Bayesian Statistics

Christian P. Robert

Université Paris Dauphine and CREST-INSEE
http://www.ceremade.dauphine.fr/~xian

January 9, 2006

Outline

Introduction

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions

Bayesian Point Estimation

Bayesian Calculations

Tests and model choice

Admissibility and Complete Classes

Hierarchical and Empirical Bayes Extensions, and the Stein Effect

Vocabulary, concepts and first examples

Introduction
- Models
- The Bayesian framework
- Prior and posterior distributions
- Improper prior distributions

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions

Bayesian Point Estimation

Bayesian Calculations

Parametric model

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

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

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]

Bayes Theorem

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

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

Bayes Theorem

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]

Bayes Theorem

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]

Actualisation principle

New perspective

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

New perspective

- *Uncertainty* on the parameter s 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 (Bayesian model)

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

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

Definition (Bayesian model)

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

Justifications

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

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

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)

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.

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.

Bayes' question

Given X, what inference can we make on p?

Modern translation:

Derive the posterior distribution of p given X, when

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

Resolution

Since

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

and

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

Resolution (2)

then

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

Resolution (2)

then

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

i.e.

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

[Beta distribution]

Prior and posterior distributions

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

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

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

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,

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

(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 observation s

Posterior distribution

central to Bayesian inference

- Operates conditional upon the observation s
- Incorporates the requirement of the Likelihood Principle

Posterior distribution

central to Bayesian inference

- Operates conditional upon the observation s
- Incorporates the requirement of the Likelihood Principle
- Avoids averaging over the unobserved values of x

Posterior distribution

central to Bayesian inference

- Operates conditional upon the observation s
- 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

Posterior distribution

central to Bayesian inference

- Operates conditional upon the observation s
- 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 (Flat prior (1))

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

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

Example (Flat prior (1))

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

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

and

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

Example (HPD region)

Natural confidence region

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

Example (HPD region)

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

Improper distributions

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

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

Improper 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

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

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

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*

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 $\mathscr{N}(0, 100^2)$ distribution

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 $\mathscr{N}(0, 100^2)$ 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

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

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

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

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

Example

If $x \sim \mathscr{N}(\theta, 1)$ 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 a $\mathscr{N}(x, 1)$ distribution.

Example

If $x \sim \mathscr{N}(\theta, 1)$ 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 a $\mathscr{N}(x, 1)$ distribution.

[independent of ω]

Warning - Warning - Warning - Warning - Warning

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

[Lindley, 1990]

Example (Flat prior (2))

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

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

for any (a, b)

Example ([Haldane prior)

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

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

[Haldane, 1931]

Example ([Haldane prior)

Consider a binomial observation, $x \sim \mathscr{B}(n,p)$, 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 $x \neq 0, n$.

Decision theory motivations

Introduction

Decision-Theoretic Foundations of Statistical Inference
- Evaluation of estimators
- Loss functions
- Minimaxity and admissibility
- Usual loss functions

From Prior Information to Prior Distributions

Bayesian Point Estimation

Bayesian Calculations

Evaluating estimators

Purpose of most inferential studies

To provide the statistician/client with a *decision* $d \in \mathscr{D}$

Evaluating estimators

Purpose of most inferential studies

To provide the statistician/client with a *decision* $d \in \mathscr{D}$
Requires an evaluation criterion for decisions and estimators

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

[a.k.a. loss function]

Bayesian Decision Theory

Three spaces/factors:

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

Bayesian Decision Theory

Three spaces/factors:

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

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

Bayesian Decision Theory

Three spaces/factors:

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

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

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

Foundations

Theorem (Existence)

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

[DeGroot, 1970]

Estimators

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

Estimators

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

Fact

Impossible to uniformly minimize (in d) the loss function

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

when θ is unknown

Frequentist Principle

Average loss (or frequentist risk)

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

Frequentist Principle

Average loss (or frequentist risk)

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

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!

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.

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

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

Bayesian principle (2)

Alternative

Integrate over the space Θ and compute *integrated risk*

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

which induces a **total** ordering on estimators.

Bayesian principle (2)

Alternative

Integrate over the space Θ and compute *integrated risk*

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

which induces a **total** ordering on estimators.

Existence of an optimal decision

Bayes estimator

Theorem (Construction of Bayes estimators)

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

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

Bayes estimator

Theorem (Construction of Bayes estimators)

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

Bayes estimator (2)

Definition (Bayes optimal procedure)

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

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

Infinite Bayes risk

Above result valid for both proper and improper priors when

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

Otherwise, **generalized Bayes estimator** that must be defined pointwise:

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

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

Infinite Bayes risk

Above result valid for both proper and improper priors when

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

Otherwise, **generalized Bayes estimator** that must be 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

Minimaxity

Frequentist insurance against the worst case and (weak) total ordering on $\mathscr{D}^*$

Minimaxity

Frequentist insurance against the worst case and (weak) total ordering on $\mathscr{D}^*$

Definition (Frequentist optimality)

The *minimax risk* associated with a loss L is

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

Minimaxity

Frequentist insurance against the worst case and (weak) total ordering on $\mathscr{D}^*$

Definition (Frequentist optimality)

The *minimax risk* associated with a loss L is

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

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

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

Criticisms

- Analysis in terms of the worst case

Criticisms

- Analysis in terms of the worst case
- Does not incorporate prior information

Criticisms

- Analysis in terms of the worst case
- Does not incorporate prior information
- Too conservative

Criticisms

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

Example (Normal mean)

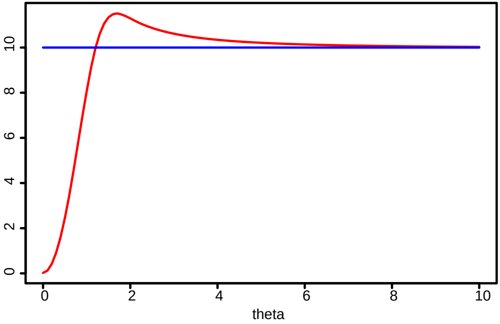
Consider

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

to estimate θ when $x \sim \mathscr{N}_p(\theta, I_p)$ 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

Minimaxity (2)

Existence

If $\mathscr{D} \subset \mathbb{R}^k$ convex and compact, and if $\mathsf{L}(\theta, d)$ 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:

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

Connection with Bayesian approach

The Bayes risks are always smaller than the minimax risk:

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

Definition

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

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

$\underline{r}$ is the *maximin risk* and the corresponding π the *favourable prior*

Maximin-ity

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

Maximin-ity (2)

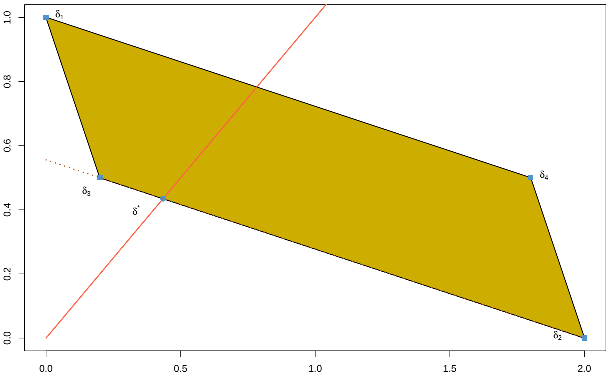
Example (Binomial probability)

Consider $x \sim \mathscr{B}e(\theta)$ 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_3
δ^*
δ_4
δ_2
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0

Risk set

Example (Binomial probability (2))

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

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

Example (Binomial probability (2))

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

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

Also randomized Bayes estimator for

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

Checking minimaxity

Theorem (Bayes & minimax)

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, then δ_0 is minimax and π_0 is the least favourable distribution

Example (Binomial probability (3))

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

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

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 (2)

Theorem (Bayes & minimax (2))

If for a sequence (π_n) of proper priors, the generalised 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 (Normal mean)

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

$$\delta_0(x) = x$$

is a generalised Bayes estimator associated with

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

Example (Normal mean)

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

$$\delta_0(x) = x$$

is a generalised 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.

Admissibility

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

Definition (Admissible estimator)

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

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

Admissibility

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

Definition (Admissible estimator)

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

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

Otherwise, δ_0 is admissible

Minimaxity & admissibility

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

The converse is false!

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!

The Bayesian perspective

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

The 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.

The 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 admissible

Example (Normal mean)

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

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

Example (Normal mean)

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.

Example (Normal mean (2))

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 (Normal mean (3))

Consider $x \sim \mathscr{N}(\theta, 1)$. Then $\delta_0(x) = x$ is a generalised 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 (Normal mean (4))

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

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

the Bayes estimator for the Lebesgue measure is

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

This estimator is not admissible because it is dominated by

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

The quadratic loss

Historically, first loss function (Legendre, Gauss)

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

The quadratic loss

Historically, first loss function (Legendre, Gauss)

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

or

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

Proper loss

Posterior mean

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

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

The absolute error loss

Alternatives to the quadratic loss:

$$\mathsf{L}(\theta, d) = |\,\theta - d\,|\,,$$

or

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

The absolute error loss

Alternatives to the quadratic loss:

$$\mathsf{L}(\theta, d) = |\,\theta - d\,| \,,$$

or

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

L_1 estimator

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

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

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

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

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

The $0-1$ loss

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

Type–one and type–two errors

Associated with the risk

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

Type–one and type–two errors

Associated with the risk

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

Theorem (Bayes test)

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

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

Intrinsic losses

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

[Ultimate invariance!]

Principle

Corresponding parameterisation-free loss functions:

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

Examples:

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

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

Examples:

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

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

2. the *Hellinger distance*

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

Example (Normal mean)

Consider $x \sim \mathscr{N}(\theta, 1)$. Then

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

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

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

in both cases.

From prior information to prior distributions

Introduction

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions
- Models
- Subjective determination
- Conjugate priors
- Noninformative prior distributions

Bayesian Point Estimation

Bayesian Calculations

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

Subjective priors

Example (Capture probabilities)

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]

Example (Capture probabilities (2))

Corresponding prior modeling

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

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

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*

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

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,

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

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

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.

