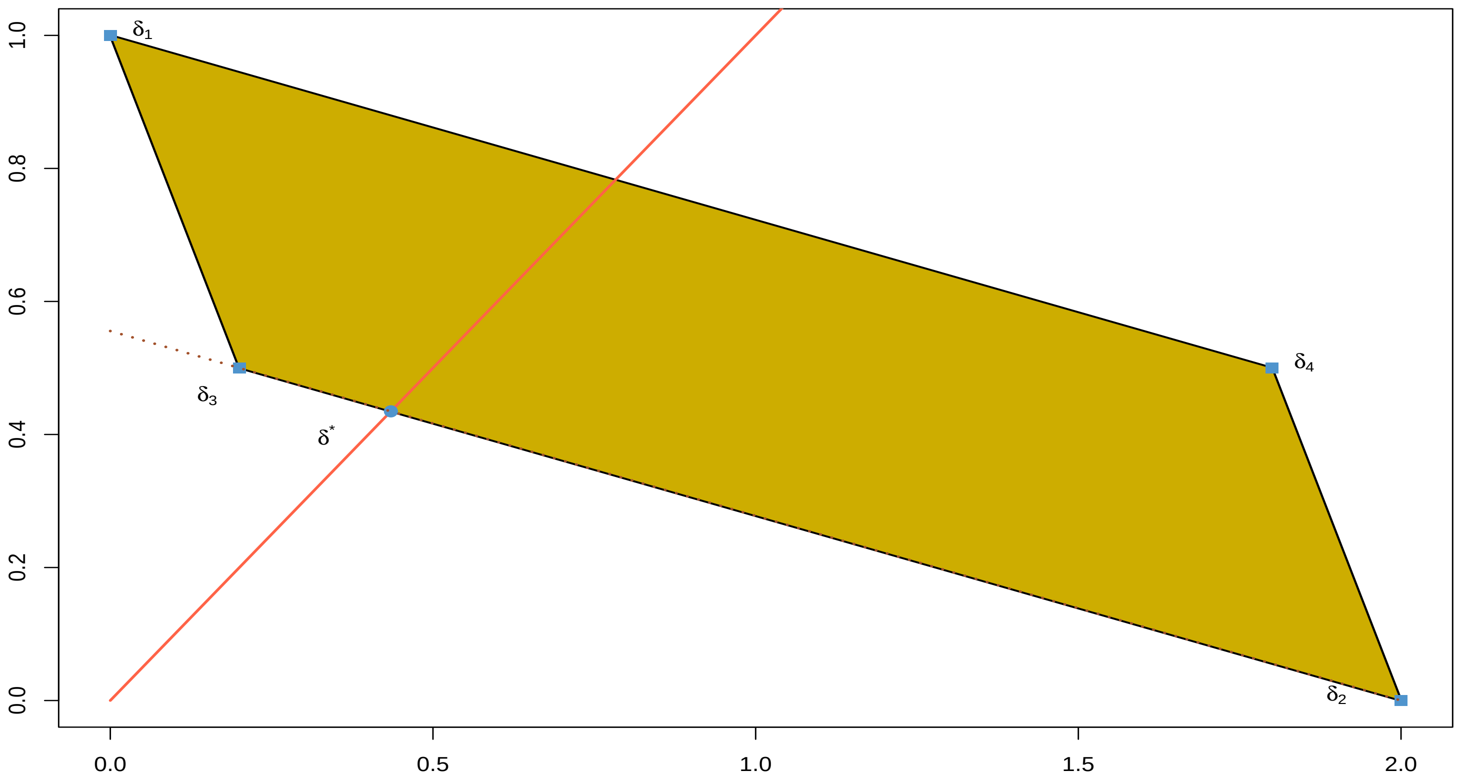
$$\theta \in \{0.1, 0.5\}$$

and

$$\begin{aligned} \delta_1(x) &= 0.1, \qquad\qquad \delta_2(x) = 0.5, \\ \delta_3(x) &= 0.1\,\mathbb{I}_{x=0} + 0.5\,\mathbb{I}_{x=1}, \quad \delta_4(x) = 0.5\,\mathbb{I}_{x=0} + 0.1\,\mathbb{I}_{x=1}. \end{aligned}$$

under

$$\mathrm{L}(\theta, d) = \begin{cases} 0 & \text{if } d = \theta \\ 1 & \text{if } (\theta, d) = (0.5, 0.1) \\ 2 & \text{if } (\theta, d) = (0.1, 0.5) \end{cases}$$


δ_1
δ_2
δ_3
δ_4
δ^*
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0

Risk set

Minimax estimator at the intersection of the diagonal of $\mathbb{R}^2$ with the lower boundary of $\mathcal{R}$:

$$\delta^*(x) = \begin{cases} \delta_3(x) & \text{with probability } \alpha = 0.87, \\ \delta_2(x) & \text{with probability } 1 - \alpha. \end{cases}$$

Also randomized Bayes estimator for

$$\pi(\theta) = 0.22\, \mathbb{I}_{0.1}(\theta) + 0.78\, \mathbb{I}_{0.5}(\theta)$$

Checking minimaxity

If δ_0 is a Bayes estimator for π_0 and if

$$R(\theta, \delta_0) \leq r(\pi_0)$$

for every θ in the support of π_0, δ_0 is minimax and π_0 is the least favourable distribution.

Example 15 Consider $x \sim \mathcal{B}(n, \theta)$ for the loss

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

When $\theta \sim \mathcal{B}e\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$, the posterior mean is

$$\delta^*(x) = \frac{x + \sqrt{n}/2}{n + \sqrt{n}}.$$

with *constant risk*

$$R(\theta, \delta^*) = 1/4(1 + \sqrt{n})^2.$$

Therefore, δ^* is minimax

[H. Rubin]

Checking minimaxity (contd.)

If for a sequence (π_n) of proper priors, the generalized Bayes estimator δ_0 satisfies

$$R(\theta, \delta_0) \leq \lim_{n \to \infty} r(\pi_n) < +\infty$$

for every $\theta \in \Theta$, then δ_0 is minimax.

Example 16 When $x \sim \mathcal{N}(\theta, 1)$,

$$\delta_0(x) = x$$

is a generalized Bayes estimator associated with

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

Since, for $\pi_n(\theta) = \exp\{-\theta^2/2n\}$,

$$\begin{aligned} R(\delta_0, \theta) &= \mathbb{E}_\theta\left[(x-\theta)^2\right] = 1 \\ &= \lim_{n\to\infty} r(\pi_n) = \lim_{n\to\infty} \frac{n}{n+1} \end{aligned}$$

δ_0 is minimax.

3.3.3 Admissibility

Reduction of the set of estimators based on “local” properties

Definition 7 *An estimator δ_0 is inadmissible if there exists an estimator δ_1 such that, for every θ,*

$$R(\theta, \delta_0) \geq R(\theta, \delta_1)$$

and, for at least one θ_0

$$R(\theta_0, \delta_0) > R(\theta_0, \delta_1)$$

Minimaxity & admissibility

If there exists a unique minimax estimator, this estimator is admissible.

The converse is false!

If δ_0 is admissible with constant risk, δ_0 is the unique minimax estimator.

The converse is false!

Bayesian perspective

Admissibility strongly related to the Bayes paradigm: Bayes estimators often constitute the class of admissible estimators

- If π is strictly positive on Θ, with

$$r(\pi) = \int_\Theta R(\theta, \delta^\pi)\pi(\theta)\, d\theta < \infty$$

  and $R(\theta, \delta)$, is continuous, then the Bayes estimator δ^π is admissible.

- If the Bayes estimator associated with a prior π is unique, it is admissible.

Regular ($\neq$generalized) Bayes estimators always are admissible

Example 17 Consider $x \sim \mathcal{N}(\theta, 1)$ and the test of $H_0 : \theta \leq 0$, i.e. the estimation of

$$\mathbb{I}_{H_0}(\theta)$$

Under the loss

$$\left(\mathbb{I}_{H_0}(\theta) - \delta(x)\right)^2,$$

the estimator (p-value)

$$\begin{aligned} p(x) &= P_0(X > x) \qquad (X \sim \mathcal{N}(0,1)) \\ &= 1 - \Phi(x), \end{aligned}$$

is Bayes under Lebesgue measure.

Indeed

$$\begin{aligned} p(x) &= \mathbb{E}^\pi[\mathbb{I}_{H_0}(\theta)|x] = P^\pi(\theta < 0|x) \\ &= P^\pi(\theta - x < -x|x) = 1 - \Phi(x). \end{aligned}$$

The Bayes risk of p is finite and $p(s)$ is admissible.

Example 18 Consider $x \sim \mathcal{N}(\theta, 1)$. Then $\delta_0(x) = x$ is a generalized Bayes estimator, is admissible, but

$$\begin{aligned} r(\pi, \delta_0) &= \int_{-\infty}^{+\infty} R(\theta, \delta_0)\, d\theta \\ &= \int_{-\infty}^{+\infty} 1\, d\theta = +\infty. \end{aligned}$$

Example 19 Consider $x \sim \mathcal{N}_p(\theta, I_p)$. If

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

the Bayes estimator for the Lebesgue measure is

$$\delta^\pi(x) = ||x||^2 + p.$$

This estimator is not admissible because it is dominated by

$$\delta_0(x) = ||x||^2 - p$$

3.4 Usual loss functions

3.4.1 The quadratic loss

Historically, first loss function (Legendre, Gauss)

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

Proper loss

The Bayes estimator δ^π associated with the prior π and with the quadratic loss is the posterior expectation

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

3.4.2 The absolute error loss

Alternatives to the quadratic loss:

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

or

$$\mathrm{L}_{k_1,k_2}(\theta, d) = \begin{cases} k_2(\theta - d) & \text{if } \theta > d\text{ ,} \\ k_1(d - \theta) & \text{otherwise.} \end{cases} \tag{1}$$

The Bayes estimator associated with π and (1) is a $(k_2/(k_1 + k_2))$ fractile of $\pi(\theta|x)$.

3.4.3 The $0-1$ loss

Neyman–Pearson loss for testing hypotheses

Test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \notin \Theta_0$.

Then

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

The $0-1$ loss

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

Associated with the risk

$$\begin{aligned} R(\theta,\delta) &= \mathbb{E}_\theta[\mathrm{L}(\theta,\delta(x))] \\ &= \begin{cases} P_\theta(\delta(x)=0) & \text{if } \theta\in\Theta_0\,, \\ P_\theta(\delta(x)=1) & \text{otherwise,} \end{cases} \end{aligned}$$

type–one and type–two errors

The Bayes estimator associated with π and with the $0-1$ loss is

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

3.4.4 Intrinsic losses

Noninformative settings w/o natural parameterisation : the estimators should be invariant under reparameterisation

[Ultimate invariance!]

Corresponding parameterisation-free loss functions:

$$\mathrm{L}(\theta, \delta) = d(f(\cdot|\theta), f(\cdot|\delta)),$$

Examples:

(1) the *entropy distance* (or *Kullback–Leibler divergence*)

$$\mathrm{L_e}(\theta,\delta) = \mathbb{E}_\theta \left[\log \left(\frac{f(x|\theta)}{f(x|\delta)} \right) \right],$$

(2) the *Hellinger distance*

$$\mathrm{L_H}(\theta,\delta) = \frac{1}{2} \mathbb{E}_\theta \left[\left(\sqrt{\frac{f(x|\delta)}{f(x|\theta)}} - 1 \right)^2 \right].$$

Example 20 Consider $x \sim \mathcal{N}(\theta, 1)$. Then

$$\begin{aligned} \mathrm{L}_{\mathrm{e}}(\theta,\delta) &= \frac{1}{2}\mathbb{E}_\theta[-(x-\theta)^2 + (x-\delta)^2] = \frac{1}{2}(\delta-\theta)^2, \\ \mathrm{L}_{\mathrm{H}}(\theta,\delta) &= 1 - \exp\{-(\delta-\theta)^2/8\}. \end{aligned}$$

When $\pi(\theta|x)$ is a $\mathcal{N}(\mu(x),\sigma^2)$ distribution, the Bayes estimator of θ is

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

in both cases.

4 Admissibility and Complete Classes

4.1 Admissibility of Bayes estimators

Bayes estimators may be inadmissible when the Bayes risk is infinite

Example 21 Consider $x \sim \mathcal{N}(\theta, 1)$ with a conjugate prior $\theta \sim \mathcal{N}(0, \sigma^2)$ and loss

$$\mathrm{L}_\alpha(\theta, \delta) = e^{\theta^2/2\alpha}(\theta - \delta)^2,$$

The associated generalized Bayes estimator is defined for $\alpha > \frac{\sigma^2}{\sigma^2+1}$ and

$$\begin{aligned} \delta_\alpha^\pi(x) &= \frac{\sigma^2 + 1}{\sigma^2}\left(\frac{\sigma^2 + 1}{\sigma^2} - \alpha^{-1}\right)^{-1} \delta^\pi(x) \\ &= \frac{\alpha}{\alpha - \frac{\sigma^2}{\sigma^2+1}} \delta^\pi(x). \end{aligned}$$

The corresponding Bayes risk is

$$r(\pi) = \int_{-\infty}^{+\infty} e^{\theta^2/2\alpha} e^{-\theta^2/2\sigma^2} d\theta,$$

that is, is infinite for $\alpha \leq \sigma^2$. Moreover, since $\delta_\alpha^\pi(x) = cx$ with $c > 1$ when

$$\alpha > \alpha \frac{\sigma^2 + 1}{\sigma^2} - 1,$$

δ_α^π is inadmissible

Formal admissibility result

If Θ is a discrete set and $\pi(\theta) > 0$ for every $\theta \in \Theta$, then there exists an admissible Bayes estimator associated with π

4.1.1 Boundary conditions

If

$$f(x|\theta) = h(x)e^{\theta.T(x)-\psi(\theta)}, \qquad \theta \in [\underline{\theta}, \bar{\theta}]$$

and π is a conjugate prior,

$$\pi(\theta|t_0, \lambda) = e^{\theta.t_0 - \lambda\psi(\theta)}$$

a sufficient condition for $\mathbb{E}^\pi[\nabla\psi(\theta)|x]$ to be admissible is that, for every $\underline{\theta} < \theta_0 < \bar{\theta}$,

$$\int_{\theta_0}^{\bar{\theta}} e^{-\gamma_0\lambda\theta+\lambda\psi(\theta)}\, d\theta = \int_{\underline{\theta}}^{\theta_0} e^{-\gamma_0\lambda\theta+\lambda\psi(\theta)}\, d\theta = +\infty.$$

Example 22 Consider $x \sim \mathcal{B}(n, p)$.

$$f(x|\theta) = \binom{n}{x} e^{(x/n)\theta} \left(1 + e^{\theta/n}\right)^{-n} \qquad \theta = n \log(p/1 - p)$$

Then the two integrals

$$\int_{-\infty}^{\theta_0} e^{-\gamma_0 \lambda\theta} \left(1 + e^{\theta/n}\right)^{\lambda n} d\theta \quad \text{and} \quad \int_{\theta_0}^{+\infty} e^{-\gamma_0 \lambda\theta} \left(1 + e^{\theta/n}\right)^{\lambda n} d\theta$$

cannot diverge simultaneously if $\lambda < 0$.

For $\lambda > 0$, second integral divergent if $\lambda(1 - \gamma_0) > 0$ and first integral divergent if $\gamma_0 \lambda \geq 0$.

Admissible Bayes estimators of p

$$\delta^\pi(x) = a\frac{x}{n} + b, \qquad 0 \leq a \leq 1, \quad b \geq 0, \quad a + b \leq 1.$$

4.1.2 Differential representations

Setting of multidimensional exponential families

$$f(x|\theta) = h(x)e^{\theta.x-\psi(\theta)}, \qquad \theta \in \mathbb{R}^p$$

Measure g such that

$$I_x(\nabla g) = \int ||\nabla g(\theta)|| e^{\theta.x-\psi(\theta)} \, d\theta < +\infty$$

Representation of the posterior mean of $\nabla\psi(\theta)$

$$\delta_g(x) = x + \frac{I_x(\nabla g)}{I_x(g)}.$$

Sufficient admissibility conditions:

$$\begin{aligned}\int_{\{||\theta||>1\}} \frac{g(\theta)}{||\theta||^2 \log^2(||\theta|| \vee 2)} d\theta &< \infty, \\ \int \frac{||\nabla g(\theta)||^2}{g(\theta)} d\theta &< \infty,\end{aligned}$$

and

$$\forall \theta \in \Theta, \qquad R(\theta, \delta_g) < \infty,$$

Consequence

If

$$\Theta = \mathbb{R}^p \qquad p \leq 2$$

the estimator

$$\delta_0(x) = x$$

is admissible.

Example 23 If $x \sim \mathcal{N}_p(\theta, I_p)$, $p \leq 2$, $\delta_0(x) = x$ is admissible.

4.1.3 Recurrence conditions

Special case of $\mathcal{N}_p(\theta, \Sigma)$:

A generalized Bayes estimator of the form

$$\delta(x) = (1 - h(||x||))x$$

is

(i) inadmissible if there exist $\epsilon > 0$ and $K < +\infty$ such that, for $||x|| > K$,

$$||x||^2 h(||x||) < p - 2 - \epsilon;$$

and

(ii) admissible if there exist K_1 and K_2 such that $h(||x||)||x|| \leq K_1$ for every x and, for $||x|| > K_2$,

$$||x||^2 h(||x||) \geq p - 2.$$

[Brown, 1971]

General case:

Estimation of a **bounded** function $g(\theta)$

For a given prior π, Markovian transition kernel

$$K(\theta|\eta) = \int_{\mathcal{X}} \pi(\theta|x) f(x|\eta)\, dx,$$

The generalized Bayes estimator of $g(\theta)$ is admissible if the associated Markov chain $(\theta^{(n)})$ is π-recurrent.

[Eaton, 1994]

Extension to the **unbounded case**, based on the (case dependent) transition kernel

$$T(\theta|\eta) = \Psi(\eta)^{-1}(\varphi(\theta) - \varphi(\eta))^2 K(\theta|\eta)\,,$$

where $\Psi(\theta)$ normalizing factor

The generalized Bayes estimator of $\varphi(\theta)$ is admissible if the associated Markov chain $(\theta^{(n)})$ is π-recurrent.

[Eaton, 1999]

4.2 Necessary and sufficient admissibility conditions

Formalisation of the statement that...

...all admissible estimators are limits of Bayes estimators...

4.2.1 Blyth's sufficient condition

If, for an estimator δ_0, there exists a sequence (π_n) of generalized prior distributions such that

(i) $r(\pi_n, \delta_0)$ is finite for every n;

(ii) for every nonempty open set $C \subset \Theta$, there exist $K > 0$ and N such that, for every $n \geq N$, $\pi_n(C) \geq K$; and

(iii) $\lim_{n \to +\infty} r(\pi_n, \delta_0) - r(\pi_n) = 0$;

then δ_0 is admissible.

Example 24 Consider $x \sim \mathcal{N}(\theta, 1)$ and $\delta_0(x) = x$

Choose π_n as the measure with density

$$g_n(x) = e^{-\theta^2/2n}$$

[condition (ii) is satisfied]

The Bayes estimator for π_n is

$$\delta_n(x) = \frac{nx}{n+1},$$

and

$$\begin{aligned} r(\pi_n) &= \int_{\mathbb{R}} \left[\frac{\theta^2}{(n+1)^2} + \frac{n^2}{(n+1)^2} \right] g_n(\theta)\, d\theta \\ &= \sqrt{2\pi n}\, \frac{n}{n+1}, \end{aligned}$$

[condition (i) is satisfied]

while

$$r(\pi_n, \delta_0) = \int_{\mathbb{R}} 1 \, g_n(\theta) \, d\theta = \sqrt{2\pi n}.$$

Moreover,

$$r(\pi_n, \delta_0) - r(\pi_n) = \sqrt{2\pi n}/(n+1)$$

converges to 0.

[condition (iii) is satisfied]

4.2.2 Stein’s necessary and sufficient condition

Assumptions

(i) $f(x|\theta)$ is continuous in θ and strictly positive on Θ; and

(ii) the loss L is strictly convex, continuous and, if $E \subset \Theta$ is compact,

$$\lim_{\|\delta\|\rightarrow+\infty} \inf_{\theta\in E} \mathrm{L}(\theta,\delta) = +\infty.$$

Stein's n&s condition:

δ is admissible **iff** there exist

1. a sequence (F_n) of increasing compact sets such that
$$\Theta = \bigcup_n F_n,$$
2. a sequence (π_n) of finite measures with support F_n, and
3. a sequence (δ_n) of Bayes estimators associated with π_n

such that

(i) there exists a compact set $E_0 \subset \Theta$ such that $\inf_n \pi_n(E_0) \geq 1$;

(ii) if $E \subset \Theta$ is compact, $\sup\limits_n \pi_n(E) < +\infty$;

(iii) $\lim\limits_n r(\pi_n, \delta) - r(\pi_n) = 0$; and

(iv) $\lim\limits_n R(\theta, \delta_n) = R(\theta, \delta)$.

4.3 Complete classes

Definition 8 *A class $\mathcal{C}$ of estimators is* complete *if, for every $\delta' \notin \mathcal{C}$, there exists $\delta \in \mathcal{C}$ that dominates δ'. The class is* essentially complete *if, for every $\delta' \notin \mathcal{C}$, there exists $\delta \in \mathcal{C}$ that is at least as good as δ'.*

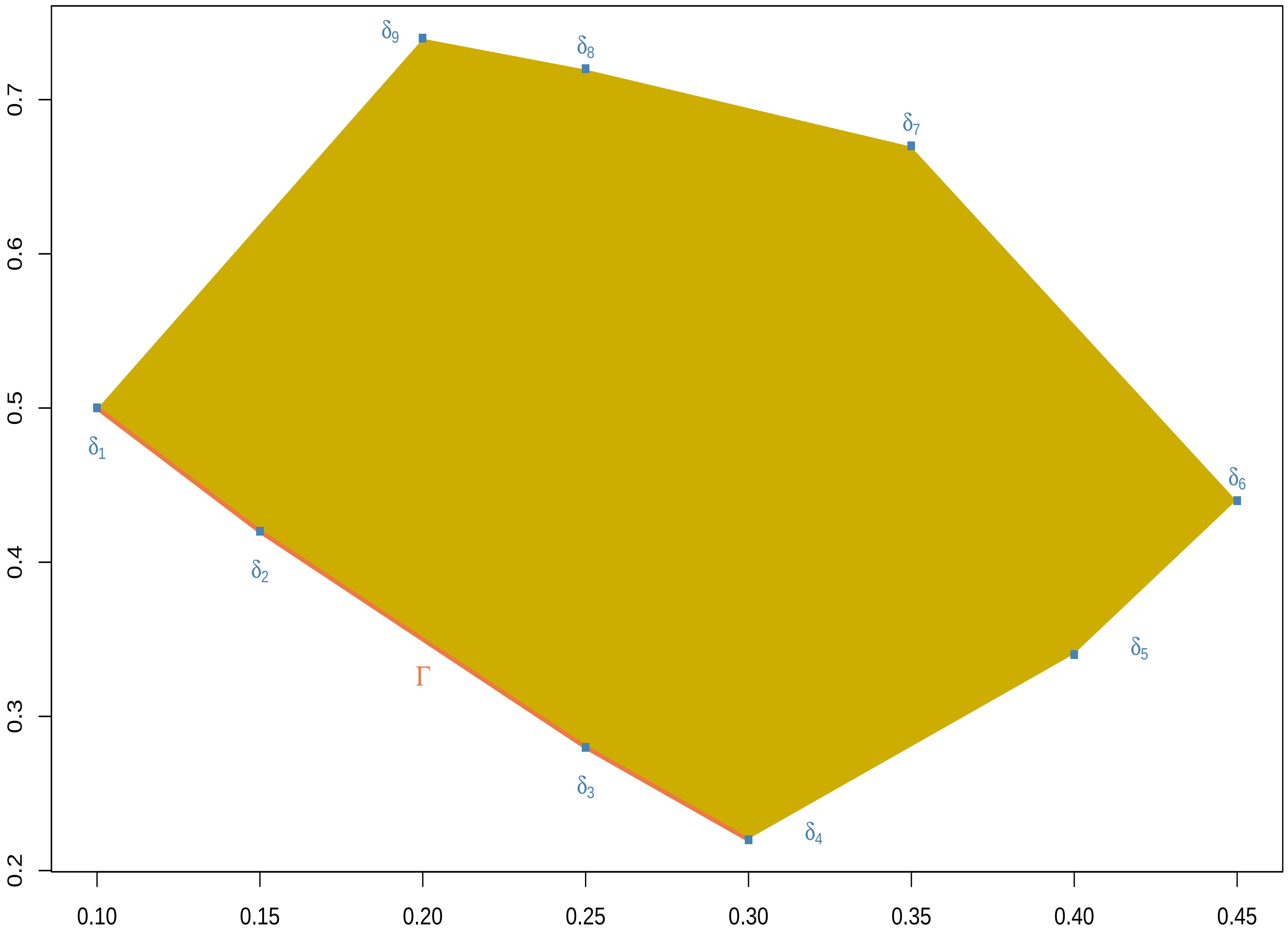
Special case:

$\Theta = \{\theta_1, \theta_2\}$, with risk set

$$\mathcal{R} = \{r = (R(\theta_1, \delta), R(\theta_2, \delta)),\ \delta \in \mathcal{D}^*\},$$

bounded and closed from below

Lower boundary, $\Gamma(\mathcal{R})$, provides the *admissible* points of $\mathcal{R}$.



δ_9
δ_8
δ_7
δ_6
δ_5
δ_4
δ_3
δ_2
δ_1
Γ
0.7
0.6
0.5
0.4
0.3
0.2
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

For every $r \in \Gamma(\mathcal{R})$, there exists a tangent line to $\mathcal{R}$ going through r, with positive slope and equation

$$p_1 r_1 + p_2 r_2 = k$$

Therefore r is a Bayes estimator for $\pi(\theta_i) = p_i \ (i = 1, 2)$

Wald's theorems

If Θ is finite and if $\mathcal{R}$ is bounded and closed from below, then the set of Bayes estimators constitutes a complete class

If Θ is compact and the risk set $\mathcal{R}$ is convex, if all estimators have a continuous risk function, the Bayes estimators constitute a complete class.

Extensions

If Θ not compact, in many cases, complete classes are made of generalized Bayes estimators

Example 25 When estimating the natural parameter θ of an exponential family

$$x \sim f(x|\theta) = e^{\theta \cdot x - \psi(\theta)} h(x), \quad x, \theta \in \mathbb{R}^k,$$

under quadratic loss, every admissible estimator is a generalized Bayes estimator.

5 Bayesian Point Estimation

Posterior distribution

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

- extensive summary of the information available on θ
- integrate simultaneously prior information and information brought by x
- unique motor of inference

5.1 Bayesian inference

5.1.1 MAP estimator

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

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

Motivations

- Associated with $0-1$ losses and L_p losses
- Penalized likelihood estimator
- Further appeal in restricted parameter spaces

Example 26 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:

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

Not always appropriate:

Example 27 Consider

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

and $\pi(\theta) = \frac{1}{2}e^{-|\theta|}$. The MAP estimator of θ is then always

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

5.1.2 Prediction

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

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

Example 28 Consider the AR(1) model

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

the predictive of x_T is then

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

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

5.2 Bayesian Decision Theory

For a loss $\mathrm{L}(\theta, \delta)$ and a prior π, the *Bayes rule* is

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

Note: Practical computation not always possible analytically.

5.2.1 Conjugate priors

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

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

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

Example 29 Consider

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

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

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

and

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

Even conjugate priors may lead to computational difficulties

Example 30 Consider $x \sim \mathcal{N}_p(\theta, I_p)$ and

$$\mathrm{L}(\theta, d) = \frac{(d - ||\theta||^2)^2}{2||\theta||^2 + p}$$

for which $\delta_0(x) = ||x||^2 - p$ has a constant risk, 1

For the conjugate distributions, $\mathcal{N}_p(0, \tau^2 I_p)$,

$$\delta^\pi(x) = \frac{\mathbb{E}^\pi[||\theta||^2/(2||\theta||^2 + p)|x]}{\mathbb{E}^\pi[1/(2||\theta||^2 + p)|x]}$$

cannot be computed analytically.

5.3 The particular case of the normal model

Importance of the normal model in many fields

$$\mathcal{N}_p(\theta, \Sigma)$$

with known Σ, normal conjugate distribution, $\mathcal{N}_p(\mu, A)$.

Under quadratic loss, the Bayes estimator is

$$\begin{aligned}\delta^\pi(x) &= x - \Sigma(\Sigma + A)^{-1}(x - \mu) \\ &= \left(\Sigma^{-1} + A^{-1}\right)^{-1}\left(\Sigma^{-1}x + A^{-1}\mu\right);\end{aligned}$$

5.3.1 Estimation of variance

If

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

the likelihood is

$$\ell(\theta, \sigma \mid \bar{x}, s^2) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\left\{s^2 + n\left(\bar{x} - \theta\right)^2\right\}\right]$$

The *Jeffreys prior* for this model is

$$\pi^*(\theta, \sigma) = \frac{1}{\sigma^2}$$

but invariance arguments lead to prefer

$$\tilde{\pi}(\theta, \sigma) = \frac{1}{\sigma}$$

In this case, the posterior distribution of (θ, σ) is

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

- Conjugate posterior distributions have the same form
- θ and σ^2 are not a priori independent.
- Requires a careful determination of the hyperparameters

5.3.2 Linear models

Usual regression model

$$y = X\beta + \epsilon, \qquad \epsilon \sim \mathcal{N}_k(0, \Sigma),\ \beta \in \mathbb{R}^p$$

Conjugate distributions of the type

$$\beta \sim \mathcal{N}_p(A\theta, C),$$

where $\theta \in \mathbb{R}^q$ ($q \leq p$).

Strong connection with random-effect models

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

Σ unknown

In this general case, the Jeffreys prior is

$$\pi^J(\beta, \Sigma) = \frac{1}{|\Sigma|^{(k+1)/2}}.$$

likelihood

$$\ell(\beta, \Sigma|y) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(y_i - X_i\beta)(y_i - X_i\beta)^t\right]\right\}$$

- suggests (inverse) Wishart distribution on Σ
- posterior marginal distribution on β only defined for sample size large enough
- no closed form expression for posterior marginal

Special case: $\epsilon \sim \mathcal{N}_k(0, \sigma^2 I_k)$

The least-squares estimator $\hat{\beta}$ has a normal distribution

$$\mathcal{N}_p(\beta, \sigma^2 (X^t X)^{-1})$$

Corresponding conjugate distributions on (β, σ^2)

$$\begin{aligned} \beta|\sigma^2 &\sim \mathcal{N}_p\left(\mu, \frac{\sigma^2}{n_0}(X^t X)^{-1}\right), \\ \sigma^2 &\sim \mathcal{IG}(\nu/2, s_0^2/2), \end{aligned}$$

since, if $s^2 = ||y - X\hat{\beta}||^2$,

$$\begin{aligned}
\beta|\hat{\beta}, s^2, \sigma^2 &\sim \mathcal{N}_p\left(\frac{n_0\mu + \hat{\beta}}{n_0 + 1}, \frac{\sigma^2}{n_0 + 1}(X^tX)^{-1}\right), \\
\sigma^2|\hat{\beta}, s^2 &\sim \mathcal{IG}\left(\frac{k - p + \nu}{2}, \frac{s^2 + s_0^2 + \frac{n_0}{n_0+1}(\mu - \hat{\beta})^t X^t X(\mu - \hat{\beta})}{2}\right).
\end{aligned}$$

5.3.3 The AR(p) model

Markovian dynamic model

$$x_t \sim \mathcal{N}\left(\mu - \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu), \sigma^2\right)$$

Appeal:

- Among the most commonly used model in dynamic settings
- More challenging than the static models (stationarity constraints)
- Different models depending on the processing of the starting value x_0

Stationarity

Stationarity constraints in the prior as a restriction on the values of θ.