μ_2	x	Minimum mean	Maximum mean	Maximum 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]

Conjugate priors

Specific parametric family with analytical properties

Definition

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

[Raiffa & Schlaifer, 1961]

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

Justifications

- Limited/finite information conveyed by x

Justifications

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

Justifications

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

Justifications

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

Justifications

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

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

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

The family of distributions

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

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

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

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 $x_0 \in \mathscr{X}$, 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

When $x \sim \mathscr{B}e(\alpha, \theta)$ 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

Coin spun on its edge, proportion θ of *heads*
When spinning n times a given coin, number of heads

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

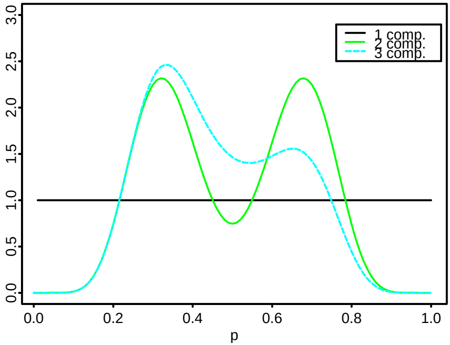
Flat prior, or mixture prior

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

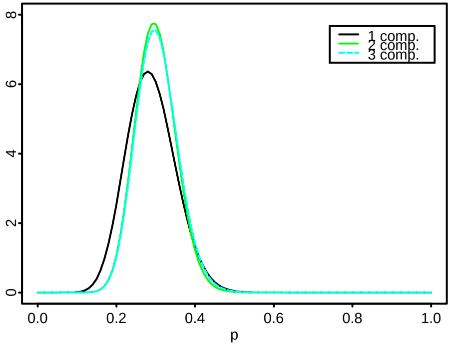
or

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

Mixtures of natural conjugate distributions also make conjugate families



Three prior distributions for a spinning-coin experiment



Posterior distributions for 50 observations

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]

Laplace's prior

Principle of *Insufficient Reason* (Laplace)

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

Extension to continuous spaces

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

- Lack of reparameterization invariance/coherence

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

- Problems of properness

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

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

Invariant priors

Principle: Agree with the natural symmetries of the problem

- Identify invariance structures as group action

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

- 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)

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

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

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

Example

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

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

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

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

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 (Neyman–Scott problem)

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

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

Example (Neyman–Scott problem)

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,

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

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 (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

Example (Linear calibration model (2))

One-sided differential equation:

$$|\beta|^{-1} s^{-1/2} \frac{\partial}{\partial x_0}\{e(x_0)\pi(\theta)\} - e^{-1/2}(x_0)\mathrm{sgn}(\beta) n^{-1} s^{1/2} \frac{\partial \pi(\theta)}{\partial x_0}$$
$$-e^{-1/2}(x_0)(x_0 - \bar{x}) s^{-1/2} \frac{\partial}{\partial \beta}\{\mathrm{sgn}(\beta)\pi(\theta)\} = 0$$

with

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

Example (Linear calibration model (3))

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

Other approaches

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

Bayesian Point Estimation

Introduction

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions

Bayesian Point Estimation
- Bayesian inference
- Bayesian Decision Theory
- The particular case of the normal model
- Dynamic models

Bayesian Calculations

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

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

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

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

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

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

Consider the AR(1) model

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

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

Bayesian Decision Theory

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

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

Note: Practical computation not always possible analytically.

Conjugate priors

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

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

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

Consider

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

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

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

and

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

Even conjugate priors may lead to computational difficulties

Example

Consider $x \sim \mathscr{N}_p(\theta, I_p)$ and

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

for which $\delta_0(x) = ||x||^2 - p$ has a constant risk, 1
For the conjugate distributions, $\mathscr{N}_p(0, \tau^2 I_p)$,

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

cannot be computed analytically.

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

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

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

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

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

Linear models

Usual regression modelregression!model

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

Conjugate distributions of the type

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

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: $\epsilon \sim \mathscr{N}_k(0, \sigma^2 I_k)$
The least-squares estimator $\hat{\beta}$ has a normal distribution

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

Corresponding conjugate distribution s on (β, σ^2)

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

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

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

The AR(p) model

Markovian dynamic model

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

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

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

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 distributiondistribution!normal on $(\mu, \varrho_1, \ldots, \rho_p)$ and an inverse gamma distributiondistribution!inverse gamma 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 $\mathscr{P}$, whose inverses are also within the unit circle.

Stationarity & priors (contd.)

Jeffreys' prior associated with the stationary representationrepresentation!stationary 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

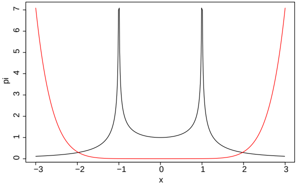
$$\pi_2^J(\mu, \sigma^2, \varrho) \propto \frac{1}{\sigma^2} \frac{1}{\sqrt{|1-\varrho^2|}} \sqrt{\left|1 - \frac{1-\varrho^{2T}}{T(1-\varrho^2)}\right|} .$$

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 $|\varrho| > 1$

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



The MA(q) model

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

Stationary but, for identifiability considerations, the polynomial

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

must have all its roots outside the unit circle.

Example

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

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

It can also be written

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

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 ($|s| \leq q$)

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

Not manageable in practice

Representations (contd.)

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

$$L(\mu, \vartheta_1, \ldots, \vartheta_q, \sigma | x_{1:T}, \epsilon_0, \ldots, \epsilon_{-q+1}) =$$
$$\sigma^{-T} \prod_{t=1}^{T} \exp\left\{ -\left(x_t - \mu + \sum_{j=1}^{q} \vartheta_j \hat{\epsilon}_{t-j} \right)^2 / 2\sigma^2 \right\},$$

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 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

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

and

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

Example

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 $\mathscr{Q}(x)$: the ϑ_j's vary in a complex space.
New reparametrization: the ψ_i's are the *inverse partial auto-correlations*

Bayesian Calculations

Introduction

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions

Bayesian Point Estimation

Bayesian Calculations
- Implementation difficulties
- Classical approximation methods
- Markov chain Monte Carlo methods

Tests and model choice

B Implementation difficulties

- Computing the posterior distribution

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

B Implementation difficulties

- Computing the posterior distribution

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

- Resolution of

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

B Implementation difficulties

- Computing the posterior distribution

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

- Resolution of

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

- Maximisation of the marginal posterior

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

B Implementation further difficulties

- Computing posterior quantities

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

B Implementation further difficulties

- Computing posterior quantities

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

- Resolution (in k) of

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

Example (Cauchy posterior)

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

with known hyperparameters μ and σ^2.

Example (Cauchy posterior)

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

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 (Mixture of two normal distributions)

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

Example (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

$$\mu_i|\sigma_i \sim \mathscr{N}(\xi_i, \sigma_i^2/n_i), \quad \sigma_i^2 \sim \mathscr{I}\mathscr{G}(\nu_i/2, s_i^2/2), \quad p \sim \mathscr{B}e(\alpha, \beta)$$

Example (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

$$\mu_i|\sigma_i \sim \mathscr{N}(\xi_i, \sigma_i^2/n_i), \quad \sigma_i^2 \sim \mathscr{I}\mathscr{G}(\nu_i/2, s_i^2/2), \quad p \sim \mathscr{B}e(\alpha, \beta)$$

Posterior

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

[$O(2^n)$]

Example (Mixture of two normal distributions (2))

For a given permutation (k_t), conditional posterior distribution

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

Example (Mixture of two normal distributions (3))

where

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

Example (Mixture of two normal distributions (3))

where

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

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

Example (Mixture of two normal distributions (4))

Bayes estimator of θ:

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

Example (Mixture of two normal distributions (4))

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

Numerical integration

▸ Switch to Monte Carlo

- Simpson's method

Numerical integration

Switch to Monte Carlo

- Simpson's method
- polynomial quadrature

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

Numerical integration

Switch to Monte Carlo

- 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*, $H_n(t)$.

Numerical integration

Switch to Monte Carlo

- 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]

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.

MC Principle

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 $\mathfrak{I}$

MC Principle

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.

[Commercial!!]

Importance function

No need to simulate from $\pi(\cdot|x)$ or from π

Importance function

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

[Importance function]

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

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

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

Requirements

- Simulation from h must be easy

Requirements

- Simulation from h must be easy
- $h(\theta)$ must be close enough to $g(\theta)\pi(\theta|x)$

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 (Cauchy Example continued)

Since $\pi(\theta)$ is $\mathscr{N}(\mu, \sigma^2)$, possible to simulate a normal sample $\theta_1, \ldots, \theta_M$ and to approximate the Bayes estimator by

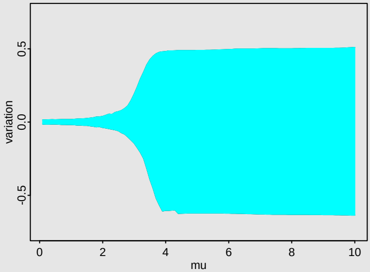
$$\frac{\sum_{t=1}^{M} \theta_t \prod_{i=1}^{n}[1 + (x_i - \theta_t)^2]^{-1}}{\sum_{t=1}^{M} \prod_{i=1}^{n}[1 + (x_i - \theta_t)^2]^{-1}}$$

The importance function may be π

Example (Cauchy Example continued)

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

$$\frac{\sum_{t=1}^{M} \theta_t \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}{\sum_{t=1}^{M} \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}$$



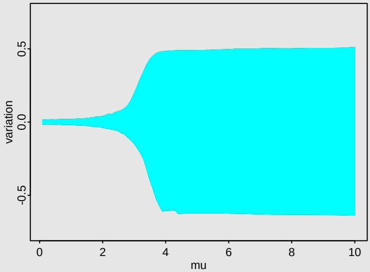
90% variation

The importance function may be π

Example (Cauchy Example continued)

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

$$\frac{\sum_{t=1}^{M} \theta_t \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}{\sum_{t=1}^{M} \prod_{i=1}^{n} [1 + (x_i - \theta_t)^2]^{-1}}$$



90% variation

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

Defensive sampling

Use a mix of prior and posterior

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

[Newton & Raftery, 1994]

Defensive sampling

Use a mix of prior and posterior

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

[Newton & Raftery, 1994]

Requires proper knowledge of normalising constants

[Bummer!]