AR(p) model stationary iff the roots of the polynomial

$$\mathcal{P}(x) = 1 - \sum_{i=1}^{p} \varrho_i x^i$$

are all outside the unit circle

Closed form likelihood

Conditional on the negative time values

$$L(\mu, \varrho_1, \ldots, \varrho_p, \sigma | x_{1:T}, x_{0:(-p+1)}) =$$
$$\sigma^{-T} \prod_{t=1}^{T} \exp \left\{ - \left(x_t - \mu + \sum_{i=1}^{p} \varrho_i (x_{t-i} - \mu) \right)^2 \Big/ 2\sigma^2 \right\} \,,$$

Natural conjugate prior for $\theta = (\mu, \varrho_1, \ldots, \varrho_p, \sigma^2)$:

a normal distribution on $(\mu, \varrho_1, \ldots, \rho_p)$ and an inverse gamma distribution on σ^2.

Stationarity & priors

Under stationarity constraint, complex parameter space

The *Durbin–Levinson recursion* proposes a *reparametrization* from the parameters ϱ_i to the *partial autocorrelations*

$$\psi_i \in [-1, 1]$$

which allow for a uniform prior.

Transform:

0. Define $\varphi^{ii} = \psi_i$ and $\varphi^{ij} = \varphi^{(i-1)j} - \psi_i \varphi^{(i-1)(i-j)}$, for $i > 1$ and $j = 1, \cdots, i-1$.

1. Take $\varrho_i = \varphi^{pi}$ for $i = 1, \cdots, p$.

Different approach via the real+complex roots of the polynomial $\mathcal{P}$, whose inverses are also within the unit circle.

Stationarity & priors (contd.)

Jeffreys' prior associated with the stationary representation is

$$\pi_1^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{1-\varrho^2}} \,.$$

Within the non-stationary region $|\varrho| > 1$, the Jeffreys prior is

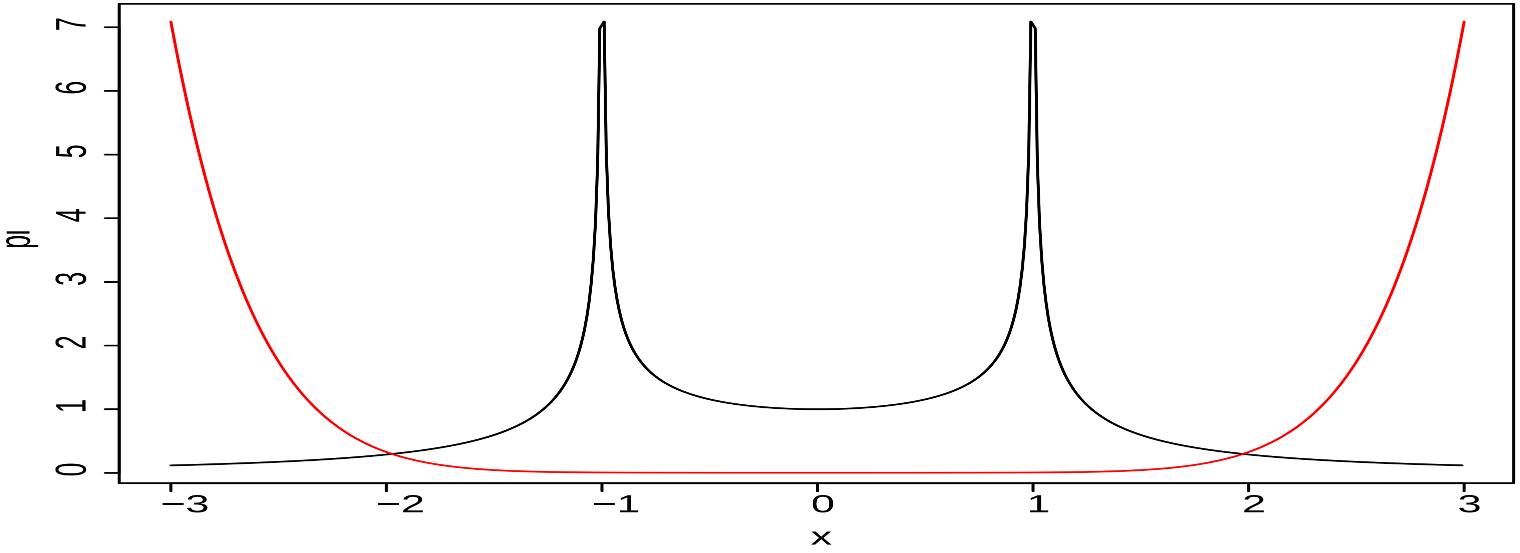
$$\pi_2^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{|1-\varrho^2|}} \sqrt{\left| 1 - \frac{1-\varrho^{2T}}{T(1-\varrho^2)} \right|} \,.$$

The dominant part of the prior is the non-stationary region!

The reference prior π_1^J is only defined when the stationary constraint holds.

Idea Symmetrise to the region $|\varrho| > 1$

$$\pi^B(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \begin{cases} 1/\sqrt{1-\varrho^2} & \text{if } |\varrho| < 1, \\ 1/|\varrho|\sqrt{\varrho^2-1} & \text{if } |\varrho| > 1, \end{cases} ,$$



5.3.4 The MA(q) model

$$x_t = \mu + \epsilon_t - \sum_{j=1}^{q} \vartheta_j \epsilon_{t-j}\,, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Stationary but, for identifiability considerations, the polynomial

$$\mathcal{Q}(x) = 1 - \sum_{j=1}^{q} \vartheta_j x^j$$

must have all its roots outside the unit circle.

Example 31 For the MA(1) model, $x_t = \mu + \epsilon_t - \vartheta_1 \epsilon_{t-1}$,

$$\mathrm{var}(x_t) = (1 + \vartheta_1^2)\sigma^2$$

It can also be written

$$x_t = \mu + \tilde{\epsilon}_{t-1} - \frac{1}{\vartheta_1}\tilde{\epsilon}_t, \quad \tilde{\epsilon} \sim \mathcal{N}(0, \vartheta_1^2\sigma^2)\,,$$

Both couples (ϑ_1, σ) and $(1/\vartheta_1, \vartheta_1\sigma)$ lead to alternative representations of the same model.

Representations

$x_{1:T}$ is a normal random variable with constant mean μ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \gamma_1 & \gamma_2 & \dots & \gamma_q & 0 & \dots & 0 & 0 \\ \gamma_1 & \sigma^2 & \gamma_1 & \dots & \gamma_{q-1} & \gamma_q & \dots & 0 & 0 \\ & & & \ddots & & & & & \\ 0 & 0 & 0 & \dots & 0 & 0 & \dots & \gamma_1 & \sigma^2 \end{pmatrix},$$

with $(|s| \leq q)$

$$\gamma_s = \sigma^2 \sum_{i=0}^{q-|s|} \vartheta_i \vartheta_{i+|s|} \gamma_s = \sigma^2 \sum_{i=0}^{q-|s|} \vartheta_i \vartheta_{i+|s|}$$

Not manageable in practice

Representations (contd.)

Conditional on $(\epsilon_0, \ldots, \epsilon_{-q+1})$,

$$L(\mu, \vartheta_1, \ldots, \vartheta_q, \sigma | x_{1:T}, \epsilon_0, \ldots, \epsilon_{-q+1}) =$$

$$\sigma^{-T} \prod_{t=1}^{T} \exp \left\{ - \left(x_t - \mu + \sum_{j=1}^{q} \vartheta_j \hat{\epsilon}_{t-j} \right)^2 \Big/ 2\sigma^2 \right\} ,$$

where $(t > 0)$

$$\hat{\epsilon}_t = x_t - \mu + \sum_{j=1}^{q} \vartheta_j \hat{\epsilon}_{t-j}, \; \hat{\epsilon}_0 = \epsilon_0, \; \ldots, \; \hat{\epsilon}_{1-q} = \epsilon_{1-q}$$

Recursive definition of the likelihood, still costly $\mathrm{O}(T \times q)$

Representations (contd.)

State-space representation

$$
\begin{aligned}
x_t &= G_y \mathbf{y}_t + \varepsilon_t \, , && (2) \\
\mathbf{y}_{t+1} &= F_t \mathbf{y}_t + \xi_t \, , && (3)
\end{aligned}
$$

(2) is the *observation equation* and (3) is the *state equation*

For the MA(q) model

$$\mathbf{y}_t = (\epsilon_{t-q}, \ldots, \epsilon_{t-1}, \epsilon_t)'$$

and

$$\begin{aligned}
\mathbf{y}_{t+1} &= \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ & & & \ldots & \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \\
x_t &= \mu - \begin{pmatrix} \vartheta_q & \vartheta_{q-1} & \ldots & \vartheta_1 & -1 \end{pmatrix} \mathbf{y}_t \,.
\end{aligned}$$

Example 32 For the MA(1) model, observation equation

$$x_t = (\,1 \quad 0\,)\mathbf{y}_t$$

with

$$\mathbf{y}_t = (\,y_{1t} \quad y_{2t}\,)'$$

directed by the state equation

$$\mathbf{y}_{t+1} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \mathbf{y}_t + \epsilon_{t+1} \begin{pmatrix} 1 \\ \vartheta_1 \end{pmatrix} .$$

Identifiability

Identifiability condition on $\mathcal{Q}(x)$: the ϑ_j's vary in a complex space.

New reparametrization: the ψ_i's are the *inverse partial auto-correlations*

6 Bayesian Calculations

6.1 Implementation difficulties

- Computing the posterior distribution

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

- Resolution of

$$\arg\min \int_\Theta \mathrm{L}(\theta,\delta)\pi(\theta) f(x|\theta) d\theta$$

- Maximisation of the marginal posterior

$$\arg\max \int_{\Theta_{-1}} \pi(\theta|x) d\theta_{-1}$$

- Computing posterior quantities

$$\delta^\pi(x) = \int_\Theta h(\theta)\ \pi(\theta|x) d\theta = \frac{\displaystyle\int_\Theta h(\theta)\ \pi(\theta) f(x|\theta) d\theta}{\displaystyle\int_\Theta \pi(\theta) f(x|\theta) d\theta}$$

- Resolution (in k) of

$$P(\pi(\theta|x) \geq k|x) = \alpha$$

Example 33 Consider

$$x_1, \ldots, x_n \sim \mathcal{C}(\theta, 1)$$

and $\theta \sim \mathcal{N}(\mu, \sigma^2)$, with known hyperparameters μ and σ^2.

The posterior distribution

$$\pi(\theta|x_1, \ldots, x_n) \propto e^{-(\theta-\mu)^2/2\sigma^2} \prod_{i=1}^{n} [1 + (x_i - \theta)^2]^{-1},$$

cannot be integrated analytically and

$$\delta^\pi(x_1, \ldots, x_n) = \frac{\int_{-\infty}^{+\infty} \theta e^{-(\theta-\mu)^2/2\sigma^2} \prod_{i=1}^{n} [1 + (x_i - \theta)^2]^{-1} d\theta}{\int_{-\infty}^{+\infty} e^{-(\theta-\mu)^2/2\sigma^2} \prod_{i=1}^{n} [1 + (x_i - \theta)^2]^{-1} d\theta}$$

requires two numerical integrations.

Example 34 Mixture of two normal distributions

$$x_1, \ldots, x_n \sim f(x|\theta) = p\varphi(x; \mu_1, \sigma_1) + (1-p)\varphi(x; \mu_2, \sigma_2)$$

Prior

$$\mu_i|\sigma_i \sim \mathcal{N}(\xi_i, \sigma_i^2/n_i), \qquad \sigma_i^2 \sim \mathcal{IG}(\nu_i/2, s_i^2/2), \qquad p \sim \mathcal{B}e(\alpha, \beta)$$

Posterior

$$\begin{aligned}\pi(\theta|x_1, \ldots, x_n) &\propto \prod_{j=1}^{n} \left\{p\varphi(x_j; \mu_1, \sigma_1) + (1-p)\varphi(x_j; \mu_2, \sigma_2)\right\} \pi(\theta) \\ &= \sum_{\ell=0}^{n} \sum_{(k_t)} \omega(k_t)\pi(\theta|(k_t))\end{aligned}$$

[O(2^n)]

For a given permutation (k_t), conditional posterior distribution

$$\begin{aligned}\pi(\theta|(k_t)) = \mathcal{N}&\left(\xi_1(k_t), \frac{\sigma_1^2}{n_1+\ell}\right) \times \mathcal{IG}((\nu_1+\ell)/2, s_1(k_t)/2)\\ &\times\mathcal{N}\left(\xi_2(k_t), \frac{\sigma_2^2}{n_2+n-\ell}\right) \times \mathcal{IG}((\nu_2+n-\ell)/2, s_2(k_t)/2)\\ &\times\mathcal{B}e(\alpha+\ell, \beta+n-\ell)\end{aligned}$$

where

$$\begin{array}{rclrcl}\bar{x}_1(k_t) & = & \frac{1}{\ell}\sum_{t=1}^{\ell} x_{k_t}, & \hat{s}_1(k_t) & = & \sum_{t=1}^{\ell}(x_{k_t} - \bar{x}_1(k_t))^2, \\ \bar{x}_2(k_t) & = & \frac{1}{n-\ell}\sum_{t=\ell+1}^{n} x_{k_t}, & \hat{s}_2(k_t) & = & \sum_{t=\ell+1}^{n}(x_{k_t} - \bar{x}_2(k_t))^2\end{array}$$

and

$$\begin{aligned}\xi_1(k_t) &= \frac{n_1\xi_1 + \ell\bar{x}_1(k_t)}{n_1+\ell}, \qquad \xi_2(k_t) = \frac{n_2\xi_2 + (n-\ell)\bar{x}_2(k_t)}{n_2+n-\ell}, \\ s_1(k_t) &= s_1^2 + \hat{s}_1^2(k_t) + \frac{n_1\ell}{n_1+\ell}(\xi_1 - \bar{x}_1(k_t))^2, \\ s_2(k_t) &= s_2^2 + \hat{s}_2^2(k_t) + \frac{n_2(n-\ell)}{n_2+n-\ell}(\xi_2 - \bar{x}_2(k_t))^2,\end{aligned}$$

posterior updates of the hyperparameters

Bayes estimator of θ:

$$\delta^\pi(x_1,\ldots,x_n) = \sum_{\ell=0}^{n} \sum_{(k_t)} \omega(k_t) \mathbb{E}^\pi[\theta|\mathbf{x},(k_t)]$$

Too costly: 2^n terms

6.2 Classical approximation methods

6.2.1 Numerical integration

- Simpson's method
- polynomial quadrature

$$\int_{-\infty}^{+\infty} e^{-t^2/2} f(t)\, dt \approx \sum_{i=1}^{n} \omega_i f(t_i),$$

where

$$\omega_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(t_i)]^2}$$

and t_i is the ith zero of the nth *Hermite polynomial*, $H_n(t)$.

- orthogonal bases
- wavelets

[Bumps into curse of dimen'ty]

6.2.2 Monte Carlo methods

Approximation of the integral

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

should take advantage of the fact that $f(x|\theta)\pi(\theta)$ is proportional to a density.

If the θ_i's are generated from $\pi(\theta)$, the average

$$\frac{1}{m}\sum_{i=1}^{m} g(\theta_i) f(x|\theta_i)$$

converges (almost surely) to $\mathfrak{I}$

Confidence regions can be derived from a normal approximation and the magnitude of the error remains of order

$$1/\sqrt{m}\,,$$

whatever the dimension of the problem.

Importance function

No need to simulate from $\pi(\cdot|x)$ or π: if h is a probability density,

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

[Importance function]

An approximation to $\mathbb{E}^\pi[g(\theta)|x]$ is given by

$$\frac{\sum_{i=1}^m g(\theta_i)\omega(\theta_i)}{\sum_{i=1}^m \omega(\theta_i)} \quad \text{with} \quad \omega(\theta_i) = \frac{f(x|\theta_i)\pi(\theta_i)}{h(\theta_i)}$$

if

$$supp(h) \subset supp(f(x|\cdot)\pi)$$

Requirements

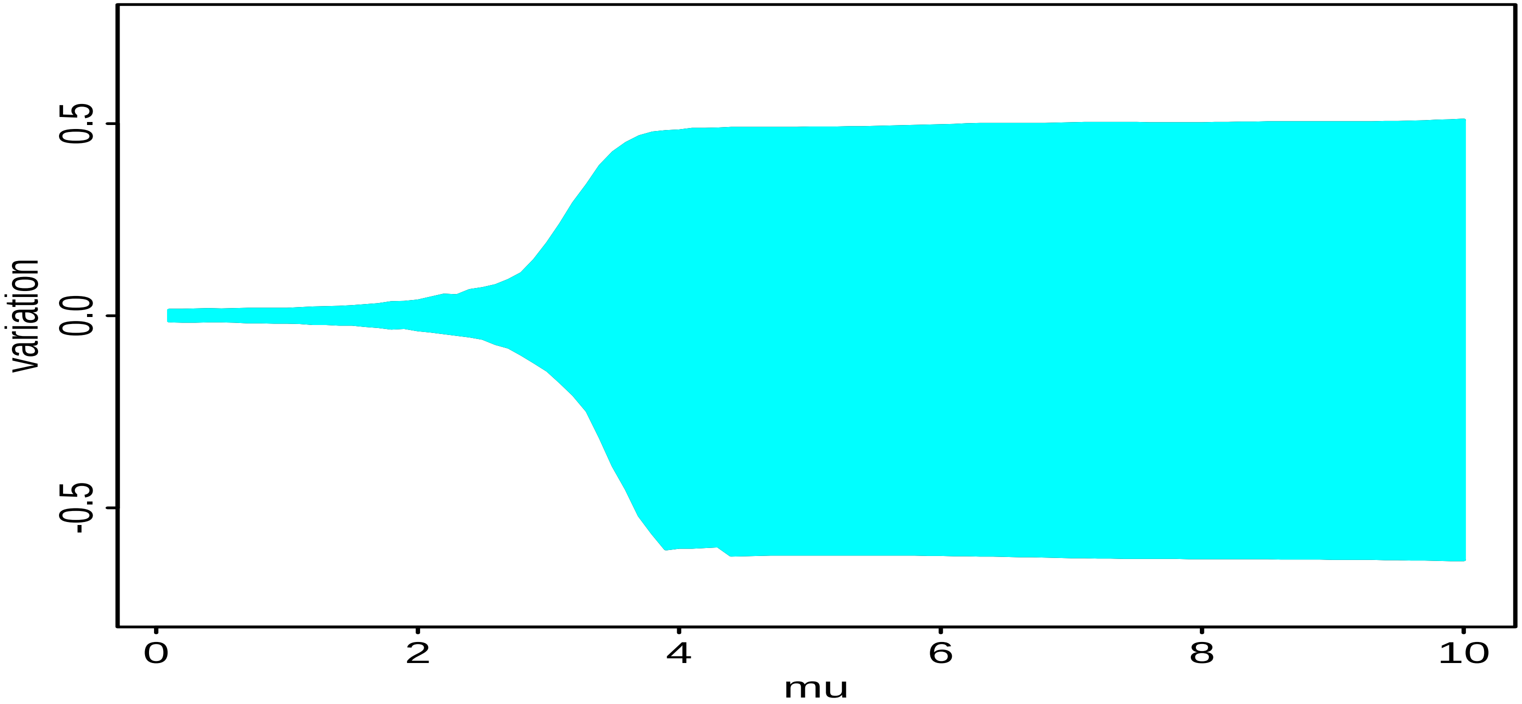
- Simulation from h must be easy
- $h(\theta)$ must be close enough to $g(\theta)\pi(\theta|x)$
- the variance of the importance sampling estimator must be finite

The importance function may be π

Example 35 (Example 33 continued) Since $\pi(\theta)$ is the normal distribution $\mathcal{N}(\mu, \sigma^2)$, it is possible to simulate a normal sample $\theta_1, \ldots, \theta_M$ and to approximate the Bayes estimator by

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

May be poor when the x_i's are all far from μ



90% range of variation of the approximation for $n = 10$ observations from $\mathcal{C}(0,1)$ distribution and $M = 1000$ simulations of θ from a $\mathcal{N}(\mu, 1)$ distribution.

Defensive sampling:

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

[Newton & Raftery, 1994]

Case of the Bayes factor

Models $\mathcal{M}_1$ vs. $\mathcal{M}_2$ compared via

$$\begin{aligned} B_{12} &= \frac{Pr(\mathcal{M}_1|x)}{Pr(\mathcal{M}_2|x)} \bigg/ \frac{Pr(\mathcal{M}_1)}{Pr(\mathcal{M}_2)} \\ &= \frac{\displaystyle\int f_1(x|\theta_1)\pi_1(\theta_1)d\theta_1}{\displaystyle\int f_2(x|\theta_2)\pi_2(\theta_2)d\theta_2} \end{aligned}$$

[Good, 1958 & Jeffreys, 1961]

Solutions

- **Bridge sampling:**

  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

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

[Chen, Shao & Ibrahim, 2000]

- **Umbrella sampling:**

$$\begin{aligned} \pi_1(\theta) &= \pi(\theta|\lambda_1) & \qquad \pi_2(\theta) &= \pi_1(\theta|\lambda_2) \\ &= \tilde{\pi}_1(\theta)/c(\lambda_1) & &= \tilde{\pi}_2(\theta)/c(\lambda_2) \end{aligned}$$

Then

$$\forall\ \pi(\lambda) \text{ on } [\lambda_1, \lambda_2], \qquad \log(c(\lambda_2)/c(\lambda_1)) = \mathbb{E}\left[\frac{\frac{d}{d\lambda}\log\tilde{\pi}(d\theta)}{\pi(\lambda)}\right]$$

and

$$\log(B_{12}) \approx \frac{1}{n}\sum_{i=1}^{n}\frac{\frac{d}{d\lambda}\log\tilde{\pi}(\theta_i|\lambda_i)}{\pi(\lambda_i)}$$

6.3 Markov chain Monte Carlo methods

Idea Given a density distribution $\pi(\cdot|x)$, produce a Markov chain $(\theta^{(t)})_t$ with stationary distribution $\pi(\cdot|x)$

Warranty:

if the Markov chains produced by MCMC algorithms are irreducible, then these chains are positive recurrent with stationary distribution $\pi(\theta|x)$ and ergodic.

Translation:

For k large enough, $\theta^{(k)}$ is approximately distributed from $\pi(\theta|x)$, no matter what the starting value $\theta^{(0)}$ is.

Practical use

- Produce an i.i.d. sample $\theta_1, \ldots, \theta_m$ from $\pi(\theta|x)$, taking the current $\theta^{(k)}$ as the new starting value
- Approximate $\mathbb{E}^\pi[g(\theta)|x]$ as

$$\frac{1}{K}\sum_{k=1}^{K} g(\theta^{(k)})$$

[Ergodic Theorem]

- Achieve quasi-independence by batch sampling
- Construct approximate posterior confidence regions

$$C_x^\pi \simeq [\theta^{(\alpha T/2)}, \theta^{(T-\alpha T/2)}]$$

6.3.1 Metropolis–Hastings algorithms

Based on a conditional density $q(\theta'|\theta)$

(i). Start with an arbitrary initial value $\theta^{(0)}$

(ii). Update from $\theta^{(m)}$ to $\theta^{(m+1)}$ $(m = 1, 2, \ldots)$ by

(a) Generate $\xi \sim q(\xi|\theta^{(m)})$

(b) Define

$$\varrho = \frac{\pi(\xi)\, q(\theta^{(m)}|\xi)}{\pi(\theta^{(m)})\, q(\xi|\theta^{(m)})} \wedge 1$$

(c) Take

$$\theta^{(m+1)} = \begin{cases} \xi & \text{with probability } \varrho, \\ \theta^{(m)} & \text{otherwise.} \end{cases}$$

Validation

Detailed balance condition

$$\pi(\theta)K(\theta'|\theta) = \pi(\theta')K(\theta|\theta')$$

with $K(\theta'|\theta)$ transition kernel

$$K(\theta'|\theta) = \varrho(\theta,\theta')q(\theta'|\theta) + \int [1-\varrho(\theta,\xi)]q(\xi|\theta)d\xi\,\delta_\theta(\theta')\,,$$

where δ Dirac mass

Random walk Metropolis–Hastings

$$q(\theta'|\theta) = f(||\theta' - \theta||)$$

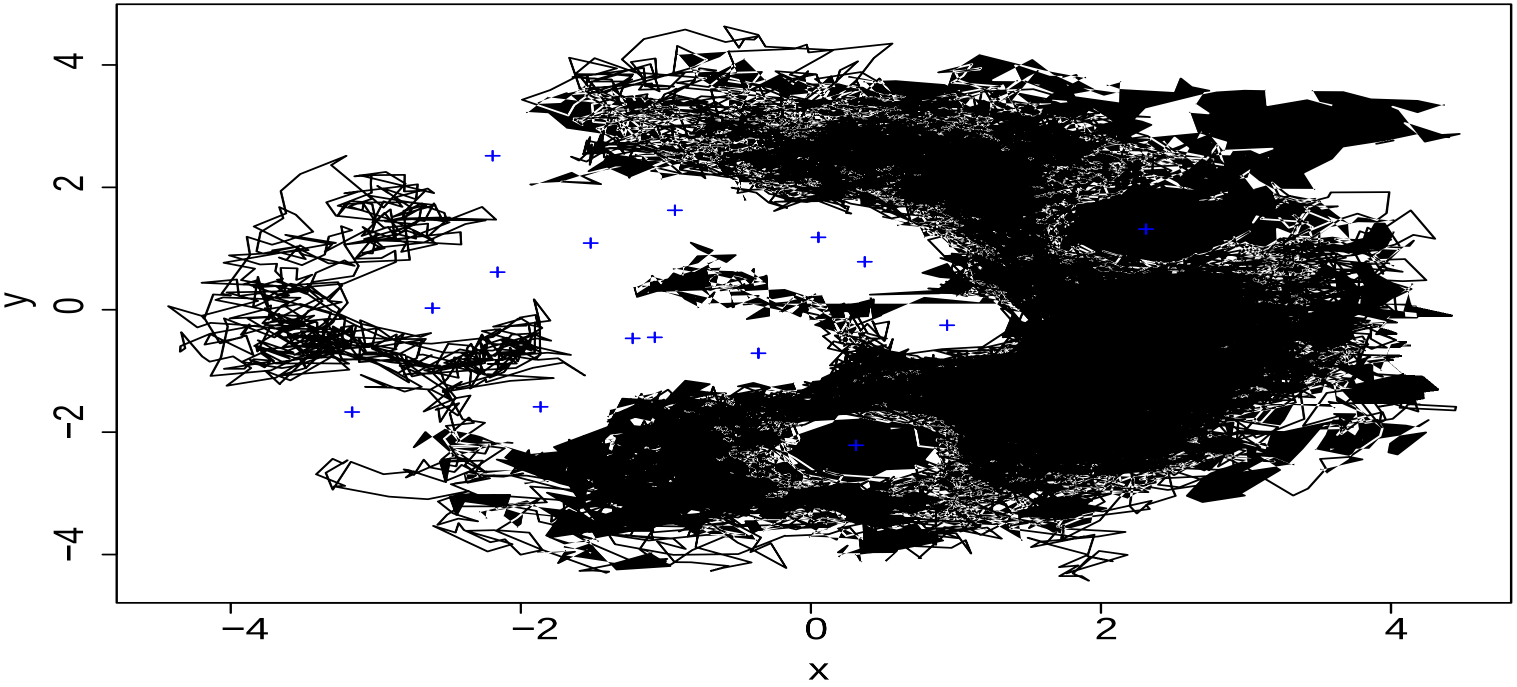
Corresponding Metropolis–Hastings acceptance ratio

$$\varrho = \frac{\pi(\xi)}{\pi(\theta^{(m)})} \wedge 1.$$

Example 36 For $\theta, x \in \mathbb{R}^2$,

$$\pi(\theta|x) \propto \exp\{-||\theta - x||^2/2\} \prod_{i=1}^{p} \exp\left\{\frac{-1}{||\theta - \mu_i||^2}\right\} ,$$

where the μ_i's are given repulsive points



Path of the Markov chain for repulsive points μ_j indicated by crosses, $x = 0$ and $p = 15$ $(5000$ iterations).

Pros & Cons

- Widely applicable
- limited tune-up requirements (scale calibrated thru acceptance)
- never uniformely ergodic

Independent proposals

Take

$$q(\theta'|\theta) = h(\theta')\,.$$

More limited applicability and closer connection with iid simulation

Examples

- prior distribution
- likelihood
- saddlepoint approximation

6.3.2 The Gibbs sampler

Takes advantage of *hierarchical structures*: if

$$\pi(\theta|x) = \int \pi_1(\theta|x,\lambda)\pi_2(\lambda|x)\,d\lambda\,,$$

simulate from the joint distribution

$$\pi_1(\theta|x,\lambda)\,\pi_2(\lambda|x)$$

Example 37 Consider $(\theta, \lambda) \in \mathbb{N} \times [0, 1]$ and

$$\pi(\theta, \lambda|x) \propto \binom{n}{\theta} \lambda^{\theta+\alpha-1}(1-\lambda)^{n-\theta+\beta-1}$$

Hierarchical structure:

$$\theta|x, \lambda \sim \mathcal{B}(n, \lambda), \qquad \lambda|x \sim \mathcal{B}e(\alpha, \beta)$$

Then

$$\pi(\theta|x) = \binom{n}{\theta} \frac{B(\alpha+\theta, \beta+n-\theta)}{B(\alpha, \beta)}$$

[beta-binomial distribution]

Difficult to work with this marginal. For instance, computation of $\mathbb{E}[\theta/(\theta+1)|x]$?

More advantageous to simulate

$$\lambda^{(i)} \sim \mathcal{B}e(\alpha, \beta) \text{ and } \theta^{(i)} \sim \mathcal{B}(n, \lambda^{(i)})$$

Then approximate $\mathbb{E}[\theta/(\theta+1)|x]$ as

$$\frac{1}{m}\sum_{i=1}^{m} \frac{\theta^{(i)}}{\theta^{(i)}+1}$$

Conditionals

Usually $\pi_2(\lambda|x)$ not available/simulable

More often, both *conditional posterior distributions*,

$$\pi_1(\theta|x,\lambda) \text{ and } \pi_2(\lambda|x,\theta)$$

can be simulated.