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]

Bridge sampling

If

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

on same space,

Bridge sampling

If

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

on same space, 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]

Further bridge sampling

Also

$$B_{12} = \frac{\displaystyle\int \tilde{\pi}_2(\theta)\alpha(\theta)\pi_1(\theta)d\theta}{\displaystyle\int \tilde{\pi}_1(\theta)\alpha(\theta)\pi_2(\theta)d\theta} \qquad \forall\ \alpha(\cdot)$$

$$\approx \frac{\displaystyle\frac{1}{n_1}\sum_{i=1}^{n_1} \tilde{\pi}_2(\theta_{1i})\alpha(\theta_{1i})}{\displaystyle\frac{1}{n_2}\sum_{i=1}^{n_2} \tilde{\pi}_1(\theta_{2i})\alpha(\theta_{2i})} \qquad \theta_{ji} \sim \pi_j(\theta)$$

Umbrella sampling

Parameterized version

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

Then

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

Umbrella sampling

Parameterized version

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

Then

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

and

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

MCMC methods

Idea

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

Formal Warranty

Convergence

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

Formal Warranty

Convergence

if the Markov chains produced by MCMC algorithms are irreducible, these chains are both 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

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

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

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

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

- Achieve quasi-independence by batch sampling

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

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

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

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

Metropolis–Hastings algorithms

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

HM Algorithm

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

Metropolis–Hastings algorithms

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

HM Algorithm

1. Start with an arbitrary initial value $\theta^{(0)}$
2. Update from $\theta^{(m)}$ to $\theta^{(m+1)}$ $(m = 1, 2, \ldots)$ by
   2.1 Generate $\xi \sim q(\xi|\theta^{(m)})$
   2.2 Define

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

Metropolis–Hastings algorithms

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

HM Algorithm

1. Start with an arbitrary initial value $\theta^{(0)}$
2. Update from $\theta^{(m)}$ to $\theta^{(m+1)}$ $(m = 1, 2, \ldots)$ by
   2.1 Generate $\xi \sim q(\xi|\theta^{(m)})$
   2.2 Define
   $$\varrho = \frac{\pi(\xi)\, q(\theta^{(m)}|\xi)}{\pi(\theta^{(m)})\, q(\xi|\theta^{(m)})} \wedge 1$$
   2.3 Take
   $$\theta^{(m+1)} = \begin{cases} \xi & \text{with probability } \varrho, \\ \theta^{(m)} & \text{otherwise.} \end{cases}$$

Validation

Detailed balance condition

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

Validation

Detailed balance condition

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

$K(\theta'|\theta)$ 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

Take

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

Random walk Metropolis–Hastings

Take

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

Corresponding Metropolis–Hastings acceptance ratio

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

Example (Repulsive normal)

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

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

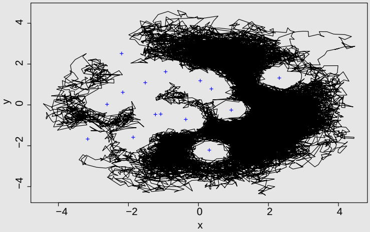
where the μ_i's are given repulsive points

Example (Repulsive normal)

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

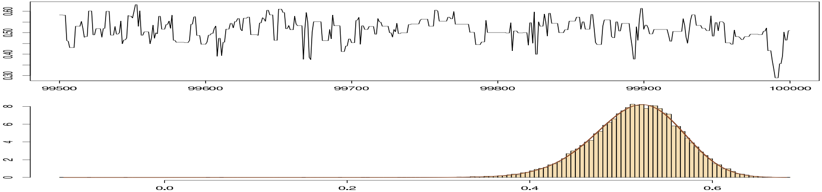


Path of the Markov chain (5000 iterations).

Pros & Cons

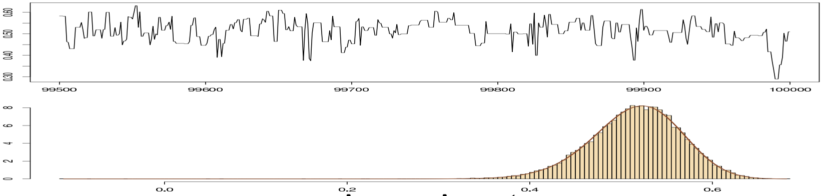
- Widely applicable
- limited tune-up requirements (scale calibrated thru acceptance)
- never uniformely ergodic

Noisy AR_1^2

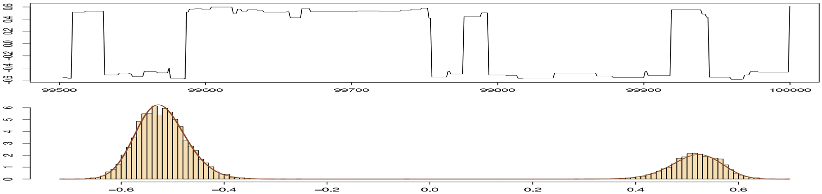


scale equal to $.1$

Noisy AR_1^2



scale equal to .1



scale equal to .5

Independent proposals

Take

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

Independent proposals

Take

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

More limited applicability and closer connection with iid simulation

Independent proposals

Take

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

More limited applicability and closer connection with iid simulation

Examples

- prior distribution
- likelihood
- saddlepoint approximation

The Gibbs sampler

Take advantage of hierarchical structures

The Gibbs sampler

Take advantage of hierarchical structures

If

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

simulate instead from the joint distribution

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

Example (beta-binomial)

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

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

Example (beta-binomial)

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:

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

then

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

[beta-binomial distribution]

Example (beta-binomial (2))

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

Example (beta-binomial (2))

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

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

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

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

Conditionals

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

Conditionals

Usually $\pi_2(\lambda|x)$ is 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

DA Algorithm

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

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

1. $\theta^{(t)}$ according to $\pi_1(\theta|x, \lambda^{(t-1)})$
2. $\lambda^{(t)}$ according to $\pi_2(\lambda|x, \theta^{(t)})$

Data augmentation

DA Algorithm

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

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

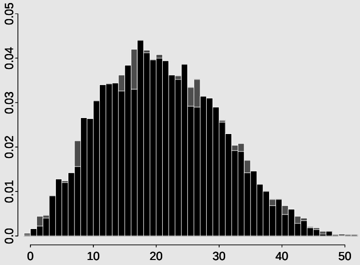
1. $\theta^{(t)}$ according to $\pi_1(\theta|x, \lambda^{(t-1)})$
2. $\lambda^{(t)}$ according to $\pi_2(\lambda|x, \theta^{(t)})$

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

Example (Beta-binomial Example cont'ed)

The conditional distributions are

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



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

Very simple example: Independent $\mathrm{N}(\mu, \sigma^2)$ obs'ions

When $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{N}(y|\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate but non-standard

Very simple example: Independent $\mathrm{N}(\mu, \sigma^2)$ obs'ions

When $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{N}(y|\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate but non-standard

But...

$$\mu|Y_{0:n}, \sigma^2 \sim \mathrm{N}\left(\mu \,\middle|\, \frac{1}{n}\sum_{i=1}^n Y_i, \frac{\sigma^2}{n}\right)$$
$$\sigma^2|Y_{1:n}, \mu \sim \mathrm{IG}\left(\sigma^2 \,\middle|\, \frac{n}{2} - 1, \frac{1}{2}\sum_{i=1}^n (Y_i - \mu)^2\right)$$

assuming constant (improper) priors on both μ and σ^2

Very simple example: Independent $\mathrm{N}(\mu, \sigma^2)$ obs'ions

When $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathrm{N}(y|\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate but non-standard

But...

$$\begin{aligned}
\mu|Y_{0:n}, \sigma^2 &\sim \mathrm{N}\left(\mu \,\middle|\, \tfrac{1}{n}\textstyle\sum_{i=1}^n Y_i, \tfrac{\sigma^2}{n}\ \right) \\
\sigma^2|Y_{1:n}, \mu &\sim \mathrm{IG}\left(\sigma^2 \,\middle|\, \tfrac{n}{2} - 1, \tfrac{1}{2}\textstyle\sum_{i=1}^n (Y_i - \mu)^2\right)
\end{aligned}$$

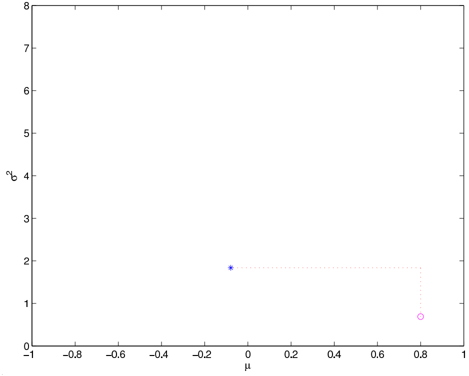
assuming constant (improper) priors on both μ and σ^2

- Hence we may use the Gibbs sampler for simulating from the posterior of (μ, σ^2)

R Gibbs Sampler for Gaussian posterior

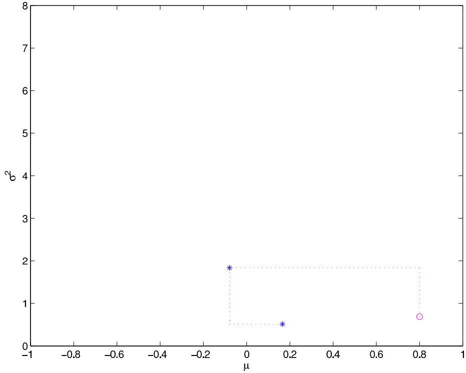
```
n = length(Y);
S = sum(Y);
mu = S/n;
for (i in 1:500)
   S2 = sum((Y-mu)^2);
   sigma2 = 1/rgamma(1,n/2-1,S2/2);
   mu = S/n + sqrt(sigma2/n)*rnorm(1);
```