Data augmentation

Initialization: Start with an arbitrary value $\lambda^{(0)}$

Iteration t: Given $\lambda^{(t-1)}$, generate

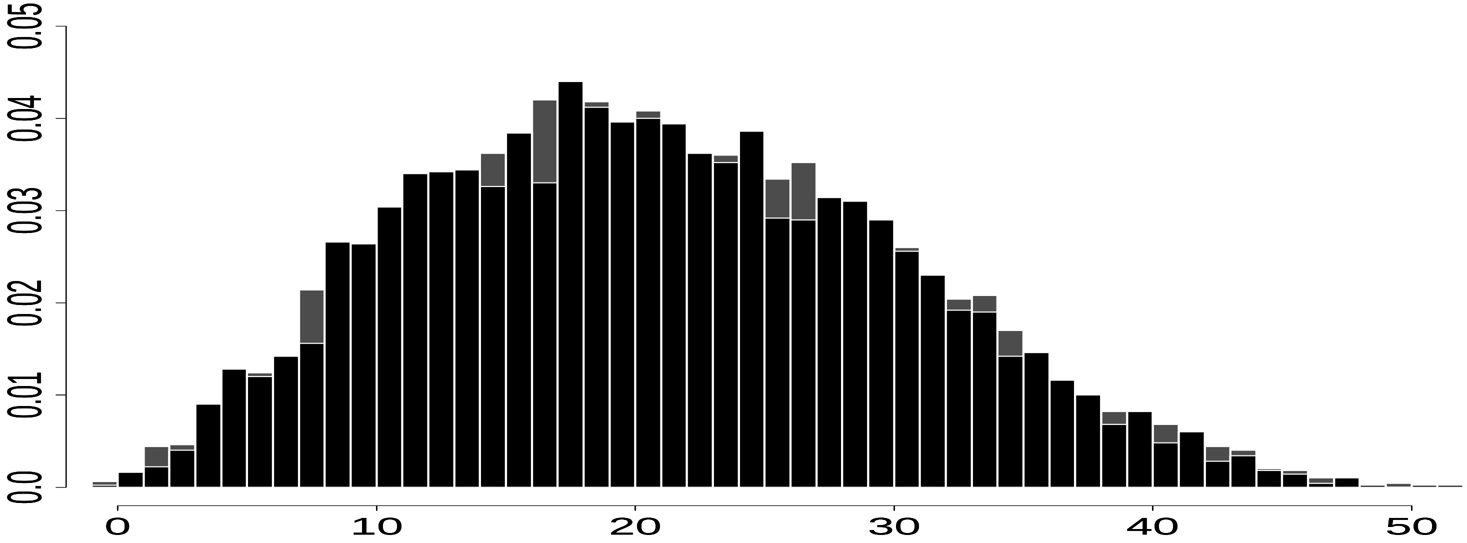
a. $\theta^{(t)}$ according to $\pi_1(\theta|x, \lambda^{(t-1)})$

b. $\lambda^{(t)}$ according to $\pi_2(\lambda|x, \theta^{(t)})$

$\pi(\theta, \lambda|x)$ **is a stationary distribution for this transition**

Example 38 (Example 37 continued) The conditional distributions are

$$\theta|x,\lambda \sim \mathcal{B}(n,\lambda), \qquad \lambda|x,\theta \sim \mathcal{B}e(\alpha+\theta,\beta+n-\theta)$$



Histograms for samples of size 5000 from the beta-binomial with $n = 54$**,** $\alpha = 3.4$**, and** $\beta = 5.2$

Rao–Blackwellization

Conditional structure of the sampling algorithm and the dual sample,

$$\lambda^{(1)}, \ldots, \lambda^{(m)},$$

should be exploited.

$\mathbb{E}^{\pi}[g(\theta)|x]$ approximated as

$$\delta_2 = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}^{\pi}[g(\theta)|x, \lambda^{(m)}],$$

instead of

$$\delta_1 = \frac{1}{m} \sum_{i=1}^{m} g(\theta^{(i)}).$$

Approximation of $\pi(\theta|x)$ by

$$\frac{1}{m} \sum_{i=1}^{m} \pi(\theta|x, \lambda_i)$$

The general Gibbs sampler

Consider several groups of parameters, $\theta, \lambda_1, \ldots, \lambda_p$, such that

$$\pi(\theta|x) = \int \ldots \int \pi(\theta, \lambda_1, \ldots, \lambda_p|x)\, d\lambda_1 \cdots d\lambda_p$$

or simply divide θ in

$$(\theta_1, \ldots, \theta_p)$$

Example 39 Consider a multinomial model,

$$y \sim \mathcal{M}_5\left(n; a_1\mu + b_1, a_2\mu + b_2, a_3\eta + b_3, a_4\eta + b_4, c(1-\mu-\eta)\right),$$

parametrized by μ and η, where

$$0 \leq a_1 + a_2 = a_3 + a_4 = 1 - \sum_{i=1}^{4} b_i = c \leq 1$$

and $c, a_i, b_i \geq 0$ are known.

This model stems from sampling according to

$$x \sim \mathcal{M}_9(n; a_1\mu, b_1, a_2\mu, b_2, a_3\eta, b_3, a_4\eta, b_4, c(1-\mu-\eta)),$$

and aggregating some coordinates:

$$y_1 = x_1 + x_2, \quad y_2 = x_3 + x_4, \quad y_3 = x_5 + x_6, \quad y_4 = x_7 + x_8,\ y_5 = x_9.$$

For the prior

$$\pi(\mu,\eta) \propto \mu^{\alpha_1-1}\eta^{\alpha_2-1}(1-\eta-\mu)^{\alpha_3-1},$$

the posterior distribution of (μ,η) cannot be derived explicitly.

Introduce $z = (x_1, x_3, x_5, x_7)$, which is not observed and

$$\begin{aligned}\pi(\eta, \mu|y, z) &= \pi(\eta, \mu|x) \\ &\propto \mu^{z_1}\mu^{z_2}\eta^{z_3}\eta^{z_4}(1-\eta-\mu)^{y_5+\alpha_3-1}\mu^{\alpha_1-1}\eta^{\alpha_2-1}\,,\end{aligned}$$

where we denote the coordinates of z as (z_1, z_2, z_3, z_4). Therefore,

$$\mu, \eta|y, z \sim \mathcal{D}(z_1 + z_2 + \alpha_1, z_3 + z_4 + \alpha_2, y_5 + \alpha_3).$$

Moreover,

$$\begin{aligned} z_i|y,\mu,\eta &\sim \mathcal{B}\left(y_i,\frac{a_i\mu}{a_i\mu+b_i}\right) \qquad (i=1,2),\\ z_i|y,\mu,\eta &\sim \mathcal{B}\left(y_i,\frac{a_i\eta}{a_i\eta+b_i}\right) \qquad (i=3,4). \end{aligned}$$

The Gibbs sampler

For a joint distribution $\pi(\theta)$ with full conditionals $\pi_1, \ldots, \pi_p$,

Given $(\theta_1^{(t)}, \ldots, \theta_p^{(t)})$, simulate

1. $\theta_1^{(t+1)} \sim \pi_1(\theta_1|\theta_2^{(t)}, \ldots, \theta_p^{(t)})$,
2. $\theta_2^{(t+1)} \sim \pi_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)})$,

$$\vdots$$

p. $\theta_p^{(t+1)} \sim \pi_p(\theta_p|\theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)})$.

6.3.3 The slice sampler

Generality of the Gibbs sampler

For

$$\pi(\theta) = \prod_{i=1}^{k} \varpi_i(\theta),$$

defined on Θ,

$$\pi(\theta) = \int \prod_{i=1}^{k} \mathbb{I}_{0 \le \omega_i \le \varpi_i(\theta)} \, d\,\omega_1 \cdots d\,\omega_k \,.$$

Corresponding slice sampler

At iteration t, simulate

1. $\omega_1^{(t+1)} \sim \mathcal{U}_{[0,\varpi_1(\theta^{(t)})]}$

   $\vdots$

k. $\omega_k^{(t+1)} \sim \mathcal{U}_{[0,\varpi_k(\theta^{(t)})]}$

k+1. $\theta^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$, with

$$A^{(t+1)} = \{\xi;\ \varpi_i(\xi) \geq \omega_i^{(t+1)},\ i = 1, \ldots, k\}.$$

Example 40 Consider

$$\pi(\alpha,\eta|x_1,\ldots,x_n) \propto \alpha^n \eta^{n+\beta-1} \left(\prod_{i=1}^n x_i\right)^\alpha \exp\left\{-\eta\sum_{i=1}^n x_i^\alpha - \alpha - \xi\eta\right\}$$

The conditional distribution $\pi_1(\eta|\alpha,x_1,\ldots,x_n)$ is

$$\mathcal{G}(\beta+n,\xi+\sum_i x_i^\alpha)$$

But $\pi_2(\alpha|\eta,x_1,\ldots,x_n)$ is much more complex

$\pi_2(\alpha|\eta, x_1, \dots, x_n)$ can be expressed as the marginal (in α) of

$$\alpha^n \mathbb{I}_{0\le\omega_0\le\chi^\alpha} \prod_{i=1}^n \mathbb{I}_{0\le\omega_i\le\exp(-\eta x_i^\alpha)} \cdot$$

Then

$$\alpha|\eta, \omega \sim \alpha^n \mathbb{I}_{\alpha\log(\chi)\le\log(\omega_0)} \prod_{i=1}^n \mathbb{I}_{\alpha\log(x_i)\le\log\{-\log(\omega_i)/\eta\}}$$

Example 41 **Mixtures of two normal distributions**

$$\pi(\theta|x) \propto \tilde{\pi}(\theta|x) = \pi(\theta) \prod_{i=1}^{n} \{p\varphi(x_i;\mu_1,\sigma_1) + (1-p)\varphi(x_i;\mu_2,\sigma_2)\} \ ,$$

Simulating from $\theta \sim \mathcal{U}_{\tilde{\pi}(\theta|x)\geq\omega}$. is impossible

Instead, introduce n auxiliary variables ω_i so that

$$\pi(\theta|x_1,\ldots,x_n) \propto \pi(\theta) \prod_{i=1}^{n} \int \mathbb{I}_{p\varphi(x_i;\mu_1,\sigma_1)+(1-p)\varphi(x_i;\mu_2,\sigma_2)\geq\omega_i\geq 0} d\omega_i$$

6.3.4 The impact on Bayesian Statistics

- Radical modification of the way people work with models and prior assumptions
- Allows for much more complex structures:
  - use of graphical models
  - exploration of latent variable models
- Removes the need for analytical processing
- Boosted hierarchical modeling
- Enables *(truly)* Bayesian model choice

6.4 An application to mixture estimation

Use of the missing data representation

$$\begin{aligned} z_j|\theta &\sim \mathcal{M}_p(1; p_1, \ldots, p_k)\,, \\ x_j|z_j, \theta &\sim \mathcal{N}\left(\prod_{i=1}^k \mu_i^{z_{ij}}, \prod_{i=1}^k \sigma_i^{2z_{ij}}\right)\,. \end{aligned}$$

Corresponding conditionals (Gibbs)

$$z_j|x_j,\theta \sim \mathcal{M}_k(1; p_1(x_j,\theta),\ldots,p_k(x_j,\theta)),$$

with $(1 \le i \le k)$

$$p_i(x_j,\theta) = \frac{p_i\varphi(x_j;\mu_i,\sigma_i)}{\sum_{t=1}^k p_t\varphi(x_j;\mu_t,\sigma_t)}$$

and

$$\mu_i|\mathbf{x},\mathbf{z},\sigma_i \sim \mathcal{N}(\xi_i(\mathbf{x},\mathbf{z}),\sigma_i^2/(n+\sigma_i^2)),$$

$$\sigma_i^{-2}|\mathbf{x},\mathbf{z} \sim \mathcal{G}\left(\frac{\nu_i+n_i}{2},\frac{1}{2}\left[s_i^2+\hat{s}_i^2(\mathbf{x},\mathbf{z})+\frac{n_i m_i(\mathbf{z})}{n_i+m_i(\mathbf{z})}(\bar{x}_i(\mathbf{z})-\xi_i)^2\right]\right),$$

$$p|\mathbf{x},\mathbf{z} \sim \mathcal{D}_k(\alpha_1+m_1(\mathbf{z}),\ldots,\alpha_k+m_k(\mathbf{z})),$$

where

$$m_i(\mathbf{z}) = \sum_{j=1}^{n} z_{ij}, \qquad \bar{x}_i(j) = \frac{1}{m_i(\mathbf{z})} \sum_{j=1}^{n} z_{ij} x_j,$$

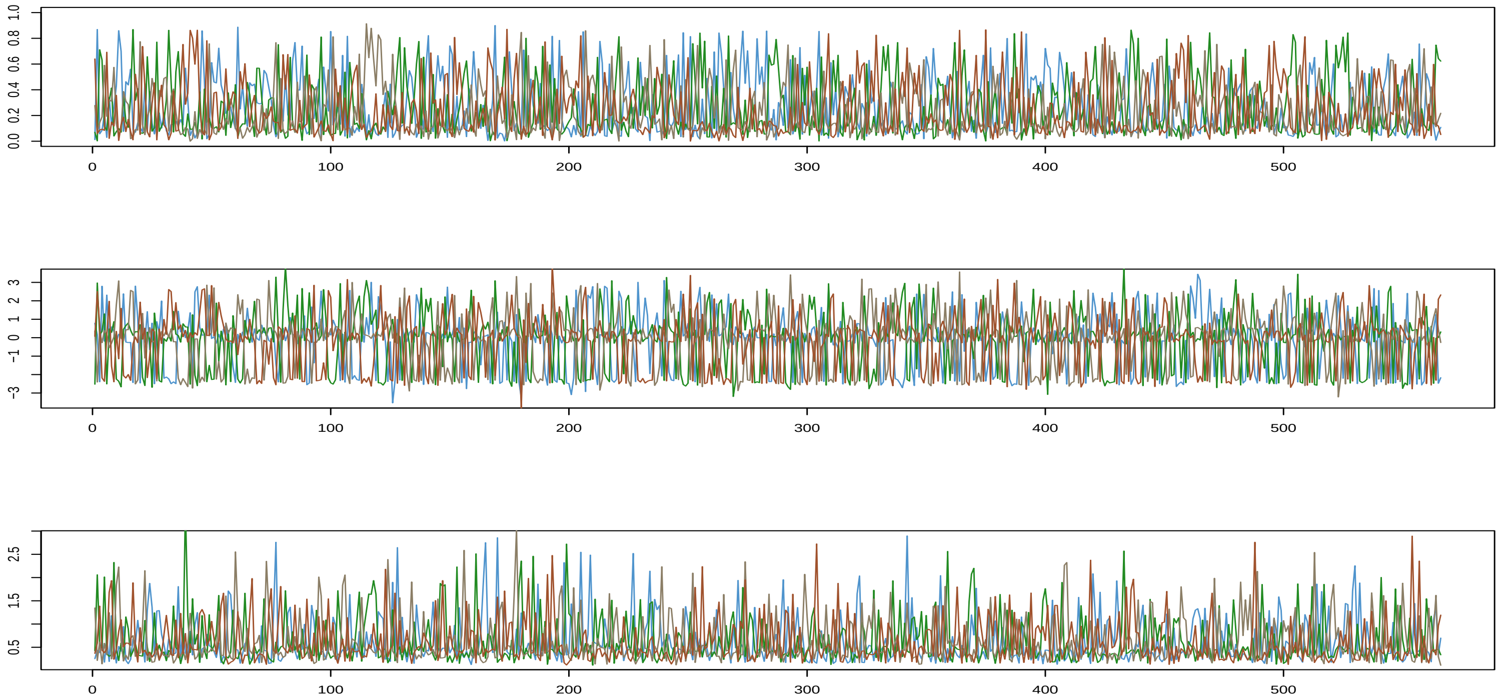
and

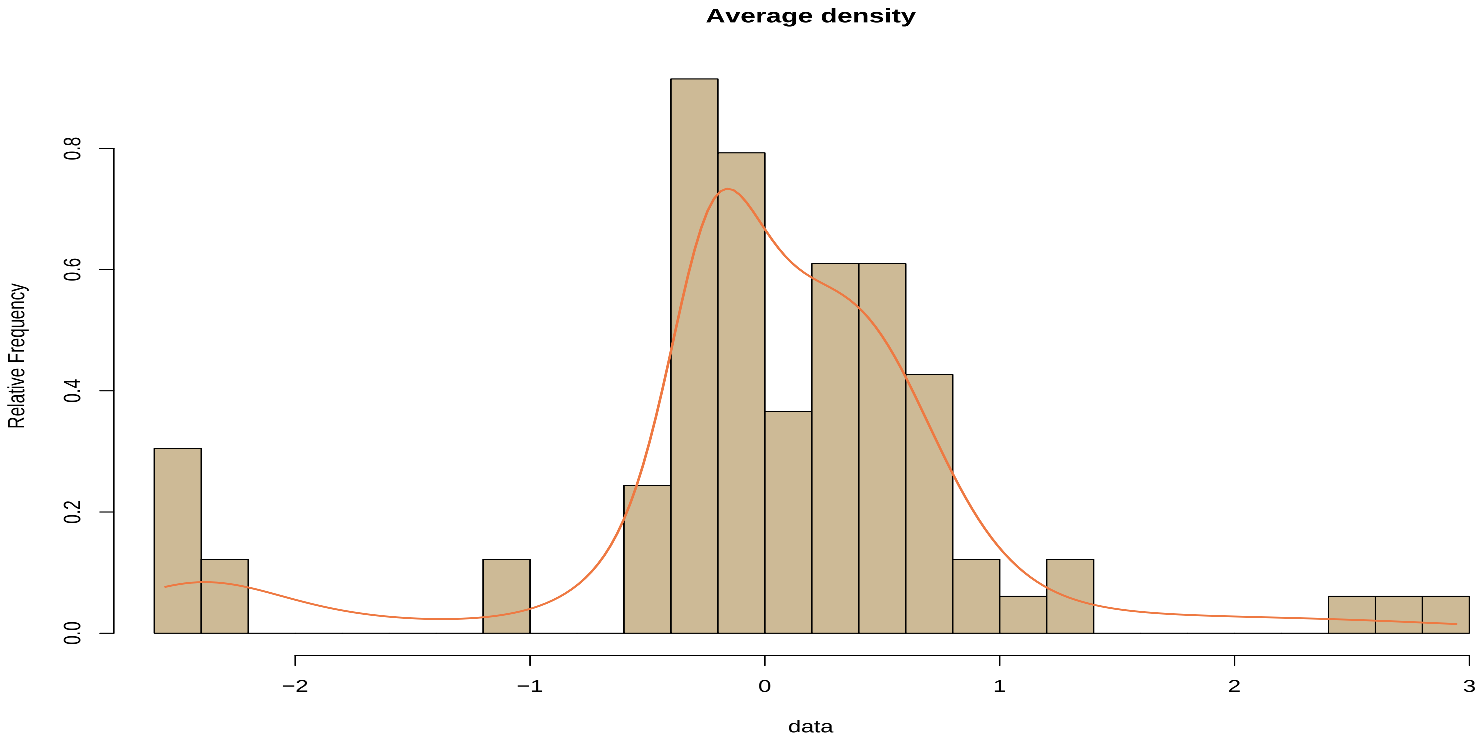
$$\xi_i(\mathbf{x}, \mathbf{z}) = \frac{n_i \xi_i + m_i(\mathbf{z}) \bar{x}_i(\mathbf{z})}{n_i + m_i(\mathbf{z})}, \qquad \hat{s}_i^2(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{n} z_{ij} (x_j - \bar{x}_i(\mathbf{z}))^2.$$

Properties

- Slow moves sometimes
- Large increase in dimension, order O(n)
- Good theoretical properties **(Duality principle)**

Example — Galaxy benchmark $(k = 4)$




Average density
Relative Frequency
0.0
0.2
0.4
0.6
0.8
−2
−1
0
1
2
3
data

Random walk Metropolis–Hastings

$$
\begin{aligned}
q(\theta_t^*|\theta_{t-1}) &= \Psi(\theta_t^* - \theta_{t-1}) \\
\rho &= \frac{\pi(\theta_t^*|x_1,\ldots,x_n)}{\pi(\theta_{t-1}|x_1,\ldots,x_n)} \wedge 1
\end{aligned}
$$

Properties

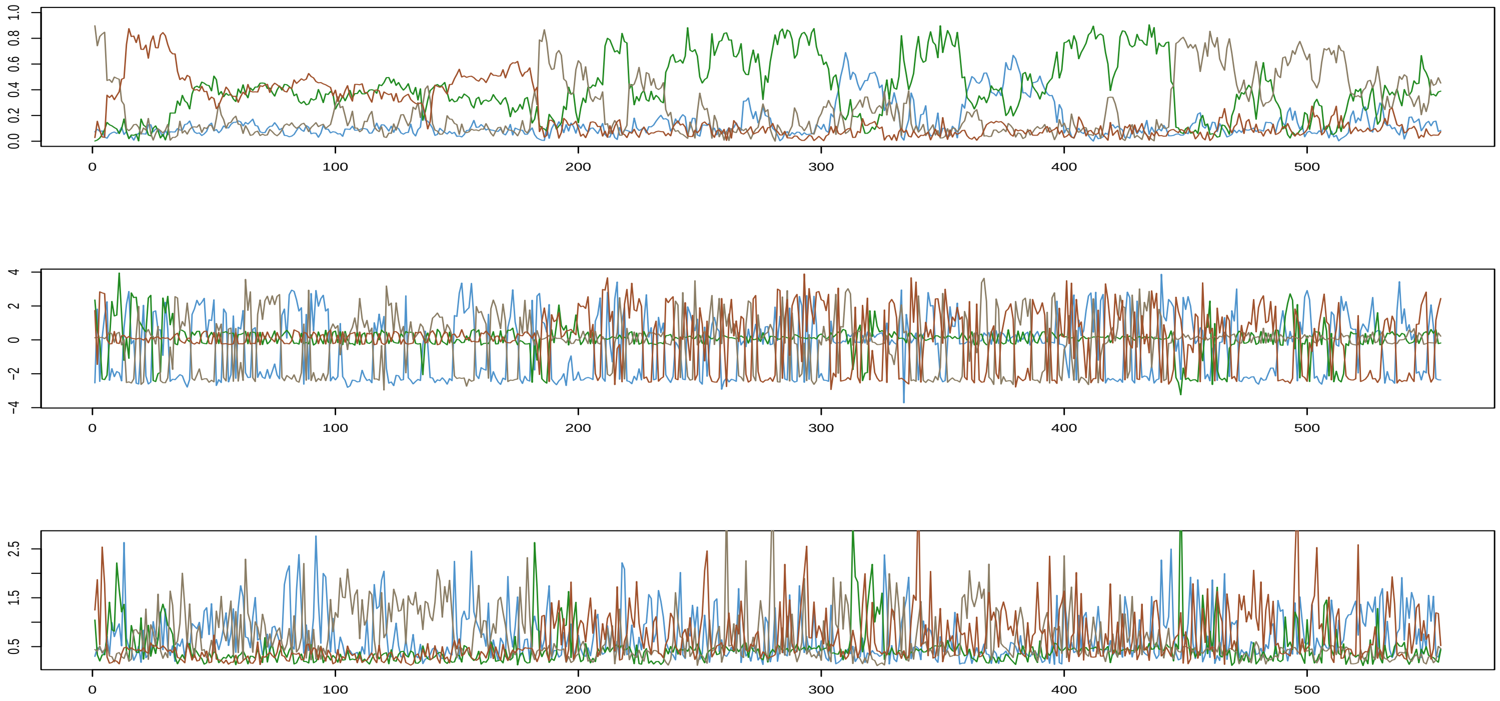
- Avoids completion
- Available (Normal vs. Cauchy vs... moves)
- Calibrated against acceptance rate
- Depends on parameterisation

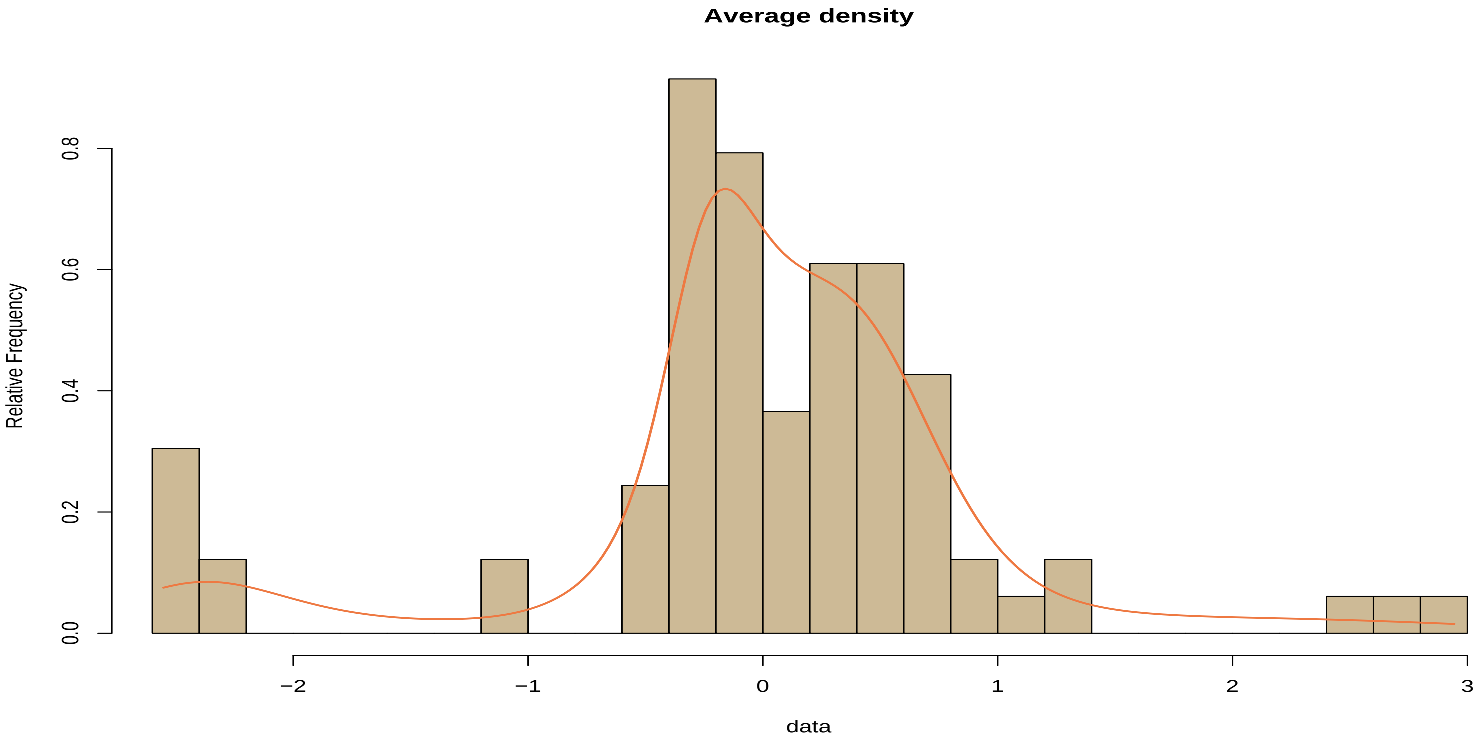
$$\lambda_j \longrightarrow \log \lambda_j \qquad\qquad p_j \longrightarrow \log(p_j/1 - p_k)$$

or

$$\theta_i \longrightarrow \frac{\exp \theta_i}{1 + \exp \theta_i}$$

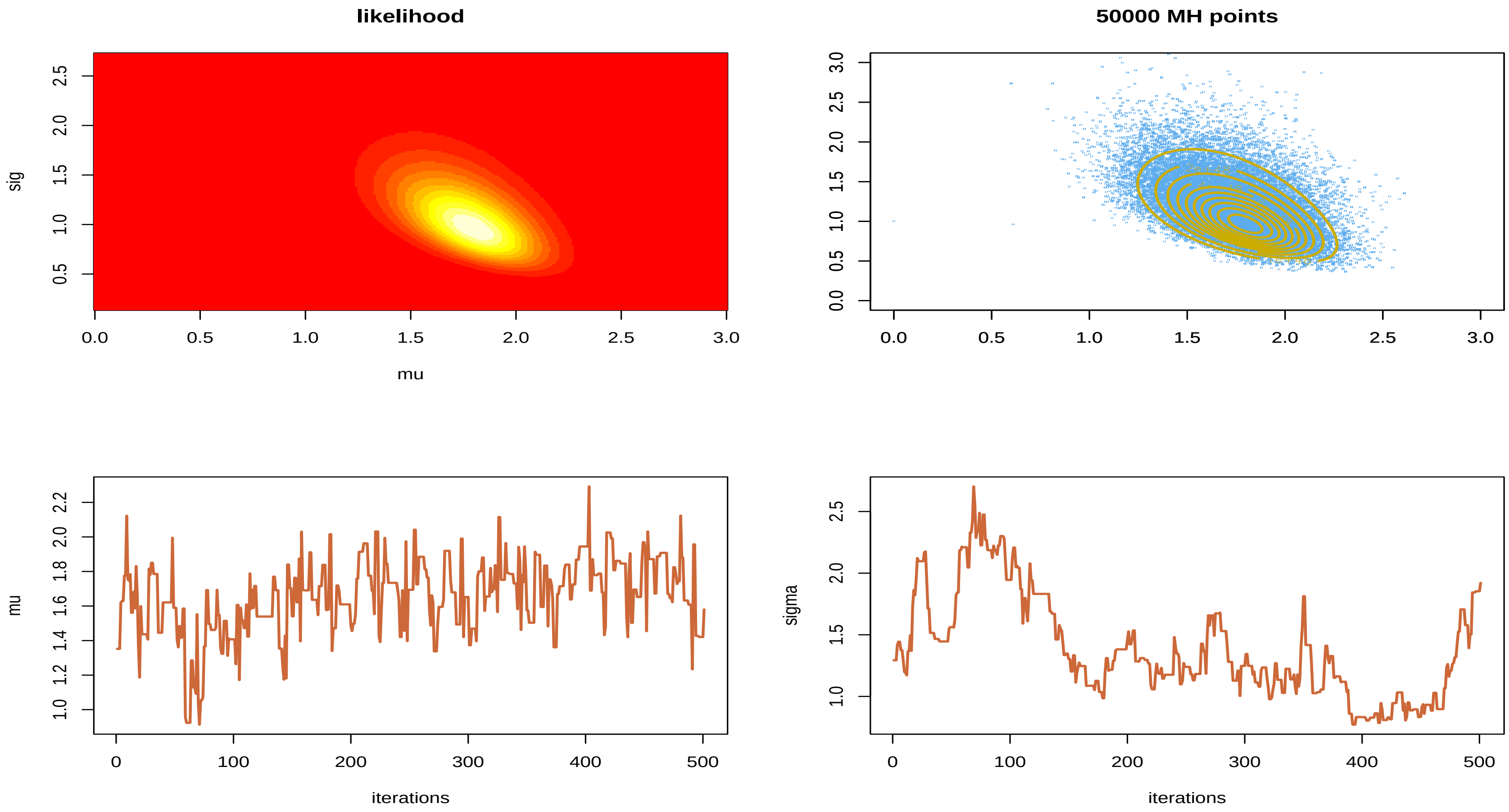
Example — Galaxy benchmark $(k = 4)$




Average density
Relative Frequency
0.0
0.2
0.4
0.6
0.8
−2
−1
0
1
2
3
data

Example — Simulated sample

$$0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(\mu,\sigma^2)$$



7 Hierarchical and Empirical Bayes Extensions, and the Stein Effect

The Bayesian analysis is sufficiently reductive to produce effective decisions, but this efficiency can also be misused.

The prior information is rarely rich enough to define a prior distribution exactly.

Uncertainty must be included within the Bayesian model:

- Further prior modelling
- Upper and lower probabilities[Dempster-Shafer]
- Imprecise probabilities [Walley]

7.1 Hierarchical Bayes analysis

Decomposition of the prior distribution into several conditional levels of distributions

Often two levels: the first-level distribution is generally a conjugate prior, with parameters distributed from the second-level distribution

Real life motivations (multiple experiments, meta-analysis, ...)

7.1.1 Hierarchical models

Definition 9 *A hierarchical Bayes model is a Bayesian statistical model, $(f(x|\theta),\ \pi(\theta))$, where*

$$\pi(\theta) = \int_{\Theta_1 \times \ldots \times \Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\cdots\pi_{n+1}(\theta_n)\, d\theta_1 \cdots d\theta_{n+1}.$$

The parameters θ_i are called hyperparameters of level i $(1 \leq i \leq n)$.

Example 42 Experiment where rats are intoxicated by a substance, then treated by either a placebo or a drug:

$$\begin{aligned}
x_{ij} &\sim \mathcal{N}(\theta_i, \sigma_c^2), & 1 \leq j \leq J_i^c\,, &\quad \text{control} \\
y_{ij} &\sim \mathcal{N}(\theta_i + \delta_i, \sigma_a^2), & 1 \leq j \leq J_i^a\,, &\quad \text{intoxication} \\
z_{ij} &\sim \mathcal{N}(\theta_i + \delta_i + \xi_i, \sigma_t^2), & 1 \leq j \leq J_i^t\,, &\quad \text{treatment}
\end{aligned}$$

Additional variable w_i, equal to 1 if the rat is treated with the drug, and 0 otherwise.