Example of results with $n = 10$ observations from the N(0, 1) distribution



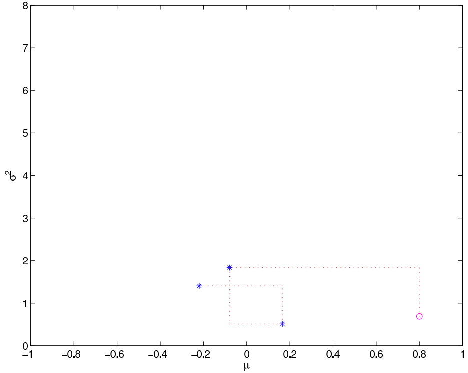
Number of Iterations 1

Example of results with $n = 10$ observations from the N(0, 1) distribution



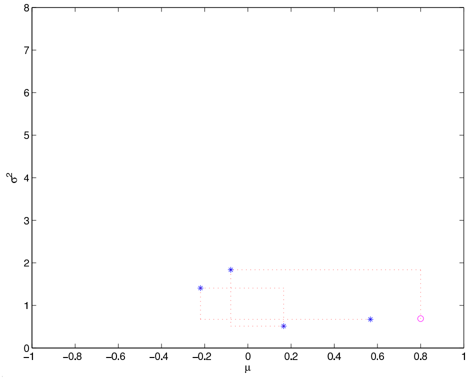
Number of Iterations 1, 2

Example of results with $n = 10$ observations from the N(0, 1) distribution



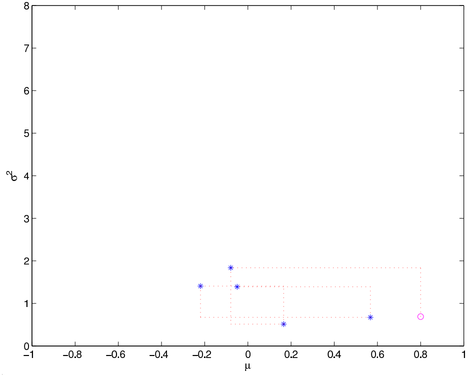
Number of Iterations 1, 2, 3

Example of results with $n = 10$ observations from the N(0, 1) distribution



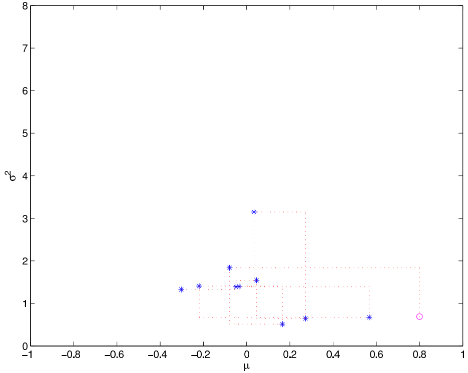
Number of Iterations 1, 2, 3, 4

Example of results with $n = 10$ observations from the N(0, 1) distribution



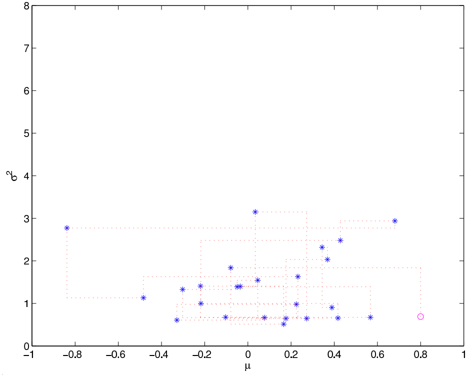
Number of Iterations 1, 2, 3, 4, 5

Example of results with $n = 10$ observations from the N(0, 1) distribution



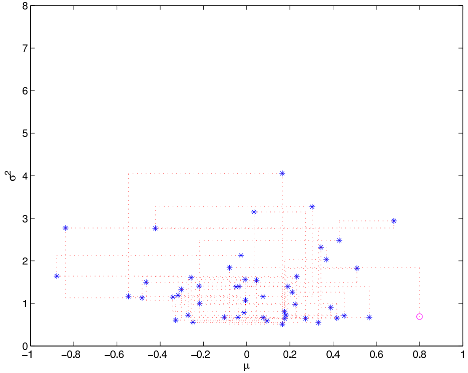
Number of Iterations 1, 2, 3, 4, 5, 10

Example of results with $n = 10$ observations from the N(0, 1) distribution



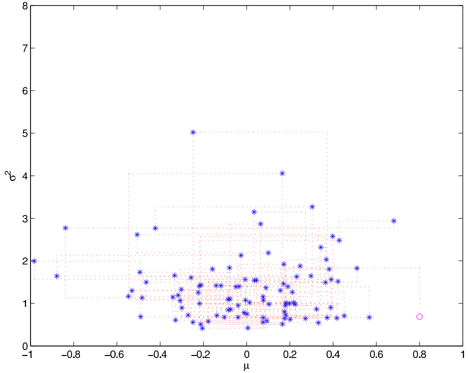
Number of Iterations 1, 2, 3, 4, 5, 10, 25

Example of results with $n = 10$ observations from the N(0, 1) distribution



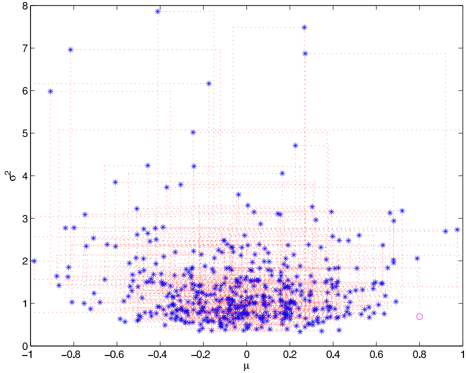
Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50

Example of results with $n = 10$ observations from the N(0, 1) distribution



Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100

Example of results with $n = 10$ observations from the N(0, 1) distribution



Number of Iterations 1, 2, 3, 4, 5, 10, 25, 50, 100, 500

Rao–Blackwellization

Conditional structure of the sampling algorithm and the dual sample,

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

should be exploited.

Rao–Blackwellization

Conditional structure of the sampling algorithm and the dual sample,

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

should be exploited.
$\mathbb{E}^\pi[g(\theta)|x]$ can be approximated as

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

Rao–Blackwellization

Conditional structure of the sampling algorithm and the dual sample,

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

should be exploited.
$\mathbb{E}^\pi[g(\theta)|x]$ can be approximated as

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

instead of

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

Rao–Black'ed density estimation

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 (Multinomial posterior)

Multinomial model

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

parametrized by μ and η, where

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

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

Example (Multinomial posterior (2))

This model stems from sampling according to

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

and aggregating some coordinates:

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

Example (Multinomial posterior (2))

This model stems from sampling according to

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

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.

Example (Multinomial posterior (3))

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).

Example (Multinomial posterior (3))

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,

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

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

An application to mixture estimation

Use of the missing data representation

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

Corresponding conditionals (Gibbs)

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

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

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

and

$$\mu_i|\mathbf{x},\mathbf{z},\sigma_i \sim \mathscr{N}(\xi_i(\mathbf{x},\mathbf{z}),\sigma_i^2/(n+\sigma_i^2)),$$
$$\sigma_i^{-2}|\mathbf{x},\mathbf{z} \sim \mathscr{G}\left(\frac{\nu_i+n_i}{2},\frac{1}{2}\left[s_i^2+\hat{s}_i^2(\mathbf{x},\mathbf{z})+\frac{n_i m_i(\mathbf{z})}{n_i+m_i(\mathbf{z})}(\bar{x}_i(\mathbf{z})-\xi_i)^2\right]\right)$$
$$p|\mathbf{x},\mathbf{z} \sim \mathscr{D}_k(\alpha_1+m_1(\mathbf{z}),\ldots,\alpha_k+m_k(\mathbf{z})),$$

Corresponding conditionals (Gibbs, 2)

where

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

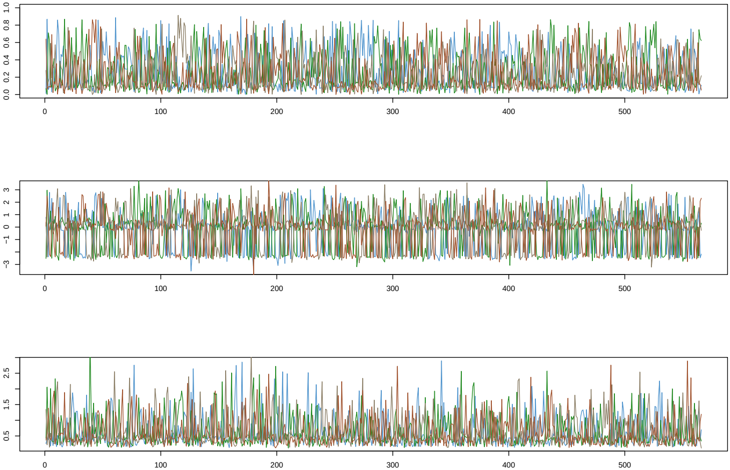
and

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

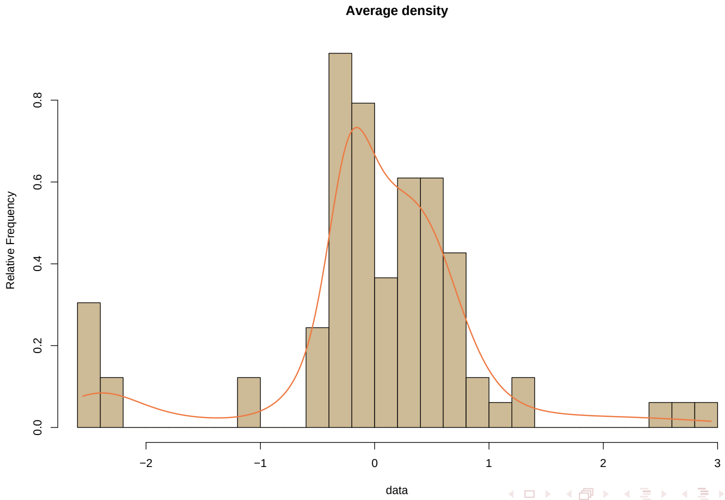
Properties

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

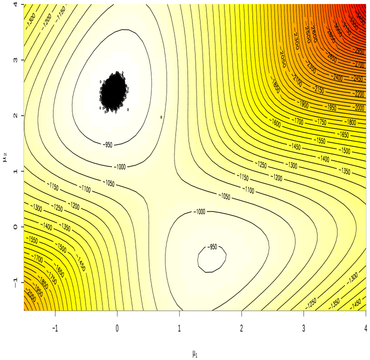
Galaxy benchmark ($k = 4$)



Galaxy benchmark ($k = 4$)

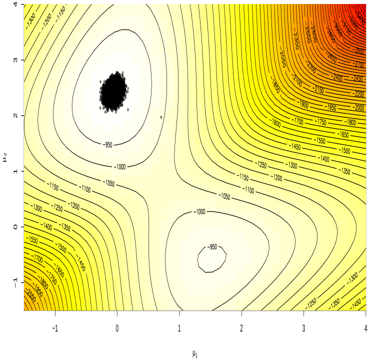


A wee problem with Gibbs on mixtures



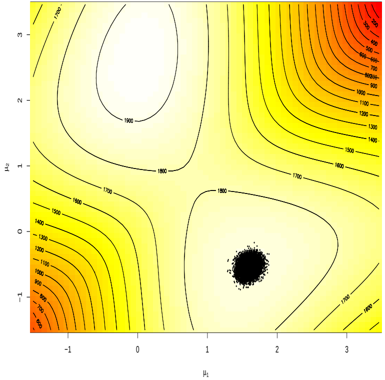
Gibbs started at random

A wee problem with Gibbs on mixtures



Gibbs started at random

Gibbs stuck at the wrong mode



[Marin, Mengersen & Robert, 2005]

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

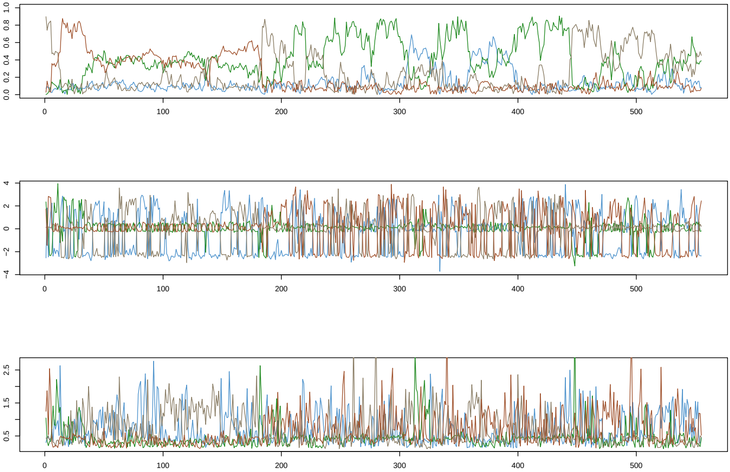
- Avoids completion
- Available (Normal vs. Cauchy vs... moves)
- Calibrated against acceptance rate
- Depends on parameterisation