Prior distributions $(1 \leq i \leq I)$,

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2), \qquad \delta_i \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2),$$

and

$$\xi_i \sim \mathcal{N}(\mu_P, \sigma_P^2) \qquad \text{or} \qquad \xi_i \sim \mathcal{N}(\mu_D, \sigma_D^2),$$

depending on whether the ith rat is treated with a placebo or a drug.

Hyperparameters of the model,

$$\mu_\theta, \mu_\delta, \mu_P, \mu_D, \sigma_c, \sigma_a, \sigma_t, \sigma_\theta, \sigma_\delta, \sigma_P, \sigma_D\ ,$$

associated with Jeffreys' noninformative priors.

7.1.2 Justifications

(i). **Objective reasons based on prior information**

Example 43 (Example 42 continued) Alternative prior

$$\delta_i \sim p\mathcal{N}(\mu_{\delta 1}, \sigma^2_{\delta 1}) + (1-p)\mathcal{N}(\mu_{\delta 2}, \sigma^2_{\delta 2}),$$

allows for two possible levels of intoxication.

(ii). Separation of structural information from subjective information

Example 44 Uncertainties about generalized linear models

$$y_i|x_i \sim \exp\{\theta_i \cdot y_i - \psi(\theta_i)\}\,, \qquad \nabla\psi(\theta_i) = \mathbb{E}[y_i|x_i] = h(x_i^t\beta)\,,$$

where h is the *link* function

The linear constraint $\nabla\psi(\theta_i) = h(x_i^t\beta)$ can move to an higher level of the hierarchy,

$$\theta_i \sim \exp\left\{\lambda\left[\theta_i \cdot \xi_i - \psi(\theta_i)\right]\right\}$$

with $\mathbb{E}[\nabla\psi(\theta_i)] = h(x_i^t\beta)$ and

$$\beta \sim \mathcal{N}_q(0, \tau^2 I_q)$$

(iii). **In noninformative settings, compromise between the Jeffreys noninformative distributions, and the conjugate distributions.**

(iv). **Robustification of the usual Bayesian analysis from a frequentist point of view**

(v). **Often simplifies Bayesian calculations**

7.1.3 Conditional decompositions

Easy decomposition of the posterior distribution

For instance, if

$$\theta|\theta_1 \sim \pi_1(\theta|\theta_1), \qquad \theta_1 \sim \pi_2(\theta_1),$$

then

$$\pi(\theta|x) = \int_{\Theta_1} \pi(\theta|\theta_1, x)\pi(\theta_1|x)\, d\theta_1,$$

where

$$
\begin{aligned}
\pi(\theta|\theta_1, x) &= \frac{f(x|\theta)\pi_1(\theta|\theta_1)}{m_1(x|\theta_1)}, \\
m_1(x|\theta_1) &= \int_\Theta f(x|\theta)\pi_1(\theta|\theta_1)\, d\theta, \\
\pi(\theta_1|x) &= \frac{m_1(x|\theta_1)\pi_2(\theta_1)}{m(x)}, \\
m(x) &= \int_{\Theta_1} m_1(x|\theta_1)\pi_2(\theta_1)\, d\theta_1 .
\end{aligned}
$$

Moreover, this decomposition works for the posterior moments, that is, for every function h,

$$\mathbb{E}^{\pi}\left[h(\theta)|x\right] = \mathbb{E}^{\pi(\theta_1|x)}\left[\mathbb{E}^{\pi_1}\left[h(\theta)|\theta_1, x\right]\right],$$

where

$$\mathbb{E}^{\pi_1}\left[h(\theta)|\theta_1, x\right] = \int_{\Theta} h(\theta)\pi(\theta|\theta_1, x)\, d\theta.$$

Example 45 (Example 42 continued) The posterior distribution of the complete parameter vector is given by

$$\pi((\theta_i, \delta_i, \xi_i)_i, \mu_\theta, \ldots, \sigma_c, \ldots |\mathcal{D}) \propto$$

$$\prod_{i=1}^{I} \{\exp -\{(\theta_i - \mu_\theta)^2/2\sigma_\theta^2 + (\delta_i - \mu_\delta)^2/2\sigma_\delta^2\}$$

$$\prod_{j=1}^{J_i^c} \exp -\{(x_{ij} - \theta_i)^2/2\sigma_c^2\} \prod_{j=1}^{J_i^a} \exp -\{(y_{ij} - \theta_i - \delta_i)^2/2\sigma_a^2\}$$

$$\prod_{j=1}^{J_i^t} \exp -\{(z_{ij} - \theta_i - \delta_i - \xi_i)^2/2\sigma_t^2\} \Big\}$$

$$\prod_{\ell_i=0} \exp -\{(\xi_i - \mu_P)^2/2\sigma_P^2\} \prod_{\ell_i=1} \exp -\{(\xi_i - \mu_D)^2/2\sigma_D^2\}$$

$$\sigma_c^{-\sum_i J_i^c - 1} \sigma_a^{-\sum_i J_i^a - 1} \sigma_t^{-\sum_i J_i^t - 1} (\sigma_\theta \sigma_\delta)^{-I-1} \sigma_D^{-I_D - 1} \sigma_P^{-I_P - 1} ,$$

Local conditioning property

For the hierarchical model

$$\pi(\theta) = \int_{\Theta_1\times\ldots\times\Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\cdots\pi_{n+1}(\theta_n)\,d\theta_1\cdots d\theta_{n+1}.$$

we have

$$\pi(\theta_i|x,\theta,\theta_1,\ldots,\theta_n) = \pi(\theta_i|\theta_{i-1},\theta_{i+1})$$

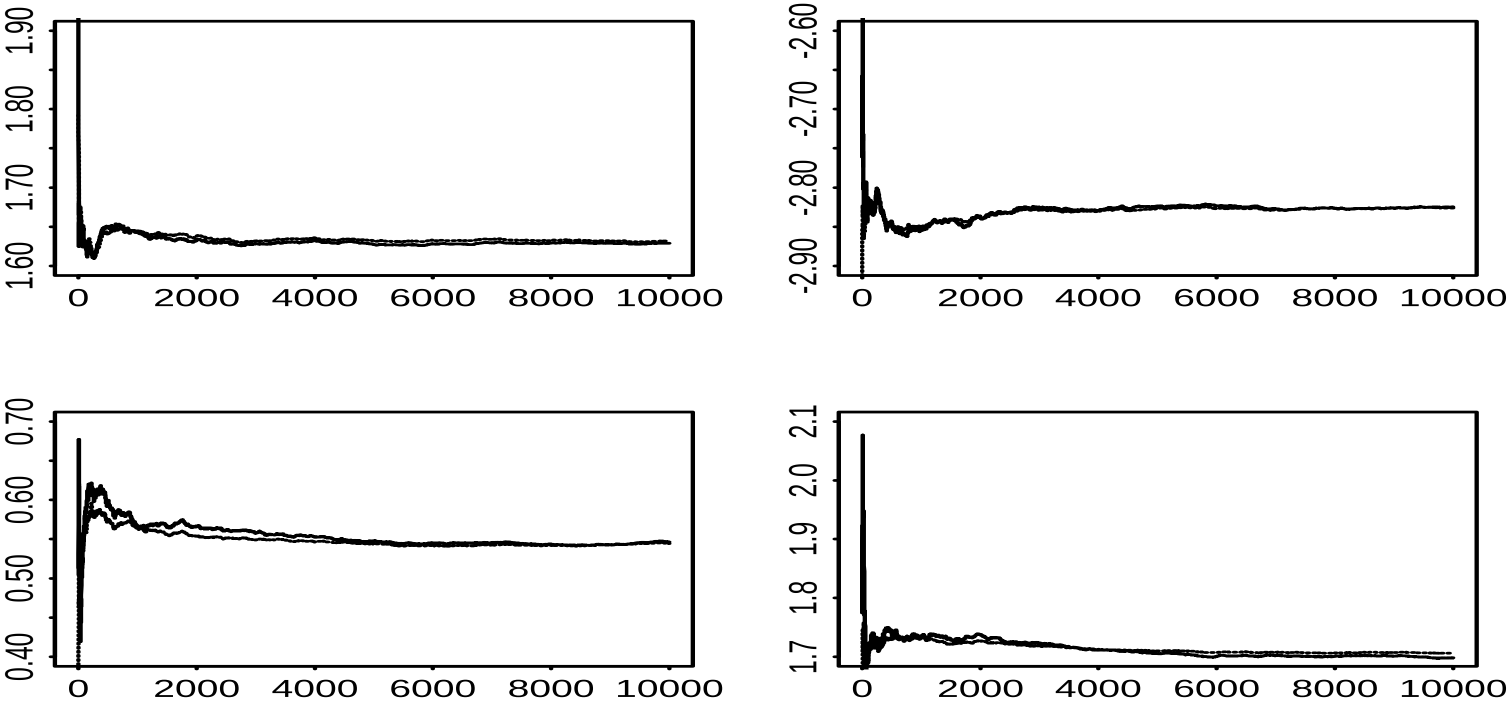
with the convention $\theta_0 = \theta$ **and** $\theta_{n+1} = 0$**.**

7.1.4 Computational issues

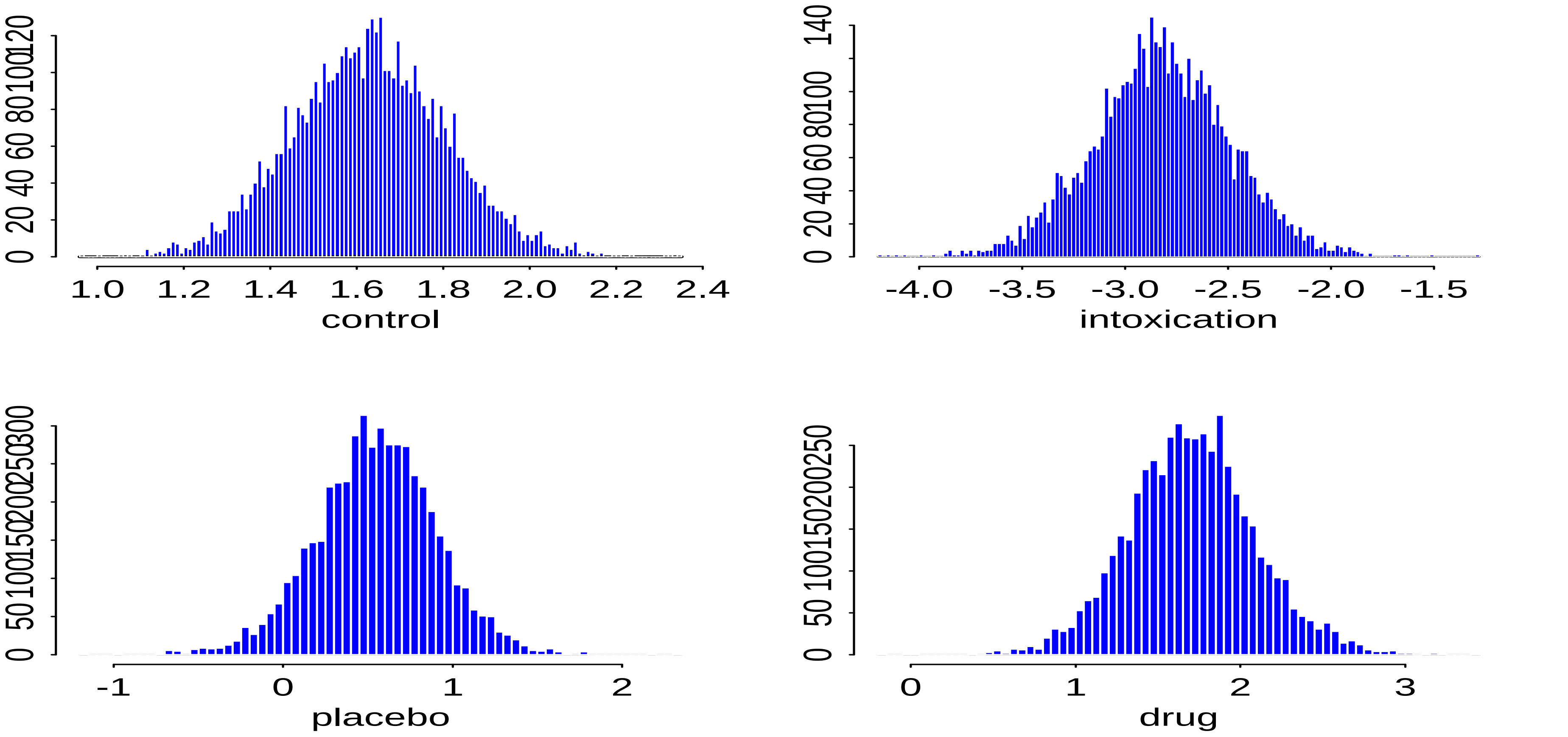
Rarely an explicit derivation of the corresponding Bayes estimators

Natural solution in hierarchical settings: use a simulation-based approach exploiting the hierarchical conditional structure

Example 46 (Example 42 continued) The full conditional distributions correspond to standard distributions and Gibbs sampling applies.



Convergence of the posterior means



Posteriors of the effects

	μ_δ	μ_D	μ_P	$\mu_D - \mu_P$
Probability	1.00	0.9998	0.94	0.985
Confidence	[-3.48,-2.17]	[0.94,2.50]	[-0.17,1.24]	[0.14,2.20]

Posterior probabilities of significant effects

7.1.5 Hierarchical extensions for the normal model

For

$$x \sim \mathcal{N}_p(\theta, \Sigma)\,, \qquad \theta \sim \mathcal{N}_p(\mu, \Sigma_\pi)$$

the hierarchical Bayes estimator is

$$\delta^\pi(x) = \mathbb{E}^{\pi_2(\mu,\Sigma_\pi|x)}[\delta(x|\mu, \Sigma_\pi)],$$

with

$$\begin{aligned}
\delta(x|\mu, \Sigma_\pi) &= x - \Sigma W(x-\mu), \\
W &= (\Sigma + \Sigma_\pi)^{-1}, \\
\pi_2(\mu, \Sigma_\pi|x) &\propto (\det W)^{1/2} \exp\{-(x-\mu)^t W(x-\mu)/2\}\pi_2(\mu, \Sigma_\pi).
\end{aligned}$$

Example 47 Consider the *exchangeable* hierarchical model

$$\begin{aligned} x|\theta &\sim \mathcal{N}_p(\theta, \sigma_1^2 I_p), \\ \theta|\xi &\sim \mathcal{N}_p(\xi \mathbf{1}, \sigma_\pi^2 I_p), \\ \xi &\sim \mathcal{N}(\xi_0, \tau^2), \end{aligned}$$

where $\mathbf{1} = (1, \ldots, 1)^t \in \mathbb{R}^p$. In this case,

$$\delta(x|\xi, \sigma_\pi) = x - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_\pi^2}(x - \xi \mathbf{1}),$$

$$\begin{aligned}
\pi_2(\xi,\sigma_\pi^2|x) &\propto (\sigma_1^2+\sigma_\pi^2)^{-p/2}\exp\{-\frac{\|x-\xi\mathbf{1}\|^2}{2(\sigma_1^2+\sigma_\pi^2)}\}e^{-(\xi-\xi_0)^2/2\tau^2}\pi_2(\sigma_\pi^2) \\
&\propto \frac{\pi_2(\sigma_\pi^2)}{(\sigma_1^2+\sigma_\pi^2)^{p/2}}\exp\left\{-\frac{p(\bar{x}-\xi)^2}{2(\sigma_1^2+\sigma_\pi^2)}-\frac{s^2}{2(\sigma_1^2+\sigma_\pi^2)}-\frac{(\xi-\xi_0)^2}{2\tau^2}\right\}
\end{aligned}$$

with $s^2=\sum_i(x_i-\bar{x})^2$. Then

$$\delta^\pi(x)=\mathbb{E}^{\pi_2(\sigma_\pi^2|x)}\left[x-\frac{\sigma_1^2}{\sigma_1^2+\sigma_\pi^2}(x-\bar{x}\mathbf{1})-\frac{\sigma_1^2+\sigma_\pi^2}{\sigma_1^2+\sigma_\pi^2+p\tau^2}(\bar{x}-\xi_0)\mathbf{1}\right]$$

and

$$\pi_2(\sigma_\pi^2|x)\propto\frac{\tau\exp-\frac{1}{2}\left[\frac{s^2}{\sigma_1^2+\sigma_\pi^2}+\frac{p(\bar{x}-\xi_0)^2}{p\tau^2+\sigma_1^2+\sigma_\pi^2}\right]}{(\sigma_1^2+\sigma_\pi^2)^{(p-1)/2}(\sigma_1^2+\sigma_\pi^2+p\tau^2)^{1/2}}\pi_2(\sigma_\pi^2).$$

Notice the particular form of the hierarchical Bayes estimator

$$\begin{aligned}\delta^\pi(x) &= x - \mathbb{E}^{\pi_2(\sigma_\pi^2|x)}\left[\frac{\sigma_1^2}{\sigma_1^2+\sigma_\pi^2}\right](x-\bar{x}\mathbf{1}) \\ &\quad -\mathbb{E}^{\pi_2(\sigma_\pi^2|x)}\left[\frac{\sigma_1^2+\sigma_\pi^2}{\sigma_1^2+\sigma_\pi^2+p\tau^2}\right](\bar{x}-\xi_0)\mathbf{1}.\end{aligned}$$

[Double shrinkage]

7.1.6 The Stein effect

If a minimax estimator is unique , it is admissible

Converse If a constant risk minimax estimator is inadmissible, every other minimax estimator has a uniformly smaller risk (!)

The Stein Paradox:

If a standard estimator $\delta^*(x) = (\delta_0(x_1), \ldots, \delta_0(x_p))$ is evaluated under weighted quadratic loss

$$\sum_{i=1}^{p} \omega_i(\delta_i - \theta_i)^2,$$

with $\omega_i > 0$ ($i = 1, \ldots, p$), there exists p_0 such that δ^* is not admissible for $p \geq p_0$, **although the components $\delta_0(x_i)$ are separately admissible to estimate the θ_i's**.

James–Stein estimator

In the normal case,

$$\delta^{JS}(x) = \left(1 - \frac{p-2}{||x||^2}\right) x,$$

dominates $\delta_0(x) = x$ under quadratic loss for $p \geq 3$, that is,

$$p = \mathbb{E}_\theta[||\delta_0(x) - \theta||^2] > \mathbb{E}_\theta[||\delta^{JS}(x) - \theta||^2].$$

And

$$\begin{aligned} \delta_c^+(x) &= \left(1 - \frac{c}{||x||^2}\right)^+ x \\ &= \begin{cases} (1 - \frac{c}{||x||^2})x & \text{if } ||x||^2 > c, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

improves on δ_0 when

$$0 < c < 2(p-2)$$

Universality

- Other distributions than the normal distribution
- Other losses other than the quadratic loss
- Connections with admissibility
- George’s multiple shrinkage
- Robustess against distribution
- Applies for confidence regions
- Applies for accuracy (or loss) estimation
- Cannot occur in finite parameter spaces

A general Stein-type domination result

Consider $z = (x^t, y^t)^t \in \mathbb{R}^p$, with distribution

$$z \sim f(||x - \theta||^2 + ||y||^2),$$

and $x \in \mathbb{R}^q$, $y \in \mathbb{R}^{p-q}$.

$$\delta_h(z) = (1 - h(||x||^2, ||y||^2))x$$

dominates δ_0 under quadratic loss if there exist $\alpha, \beta > 0$ such that:

(1) $t^\alpha h(t, u)$ **is a nondecreasing function of t for every u;**

(2) $u^{-\beta} h(t, u)$ **is a nonincreasing function of u for every t; and**

(3) $0 \le (t/u)h(t, u) \le \dfrac{2(q-2)\alpha}{p - q - 2 + 4\beta}$.

7.1.7 Optimality of hierarchical Bayes estimators

Consider

$$x \sim \mathcal{N}_p(\theta, \Sigma)$$

where Σ is known.

Prior distribution on θ is $\theta \sim \mathcal{N}_p(\mu, \Sigma_\pi)$.

The prior distribution π_2 of the hyperparameters

$$(\mu, \Sigma_\pi)$$

is decomposed as

$$\pi_2(\mu, \Sigma_\pi) = \pi_2^1(\Sigma_\pi|\mu)\pi_2^2(\mu).$$

In this case,

$$m(x) = \int_{\mathbb{R}^p} m(x|\mu)\pi_2^2(\mu)\, d\mu,$$

with

$$m(x|\mu) = \int f(x|\theta)\pi_1(\theta|\mu, \Sigma_\pi)\pi_2^1(\Sigma_\pi|\mu)\, d\theta\, d\Sigma_\pi.$$

Moreover, the Bayes estimator

$$\delta^\pi(x) = x + \Sigma \nabla \log m(x)$$

can be written

$$\delta^\pi(x) = \int \delta(x|\mu) \pi_2^2(\mu|x) \, d\mu,$$

with

$$\begin{aligned} \delta(x|\mu) &= x + \Sigma \nabla \log m(x|\mu), \\ \pi_2^2(\mu|x) &= \frac{m(x|\mu)\pi_2^2(\mu)}{m(x)}. \end{aligned}$$

A sufficient condition

An estimator δ is minimax under the loss

$$\mathrm{L}_Q(\theta,\delta) = (\theta-\delta)^t Q(\theta-\delta).$$

if it satisfies

$$R(\theta,\delta) = \mathbb{E}_\theta[\mathrm{L}_Q(\theta,\delta(x))] \leq \mathrm{tr}(\Sigma Q)$$

A sufficient condition (contd.)

If $m(x)$ **satisfies the three conditions**

$$(1)\ \mathbb{E}_\theta \|\nabla \log m(x)\|^2 < +\infty; \qquad (2)\ \mathbb{E}_\theta \left| \frac{\partial^2 m(x)}{\partial x_i \partial x_j} \Big/ m(x) \right| < +\infty;$$

and $(1 \le i \le p)$

$$(3)\ \lim_{|x_i| \to +\infty} \left|\nabla \log m(x)\right| \exp\{-(1/2)(x-\theta)^t \Sigma^{-1} (x-\theta)\} = 0,$$

the unbiased estimator of the risk of δ^π is given by

$$\begin{aligned}\mathcal{D}\delta^\pi(x) &= \mathrm{tr}(Q\Sigma)\\ &\quad+\frac{2}{m(x)}\mathrm{tr}(H_m(x)\tilde{Q}) - (\nabla\log m(x))^t\tilde{Q}(\nabla\log m(x))\end{aligned}$$

where

$$\tilde{Q} = \Sigma Q \Sigma, \qquad H_m(x) = \left(\frac{\partial^2 m(x)}{\partial x_i \partial x_j}\right)$$

and...

δ^π **is minimax if**

$$\mathrm{div}\left(\tilde{Q}\nabla\sqrt{m(x)}\right) \leq 0,$$

When $\Sigma = Q = I_p$, this condition is

$$\Delta\sqrt{m(x)} = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}(\sqrt{m(x)}) \leq 0$$

[$\sqrt{m(x)}$ superharmonic]

Superharmonicity condition

δ^π is minimax if

$$\operatorname{div}\left(\tilde{Q}\nabla m(x|\mu)\right) \leq 0.$$

N&S condition that does not depend on $\pi_2^2(\mu)$!

7.2 The empirical Bayes alternative

Strictly speaking, **not** a Bayesian method !

(i) can be perceived as a dual method of the hierarchical Bayes analysis;

(ii) *asymptotically* equivalent to the Bayesian approach;

(iii) usually classified as Bayesian by others; and

(iv) may be acceptable in problems for which a genuine Bayes modeling is too complicated/costly.

7.2.1 Parametric empirical Bayes

When hyperparameters from a conjugate prior $\pi(\theta|\lambda)$ are unavailable, estimate these hyperparameters from the marginal distribution

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

by $\hat{\lambda}(x)$ and to use $\pi(\theta|\hat{\lambda}(x), x)$ as a pseudo-posterior

Fundamental ad-hocquery

Which estimate $\hat{\lambda}(x)$ for λ ?

Moment method or maximum likelihood or Bayes or &tc...

Example 48 Consider x_i distributed according to $\mathcal{P}(\theta_i)$ $(i = 1, \ldots, n)$. When $\pi(\theta|\lambda)$ is $\mathcal{E}xp(\lambda)$,

$$\begin{aligned} m(x_i|\lambda) &= \int_0^{+\infty} e^{-\theta} \frac{\theta^{x_i}}{x_i!} \lambda e^{-\theta\lambda} d\theta \\ &= \frac{\lambda}{(\lambda+1)^{x_i+1}} = \left(\frac{1}{\lambda+1}\right)^{x_i} \frac{\lambda}{\lambda+1}, \end{aligned}$$

i.e. $x_i|\lambda \sim \mathcal{G}eo(\lambda/\lambda+1)$. Then

$$\hat{\lambda}(x) = 1/\bar{x}$$

and the empirical Bayes estimator of θ_{n+1} is

$$\delta^{\mathrm{EB}}(x_{n+1}) = \frac{x_{n+1}+1}{\hat{\lambda}+1} = \frac{\bar{x}}{\bar{x}+1}(x_{n+1}+1),$$

7.2.2 Empirical Bayes justifications of the Stein effect

A way to unify the different occurrences of this paradox and show its Bayesian roots

a. Point estimation

Example 49 Consider $x \sim \mathcal{N}_p(\theta, I_p)$ and $\theta_i \sim \mathcal{N}(0, \tau^2)$. The marginal distribution of x is then

$$x|\tau^2 \sim \mathcal{N}_p(0, (1+\tau^2)I_p)$$

and the maximum likelihood estimator of τ^2 is

$$\hat{\tau}^2 = \begin{cases} (||x||^2/p) - 1 & \text{if } ||x||^2 > p, \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding empirical Bayes estimator of θ_i is then

$$\begin{aligned} \delta^{\mathrm{EB}}(x) &= \frac{\hat{\tau}^2 x}{1+\hat{\tau}^2} \\ &= \left(1 - \frac{p}{||x||^2}\right)^{+} x. \end{aligned}$$

[truncated James–Stein]

Normal model

$$\begin{aligned} x|\theta &\sim \mathcal{N}_p(\theta, \Lambda), \\ \theta|\beta, \sigma_\pi^2 &\sim \mathcal{N}_p(Z\beta, \sigma_\pi^2 I_p), \end{aligned}$$

with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and Z a $(p \times q)$ full rank matrix.

The marginal distribution of x is

$$x_i|\beta, \sigma_\pi^2 \sim \mathcal{N}(z_i'\beta, \sigma_\pi^2 + \lambda_i)$$

and the posterior distribution of θ is

$$\theta_i|x_i, \beta, \sigma_\pi^2 \sim \mathcal{N}\left((1 - b_i)x_i + b_i z_i'\beta, \lambda_i(1 - b_i)\right),$$

with $b_i = \lambda_i/(\lambda_i + \sigma_\pi^2)$.

If

$$\lambda_1 = \ldots = \lambda_n = \sigma^2$$

the best equivariant estimators of β and b are given by

$$\hat{\beta} = (Z^t Z)^{-1} Z^t x \qquad \text{and} \qquad \hat{b} = \frac{(p-q-2)\sigma^2}{s^2},$$

with $s^2 = \sum_{i=1}^{p} (x_i - z_i' \hat{\beta})^2$.

The corresponding empirical Bayes estimator of θ are

$$\delta^{\mathrm{EB}}(x) = Z\hat{\beta} + \left(1 - \frac{(p-q-2)\sigma^2}{||x - Z\hat{\beta}||^2}\right)(x - Z\hat{\beta}),$$

which is of the form of the general Stein estimators.

When the means are assumed to be identical (exchangeability), the matrix Z reduces to the vector $\mathbf{1}$ and $\beta \in \mathbb{R}$

The empirical Bayes estimator is then

$$\delta^{\mathrm{EB}}(x) = \bar{x}\mathbf{1} + \left(1 - \frac{(p-3)\sigma^2}{||x - \bar{x}\mathbf{1}||^2}\right)(x - \bar{x}\mathbf{1}).$$

b. Variance evaluation

Estimation of the hyperparameters β and σ_π^2 considerably modifies the behavior of the procedures.

Point estimation generally efficient, but estimation of the posterior variance of $\pi(\theta|x, \beta, b)$ by the empirical variance,

$$\mathrm{var}(\theta_i|x, \hat{\beta}, \hat{b})$$

induces an underestimation of this variance

Morris' correction

$$\begin{aligned} \delta^{\mathrm{EB}}(x) &= x - \tilde{B}(x - \bar{x}\mathbf{1}), \\ V_i^{\mathrm{EB}}(x) &= \left(\sigma^2 - \frac{p-1}{p}\tilde{B}\right) + \frac{2}{p-3}\hat{b}(x_i - \bar{x})^2, \end{aligned}$$

with

$$\hat{b} = \frac{p-3}{p-1}\,\frac{\sigma^2}{\sigma^2 + \hat{\sigma}_\pi^2}, \qquad \hat{\sigma}_\pi^2 = \max\left(0, \frac{||x - \bar{x}\mathbf{1}||^2}{p-1} - \sigma_\pi^2\right)$$

and

$$\tilde{B} = \frac{p-3}{p-1}\,\min\left(1, \frac{\sigma^2(p-1)}{||x - \bar{x}\mathbf{1}||^2}\right).$$

8 A Defense of the Bayesian Choice

Unlimited range of applications

- artificial intelligence
- biostatistic
- econometrics
- epidemiology
- environmetrics
- finance

- genomics
- geostatistics
- image processing and pattern recognition
- neural networks
- signal processing
- Bayesian networks

(1). **Choosing a probabilistic representation**

Bayesian Statistics appears as the calculus of uncertainty

Reminder:

A probabilistic model is nothing but an *interpretation* of a given phenomenon

(2). **Conditioning on the data**

At the basis of statistical inference lies an *inversion process* between cause and effect. Using a prior distribution brings a necessary balance between observations and parameters and enable to operate *conditional upon* x

(3). **Exhibiting the true likelihood**

Provides a complete *quantitative inference* on the parameters and predictive that points out inadequacies of frequentist statistics, while implementing the Likelihood Principle.

(4). Using priors as tools and summaries

The choice of a prior distribution π does not require any kind of *belief* in this distribution: rather consider it as a *tool* that *summarizes* the available prior information *and* the uncertainty surrounding this information

(5). **Accepting the subjective basis of knowledge**

Knowledge is a critical confrontation between *a prioris* and experiments.

Ignoring these *a prioris* impoverishes analysis.

We have, for one thing, to use a language and our language is entirely made of preconceived ideas and has to be so. However, these are unconscious preconceived ideas, which are a million times more dangerous than the other ones. Were we to assert that if we are including other preconceived ideas, consciously stated, we would aggravate the evil! I do not believe so: I rather maintain that they would balance one another.

Henri Poincaré, 1902

(6). **Choosing a coherent system of inference**

To force inference into a decision-theoretic mold allows for a clarification of the way inferential tools should be evaluated, and therefore implies a conscious (although subjective) choice of the *retained optimality*.

Logical inference process Start with requested properties, i.e. loss function and prior distribution, then derive the best solution satisfying these properties.

(7). **Looking for optimal frequentist procedures**

Bayesian inference widely intersects with the three notions of minimaxity, admissibility and equivariance. Looking for an optimal estimator most often ends up finding a Bayes estimator.

Optimality is easier to attain through the Bayes “filter”.

(8). **Solving the actual problem**

Frequentist methods justified on a *long-term* basis, i.e., from the statistician viewpoint. From a decision-maker's point of view, only the problem at hand matters! That is, he/she calls for an inference *conditional* on x.

(9). Providing a universal system of inference

Given the three factors

$$(\mathcal{X}, f(x|\theta), \quad (\Theta, \pi(\theta)), \quad (\mathcal{D}, \mathrm{L}(\theta, d)) \, ,$$

the Bayesian approach validates one and only one inferential procedure

(10). **Computing procedures as a minimization problem**

Bayesian procedures are *easier to compute* than procedures of alternative theories, in the sense that there exists a *universal method* for the computation of Bayes estimators

In practice, the *effective* calculation of the Bayes estimators is often more delicate but this defect is of another magnitude.