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

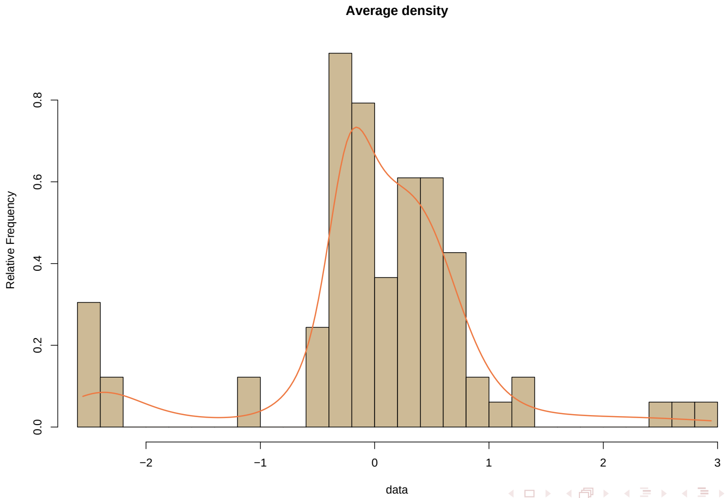
or

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

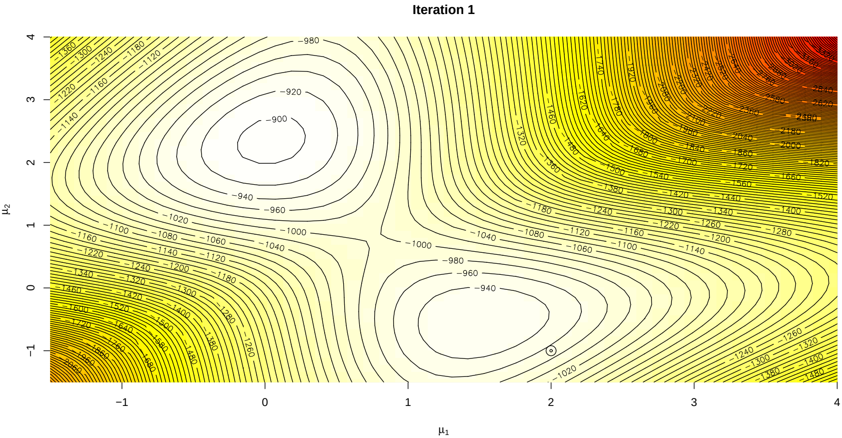
Galaxy benchmark ($k = 4$)



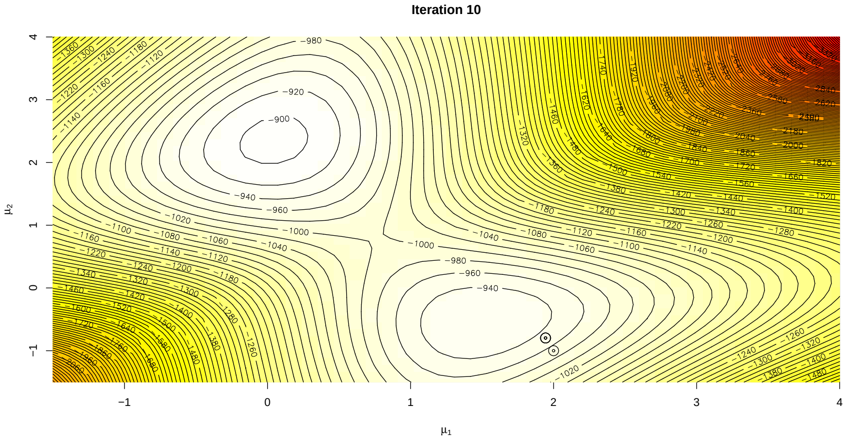
Galaxy benchmark ($k = 4$)



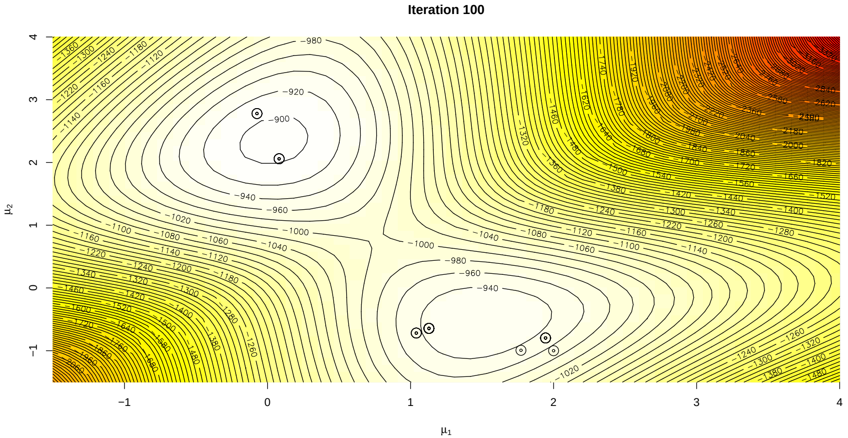
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale 1



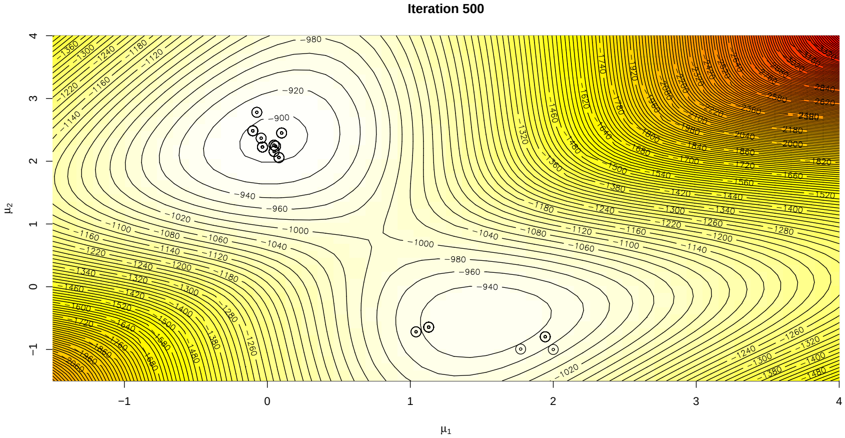
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale 1



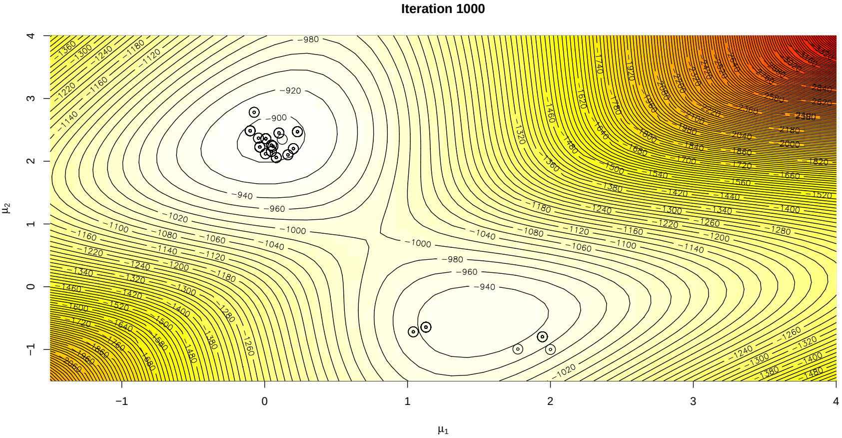
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale 1



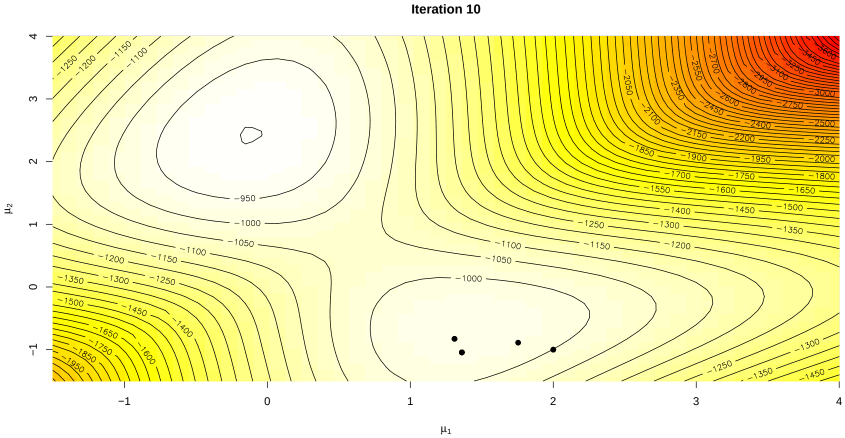
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale 1



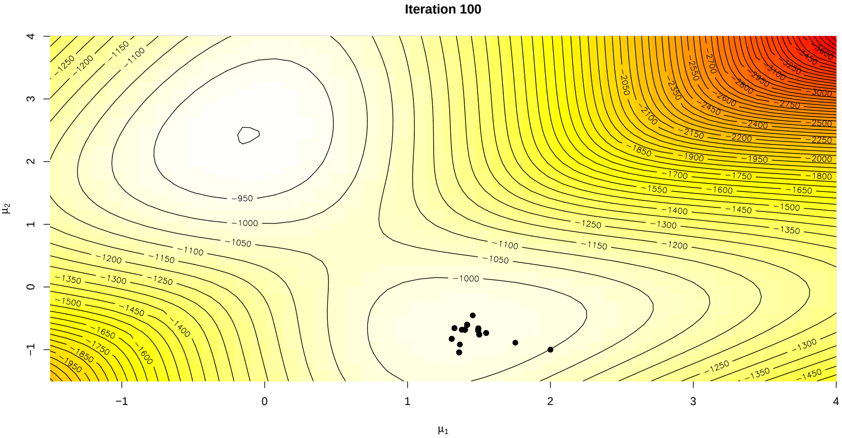
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale 1



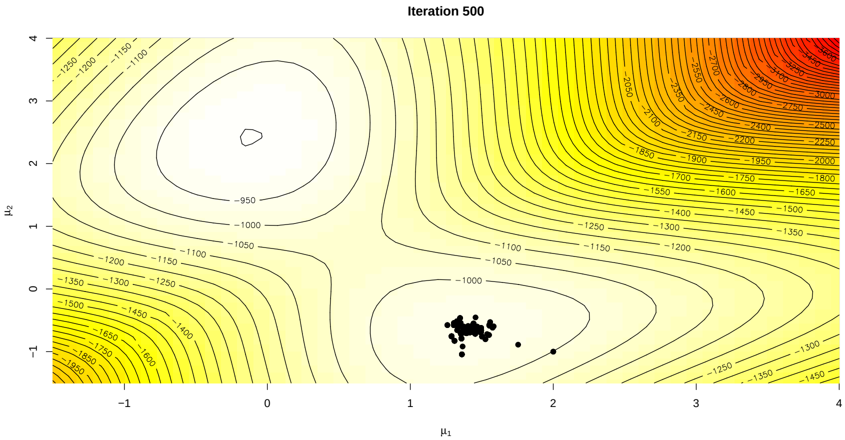
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale $\sqrt{.1}$



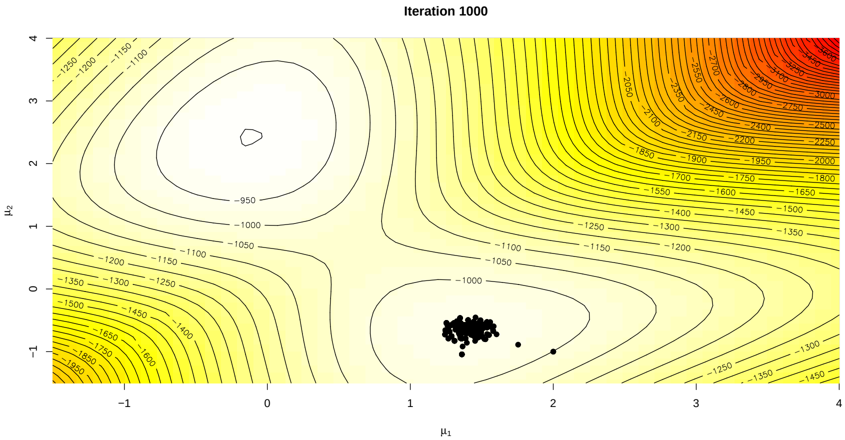
Random walk MCMC output for $.7\mathcal{N}(\mu_1,1)+.3\mathcal{N}(\mu_2,1)$ and scale $\sqrt{.1}$



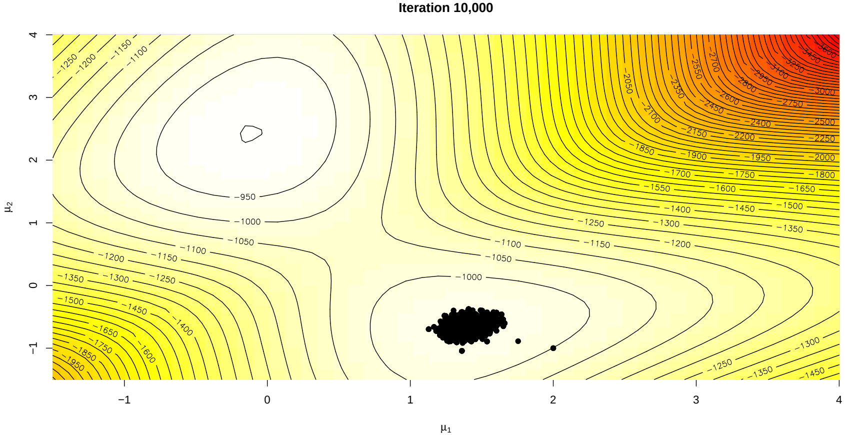
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale $\sqrt{.1}$



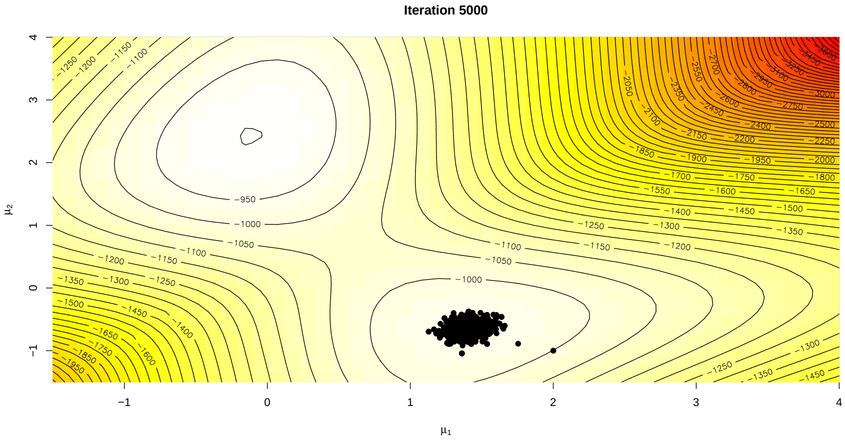
Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale $\sqrt{.1}$



Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale $\sqrt{.1}$



Random walk MCMC output for $.7\mathcal{N}(\mu_1, 1) + .3\mathcal{N}(\mu_2, 1)$ and scale $\sqrt{.1}$



Tests and model choice

Introduction

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions

Bayesian Point Estimation

Bayesian Calculations

Tests and model choice
- Bayesian tests
- Bayes factors
- Pseudo-Bayes factors

Construction of Bayes tests

Definition (Test)

Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta_0$ of a statistical model, a **test** is a statistical procedure that takes its values in $\{0, 1\}$.

Construction of Bayes tests

Definition (Test)

Given an hypothesis $H_0 : \theta \in \Theta_0$ on the parameter $\theta \in \Theta_0$ of a statistical model, a **test** is a statistical procedure that takes its values in $\{0, 1\}$.

Example (Normal mean)

For $x \sim \mathscr{N}(\theta, 1)$, decide whether or not $\theta \leq 0$.

Decision-theoretic perspective

Theorem (Optimal Bayes decision)

Under the $0-1$ loss function

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

Decision-theoretic perspective

Theorem (Optimal Bayes decision)

Under the $0-1$ loss function

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

the Bayes procedure is

$$\delta^\pi(x) = \begin{cases} 1 & \text{if } \Pr^\pi(\theta \in \Theta_0 | x) \geq a_0/(a_0 + a_1) \\ 0 & \text{otherwise} \end{cases}$$

Bound comparison

Determination of a_0/a_1 depends on consequences of "wrong decision" under both circumstances

Bound comparison

Determination of a_0/a_1 depends on consequences of "wrong decision" under both circumstances
Often difficult to assess in practice and replacement with "golden" bounds like .05, biased towards H_0

Bound comparison

Determination of a_0/a_1 depends on consequences of "wrong decision" under both circumstances

Often difficult to assess in practice and replacement with "golden" bounds like .05, biased towards H_0

Example (Binomial probability)

Consider $x \sim \mathscr{B}(n, p)$ and $\Theta_0 = [0, 1/2]$. Under the uniform prior $\pi(p) = 1$, the posterior probability of H_0 is

$$\begin{aligned} P^\pi(p \le 1/2|x) &= \frac{\int_0^{1/2} p^x (1-p)^{n-x} \mathrm{d}p}{B(x+1, n-x+1)} \\ &= \frac{(1/2)^{n+1}}{B(x+1, n-x+1)} \left\{ \frac{1}{x+1} + \ldots + \frac{(n-x)!x!}{(n+1)!} \right\} \end{aligned}$$

Loss/prior duality

Decomposition

$$
\begin{aligned}
\Pr^\pi(\theta \in \Theta_0|x) &= \int_{\Theta_0} \pi(\theta|x)\,\mathrm{d}\theta \\
&= \frac{\int_{\Theta_0} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}{\int_{\Theta} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}
\end{aligned}
$$

Loss/prior duality

Decomposition

$$\begin{aligned}
\Pr^\pi(\theta \in \Theta_0|x) &= \int_{\Theta_0} \pi(\theta|x)\,\mathrm{d}\theta \\
&= \frac{\int_{\Theta_0} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}{\int_{\Theta} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}
\end{aligned}$$

suggests representation

$$\pi(\theta) = \pi(\Theta_0)\pi_0(\theta) + (1 - \pi(\Theta_0))\pi_1(\theta)$$

Loss/prior duality

Decomposition

$$\begin{aligned}\Pr^\pi(\theta \in \Theta_0|x) &= \int_{\Theta_0} \pi(\theta|x)\,\mathrm{d}\theta \\ &= \frac{\int_{\Theta_0} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}{\int_{\Theta} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}\end{aligned}$$

suggests representation

$$\pi(\theta) = \pi(\Theta_0)\pi_0(\theta) + (1 - \pi(\Theta_0))\pi_1(\theta)$$

and decision

$$\delta^\pi(x) = 1 \text{ iff } \frac{\pi(\Theta_0)}{(1 - \pi(\Theta_0))} \frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)\,\mathrm{d}\theta}{\int_{\Theta_0^c} f(x|\theta_0)\pi_1(\theta)\,\mathrm{d}\theta} \geq \frac{a_0}{a_1}$$

Loss/prior duality

Decomposition

$$\begin{aligned}\Pr^\pi(\theta \in \Theta_0|x) &= \int_{\Theta_0} \pi(\theta|x)\,\mathrm{d}\theta \\ &= \frac{\int_{\Theta_0} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}{\int_{\Theta} f(x|\theta_0)\pi(\theta)\,\mathrm{d}\theta}\end{aligned}$$

suggests representation

$$\pi(\theta) = \pi(\Theta_0)\pi_0(\theta) + (1 - \pi(\Theta_0))\pi_1(\theta)$$

and decision

$$\delta^\pi(x) = 1 \text{ iff } \frac{\pi(\Theta_0)}{(1-\pi(\Theta_0))} \frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)\,\mathrm{d}\theta}{\int_{\Theta_0^c} f(x|\theta_0)\pi_1(\theta)\,\mathrm{d}\theta} \ge \frac{a_0}{a_1}$$

©What matters is $(\pi(\Theta_0)/a_0, (1-\pi(\Theta_0))/a_1)$

A function of posterior probabilities

Definition (Bayes factors)

For hypotheses $H_0 : \theta \in \Theta_0$ vs. $H_a : \theta \notin \Theta_0$

$$B_{01} = \frac{\pi(\Theta_0|x)}{\pi(\Theta_0^c|x)} \bigg/ \frac{\pi(\Theta_0)}{\pi(\Theta_0^c)} = \frac{\displaystyle\int_{\Theta_0} f(x|\theta)\pi_0(\theta)\mathrm{d}\theta}{\displaystyle\int_{\Theta_0^c} f(x|\theta)\pi_1(\theta)\mathrm{d}\theta}$$

[Good, 1958 & Jeffreys, 1961]

Goto Poisson example

Equivalent to Bayes rule: acceptance if
$B_{01} > \{(1 - \pi(\Theta_0))/a_1\}/\{\pi(\Theta_0)/a_0\}$

Self-contained concept

Outside decision-theoretic environment:

- eliminates choice of $\pi(\Theta_0)$

Self-contained concept

Outside decision-theoretic environment:

- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)

Self-contained concept

Outside decision-theoretic environment:

- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)
- Bayesian/marginal equivalent to the likelihood ratio

Self-contained concept

Outside decision-theoretic environment:

- eliminates choice of $\pi(\Theta_0)$
- but depends on the choice of (π_0, π_1)
- Bayesian/marginal equivalent to the likelihood ratio
- Jeffreys' scale of evidence:
  - if $\log_{10}(B_{10}^\pi)$ between 0 and 0.5, evidence against H_0 *weak*,
  - if $\log_{10}(B_{10}^\pi)$ 0.5 and 1, evidence *substantial*,
  - if $\log_{10}(B_{10}^\pi)$ 1 and 2, evidence *strong* and
  - if $\log_{10}(B_{10}^\pi)$ above 2, evidence *decisive*

Hot hand

Example (Binomial homogeneity)

Consider $H_0 : y_i \sim \mathscr{B}(n_i, p)$ $(i = 1, \ldots, G)$ vs. $H_1 : y_i \sim \mathscr{B}(n_i, p_i)$. Conjugate priors $p_i \sim \mathscr{B}e(\xi/\omega, (1-\xi)/\omega)$, with a uniform prior on $\mathbb{E}[p_i|\xi, \omega] = \xi$ and on p (ω is fixed)

Hot hand

Example (Binomial homogeneity)

Consider $H_0 : y_i \sim \mathscr{B}(n_i, p)$ $(i = 1, \ldots, G)$ vs. $H_1 : y_i \sim \mathscr{B}(n_i, p_i)$. Conjugate priors $p_i \sim \mathscr{B}e(\xi/\omega, (1-\xi)/\omega)$, with a uniform prior on $\mathbb{E}[p_i|\xi,\omega] = \xi$ and on p (ω is fixed)

$$B_{10} = \frac{\displaystyle\int_0^1 \prod_{i=1}^G \int_0^1 p_i^{y_i}(1-p_i)^{n_i-y_i} p_i^{\alpha-1}(1-p_i)^{\beta-1} \mathrm{d}\, p_i \times \Gamma(1/\omega)/[\Gamma(\xi/\omega)\Gamma((1-\xi)/\omega)] \mathrm{d}\xi}{\int_0^1 p^{\sum_i y_i}(1-p)^{\sum_i (n_i-y_i)} \mathrm{d}\, p}$$

where $\alpha = \xi/\omega$ and $\beta = (1-\xi)/\omega$.

Hot hand

Example (Binomial homogeneity)

Consider $H_0 : y_i \sim \mathscr{B}(n_i, p)$ $(i = 1, \ldots, G)$ vs. $H_1 : y_i \sim \mathscr{B}(n_i, p_i)$. Conjugate priors $p_i \sim \mathscr{B}e(\xi/\omega, (1-\xi)/\omega)$, with a uniform prior on $\mathbb{E}[p_i|\xi, \omega] = \xi$ and on p (ω is fixed)

$$B_{10} = \frac{\int_0^1 \prod_{i=1}^G \int_0^1 p_i^{y_i}(1-p_i)^{n_i-y_i} p_i^{\alpha-1}(1-p_i)^{\beta-1} \mathrm{d}\, p_i \times \Gamma(1/\omega)/[\Gamma(\xi/\omega)\Gamma((1-\xi)/\omega)] \mathrm{d}\xi}{\int_0^1 p^{\sum_i y_i}(1-p)^{\sum_i (n_i - y_i)} \mathrm{d}\, p}$$

where $\alpha = \xi/\omega$ and $\beta = (1-\xi)/\omega$.
For instance, $\log_{10}(B_{10}) = -0.79$ for $\omega = 0.005$ and $G = 138$ slightly favours H_0.

A major modification

When the null hypothesis is supported by a set of measure 0, $\pi(\Theta_0) = 0$

[End of the story?!]

A major modification

When the null hypothesis is supported by a set of measure 0, $\pi(\Theta_0) = 0$

[End of the story?!]

Requirement

Defined prior distributions under both assumptions,

$$\pi_0(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_0}(\theta), \quad \pi_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta),$$

(under the standard dominating measures on Θ_0 and Θ_1)

A major modification

When the null hypothesis is supported by a set of measure 0, $\pi(\Theta_0) = 0$

[End of the story?!]

Requirement

Defined prior distributions under both assumptions,

$$\pi_0(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_0}(\theta), \quad \pi_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta),$$

(under the standard dominating measures on Θ_0 and Θ_1)

Using the prior probabilities $\pi(\Theta_0) = \varrho_0$ and $\pi(\Theta_1) = \varrho_1$,

$$\pi(\theta) = \varrho_0 \pi_0(\theta) + \varrho_1 \pi_1(\theta).$$

Note If $\Theta_0 = \{\theta_0\}$, π_0 is the Dirac mass in θ_0

Point null hypotheses

Particular case $H_0: \ \theta = \theta_0$
Take $\rho_0 = \text{Pr}^\pi(\theta = \theta_0)$ and g_1 prior density under H_a.

Point null hypotheses

Particular case $H_0: \ \theta = \theta_0$
Take $\rho_0 = \mathrm{Pr}^\pi(\theta = \theta_0)$ and g_1 prior density under H_a.
Posterior probability of H_0

$$\pi(\Theta_0|x) = \frac{f(x|\theta_0)\rho_0}{\int f(x|\theta)\pi(\theta)\,d\theta} = \frac{f(x|\theta_0)\rho_0}{f(x|\theta_0)\rho_0 + (1-\rho_0)m_1(x)}$$

and marginal under H_a

$$m_1(x) = \int_{\Theta_1} f(x|\theta)g_1(\theta)\,d\theta.$$

Point null hypotheses (cont'd)

Dual representation

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

and

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)\rho_0}{m_1(x)(1-\rho_0)} \Bigg/ \frac{\rho_0}{1-\rho_0} = \frac{f(x|\theta_0)}{m_1(x)}$$

Point null hypotheses (cont'd)

Dual representation

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{m_1(x)}{f(x|\theta_0)}\right]^{-1}.$$

and

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)\rho_0}{m_1(x)(1-\rho_0)} \Big/ \frac{\rho_0}{1-\rho_0} = \frac{f(x|\theta_0)}{m_1(x)}$$

Connection

$$\pi(\Theta_0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\frac{1}{B_{01}^{\pi}(x)}\right]^{-1}.$$

Point null hypotheses (cont'd)

Example (Normal mean)

Test of $H_0 : \theta = 0$ when $x \sim \mathscr{N}(\theta, 1)$: we take π_1 as $\mathscr{N}(0, \tau^2)$

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

and

$$\pi(\theta = 0|x) = \left[1 + \frac{1-\rho_0}{\rho_0}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}} \exp\left(\frac{\tau^2 x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right]^{-1}$$

Point null hypotheses (cont'd)

Example (Normal mean)

Influence of τ:

τ/x	0	0.68	1.28	1.96
1	0.586	0.557	0.484	0.351
10	0.768	0.729	0.612	0.366

A fundamental difficulty

Improper priors are not allowed here

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then either π_1 or π_2 cannot be coherently normalised

A fundamental difficulty

Improper priors are not allowed here

If

$$\int_{\Theta_1} \pi_1(\mathrm{d}\theta_1) = \infty \quad \text{or} \quad \int_{\Theta_2} \pi_2(\mathrm{d}\theta_2) = \infty$$

then either π_1 or π_2 cannot be coherently normalised **but** the normalisation matters in the Bayes factor Recall Bayes factor

Constants matter

Example (Poisson versus Negative binomial)

If $\mathfrak{M}_1$ is a $\mathscr{P}(\lambda)$ distribution and $\mathfrak{M}_2$ is a $\mathscr{N}\mathscr{B}(m,p)$ distribution, we can take

$$\begin{aligned} \pi_1(\lambda) &= 1/\lambda \\ \pi_2(m,p) &= \frac{1}{M}\,\mathbb{I}_{\{1,\cdots,M\}}(m)\,\mathbb{I}_{[0,1]}(p) \end{aligned}$$

Constants matter (cont'd)

Example (Poisson versus Negative binomial (2))

then

$$
\begin{aligned}
B_{12}^{\pi} &= \frac{\displaystyle\int_0^\infty \frac{\lambda^{x-1}}{x!} e^{-\lambda} \mathrm{d}\lambda}{\displaystyle\frac{1}{M}\sum_{m=1}^{M}\int_0^\infty \binom{m}{x-1} p^x (1-p)^{m-x} dp} \\
&= 1\Bigg/\frac{1}{M}\sum_{m=x}^{M}\binom{m}{x-1}\frac{x!(m-x)!}{m!} \\
&= 1\Bigg/\frac{1}{M}\sum_{m=x}^{M} x/(m-x+1)
\end{aligned}
$$

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**
- same thing when both priors are improper

Constants matter (cont'd)

Example (Poisson versus Negative binomial (3))

- does not make sense because $\pi_1(\lambda) = 10/\lambda$ leads to a different answer, **ten times larger!**
- same thing when both priors are improper

Improper priors on common (nuisance) parameters do not matter (so much)

Normal illustration

Take $x \sim \mathscr{N}(\theta, 1)$ and $H_0 : \theta = 0$

Influence of the constant

$\pi(\theta)/x$	0.0	1.0	1.65	1.96	2.58
1	0.285	0.195	0.089	0.055	0.014
10	0.0384	0.0236	0.0101	0.00581	0.00143

Vague proper priors are not the solution

Taking a proper prior and take a “very large” variance (e.g., BUGS)

Vague proper priors are not the solution

Taking a proper prior and take a "very large" variance (e.g., BUGS) will most often result in an undefined or ill-defined limit

Vague proper priors are not the solution

Taking a proper prior and take a "very large" variance (e.g., BUGS) will most often result in an undefined or ill-defined limit

Example (Lindley's paradox)

If testing $H_0 : \theta = 0$ when observing $x \sim \mathcal{N}(\theta, 1)$, under a normal $\mathcal{N}(0, \alpha)$ prior $\pi_1(\theta)$,

$$B_{01}(x) \xrightarrow{\alpha \longrightarrow \infty} 0$$

Vague proper priors are not the solution (cont'd)

Example (Poisson versus Negative binomial (4))

$$
\begin{aligned}
B_{12} &= \frac{\displaystyle\int_0^1 \frac{\lambda^{\alpha+x-1}}{x!} e^{-\lambda\beta} \mathrm{d}\lambda}{\displaystyle\frac{1}{M}\sum_m \frac{x}{m-x+1}\,\frac{\beta^\alpha}{\Gamma(\alpha)}} \quad \text{if } \lambda \sim \mathcal{G}a(\alpha,\beta) \\
&= \frac{\Gamma(\alpha+x)}{x!\,\Gamma(\alpha)}\beta^{-x} \Big/ \frac{1}{M}\sum_m \frac{x}{m-x+1} \\
&= \frac{(x+\alpha-1)\cdots\alpha}{x(x-1)\cdots 1}\beta^{-x} \Big/ \frac{1}{M}\sum_m \frac{x}{m-x+1}
\end{aligned}
$$

Vague proper priors are not the solution (cont'd)

Example (Poisson versus Negative binomial (4))

$$B_{12} = \frac{\displaystyle\int_0^1 \frac{\lambda^{\alpha+x-1}}{x!} e^{-\lambda\beta} \mathrm{d}\lambda}{\displaystyle\frac{1}{M}\sum_m \frac{x}{m-x+1}\,\frac{\beta^\alpha}{\Gamma(\alpha)}} \quad \text{if } \lambda \sim \mathcal{G}a(\alpha,\beta)$$

$$= \frac{\Gamma(\alpha+x)}{x!\,\Gamma(\alpha)}\beta^{-x} \Big/ \frac{1}{M}\sum_m \frac{x}{m-x+1}$$

$$= \frac{(x+\alpha-1)\cdots\alpha}{x(x-1)\cdots 1}\beta^{-x} \Big/ \frac{1}{M}\sum_m \frac{x}{m-x+1}$$

depends on choice of $\alpha(\beta)$ or $\beta(\alpha) \longrightarrow 0$

Learning from the sample

Definition (Learning sample)

Given an improper prior π, $(x_1, \ldots, x_n)$ is a *learning sample* if $\pi(\cdot|x_1, \ldots, x_n)$ is proper and a *minimal learning sample* if none of its subsamples is a learning sample

Learning from the sample

Definition (Learning sample)

Given an improper prior π, $(x_1, \ldots, x_n)$ is a *learning sample* if $\pi(\cdot|x_1, \ldots, x_n)$ is proper and a *minimal learning sample* if none of its subsamples is a learning sample

There is just enough information in a minimal learning sample to make inference about θ under the prior π

Pseudo-Bayes factors

Idea

Use one part $x_{[i]}$ of the data x to make the prior proper:

Pseudo-Bayes factors

Idea

Use one part $x_{[i]}$ of the data x to make the prior proper:

- π_i improper but $\pi_i(\cdot|x_{[i]})$ proper
- and
$$\frac{\int f_i(x_{[n/i]}|\theta_i)\,\pi_i(\theta_i|x_{[i]})\mathrm{d}\theta_i}{\int f_j(x_{[n/i]}|\theta_j)\,\pi_j(\theta_j|x_{[i]})\mathrm{d}\theta_j}$$
independent of normalizing constant

Pseudo-Bayes factors

Idea

Use one part $x_{[i]}$ of the data x to make the prior proper:

- π_i improper but $\pi_i(\cdot|x_{[i]})$ proper
- and
$$\frac{\int f_i(x_{[n/i]}|\theta_i)\, \pi_i(\theta_i|x_{[i]})\mathrm{d}\theta_i}{\int f_j(x_{[n/i]}|\theta_j)\, \pi_j(\theta_j|x_{[i]})\mathrm{d}\theta_j}$$
independent of normalizing constant
- Use remaining $x_{[n/i]}$ to run test as if $\pi_j(\theta_j|x_{[i]})$ is the true prior

Motivation

- Provides a working principle for improper priors

Motivation

- Provides a working principle for improper priors
- Gather enough information from data to achieve properness
- and use this properness to run the test on remaining data

Motivation

- Provides a working principle for improper priors
- Gather enough information from data to achieve properness
- and use this properness to run the test on remaining data
- does not use x twice as in Aitkin's (1991)

Details

Since $\pi_1(\theta_1|x_{[i]}) = \dfrac{\pi_1(\theta_1) f^1_{[i]}(x_{[i]}|\theta_1)}{\int \pi_1(\theta_1) f^1_{[i]}(x_{[i]}|\theta_1) \mathrm{d}\theta_1}$

$$
\begin{aligned}
B_{12}(x_{[n/i]}) &= \frac{\int f^1_{[n/i]}(x_{[n/i]}|\theta_1)\pi_1(\theta_1|x_{[i]})\mathrm{d}\theta_1}{\int f^2_{[n/i]}(x_{[n/i]}|\theta_2)\pi_2(\theta_2|x_{[i]})\mathrm{d}\theta_2} \\
&= \frac{\int f_1(x|\theta_1)\pi_1(\theta_1)\mathrm{d}\theta_1}{\int f_2(x|\theta_2)\pi_2(\theta_2)\mathrm{d}\theta_2} \frac{\int \pi_2(\theta_2) f^2_{[i]}(x_{[i]}|\theta_2)\mathrm{d}\theta_2}{\int \pi_1(\theta_1) f^1_{[i]}(x_{[i]}|\theta_1)\mathrm{d}\theta_1} \\
&= B^N_{12}(x) B_{21}(x_{[i]})
\end{aligned}
$$

$\boxed{\text{© Independent of scaling factor!}}$

Unexpected problems!

- depends on the choice of $x_{[i]}$

Unexpected problems!

- depends on the choice of $x_{[i]}$
- many ways of combining pseudo-Bayes factors
  - $\text{AIBF} = B_{ji}^N \frac{1}{L} \sum_\ell B_{ij}(x_{[\ell]})$
  - $\text{MIBF} = B_{ji}^N \text{ med}[B_{ij}(x_{[\ell]})]$
  - $\text{GIBF} = B_{ji}^N \text{ exp} \frac{1}{L} \sum_\ell \log B_{ij}(x_{[\ell]})$
- not often an exact Bayes factor
- and thus lacking inner coherence

$$B_{12} \neq B_{10} B_{02} \qquad \text{and} \qquad B_{01} \neq 1/B_{10}\,.$$

[Berger & Pericchi, 1996]

Unexpec'd problems (cont'd)

Example (Mixtures)

There is no sample size that proper-ises improper priors, except if a training sample is allocated to *each* component

Unexpec'd problems (cont'd)

Example (Mixtures)

There is no sample size that proper-ises improper priors, except if a training sample is allocated to *each* component

Reason If

$$x_1, \ldots, x_n \sim \sum_{i=1}^{k} p_i f(x|\theta_i)$$

and

$$\pi(\theta) = \prod_i \pi_i(\theta_i) \text{ with } \int \pi_i(\theta_i) \mathrm{d}\theta_i = +\infty\,,$$

the posterior is never defined, because

$$\Pr(\text{“no observation from } f(\cdot|\theta_i)\text{”}) = (1 - p_i)^n$$

Intrinsic priors

There may exist a true prior that provides the same Bayes factor

Intrinsic priors

There may exist a true prior that provides the same Bayes factor

Example (Normal mean)

Take $x \sim \mathcal{N}(\theta, 1)$ with either $\theta = 0$ ($\mathfrak{M}_1$) or $\theta \neq 0$ ($\mathfrak{M}_2$) and $\pi_2(\theta) = 1$.
Then

$$
\begin{array}{llll}
B_{21}^{AIBF} & = B_{21} \frac{1}{\sqrt{2\pi}} \frac{1}{n} \sum_{i=1}^{n} e^{-x_1^2/2} & \approx B_{21} & \text{for } \mathcal{N}(0,2) \\
B_{21}^{MIBF} & = B_{21} \frac{1}{\sqrt{2\pi}} e^{-\text{med}(x_1^2)/2} & \approx 0.93 B_{21} & \text{for } \mathcal{N}(0,1.2)
\end{array}
$$

[Berger and Pericchi, 1998]

Intrinsic priors

There may exist a true prior that provides the same Bayes factor

Example (Normal mean)

Take $x \sim \mathcal{N}(\theta, 1)$ with either $\theta = 0$ ($\mathfrak{M}_1$) or $\theta \neq 0$ ($\mathfrak{M}_2$) and $\pi_2(\theta) = 1$.
Then

$$
\begin{array}{llll}
B_{21}^{AIBF} & = B_{21} \frac{1}{\sqrt{2\pi}} \frac{1}{n} \sum_{i=1}^{n} e^{-x_1^2/2} & \approx B_{21} & \text{for } \mathcal{N}(0,2) \\
B_{21}^{MIBF} & = B_{21} \frac{1}{\sqrt{2\pi}} e^{-\text{med}(x_1^2)/2} & \approx 0.93 B_{21} & \text{for } \mathcal{N}(0,1.2)
\end{array}
$$

[Berger and Pericchi, 1998]

When such a prior exists, it is called an **intrinsic prior**

Intrinsic priors (cont'd)

Intrinsic priors (cont'd)

Example (Exponential scale)

Take $$x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \exp(\theta - x)\mathbb{I}_{x \geq \theta}$$
and $H_0 : \theta = \theta_0,\ H_1 : \theta > \theta_0$, with $\pi_1(\theta) = 1$
Then
$$B_{10}^A = B_{10}(x) \frac{1}{n} \sum_{i=1}^n \left[e^{x_i - \theta_0} - 1 \right]^{-1}$$
is the Bayes factor for
$$\pi_2(\theta) = e^{\theta_0 - \theta} \left\{ 1 - \log\left(1 - e^{\theta_0 - \theta} \right) \right\}$$

Intrinsic priors (cont'd)

Example (Exponential scale)

Take $$x_1, \ldots, x_n \stackrel{\text{i.i.d.}}{\sim} \exp(\theta - x)\mathbb{I}_{x \geq \theta}$$
and $H_0 : \theta = \theta_0,\ H_1 : \theta > \theta_0$, with $\pi_1(\theta) = 1$
Then

$$B_{10}^A = B_{10}(x) \frac{1}{n} \sum_{i=1}^{n} \left[e^{x_i - \theta_0} - 1 \right]^{-1}$$

is the Bayes factor for

$$\pi_2(\theta) = e^{\theta_0 - \theta} \left\{ 1 - \log\left(1 - e^{\theta_0 - \theta}\right) \right\}$$

Most often, however, the pseudo-Bayes factors do not correspond to any true Bayes factor

[Berger and Pericchi, 2001]

Fractional Bayes factor

Idea

use directly the likelihood to separate training sample from testing sample

$$B_{12}^F = B_{12}(x) \frac{\displaystyle\int L_2^b(\theta_2)\pi_2(\theta_2)d\theta_2}{\displaystyle\int L_1^b(\theta_1)\pi_1(\theta_1)d\theta_1}$$

[O'Hagan, 1995]

Fractional Bayes factor

Idea

use directly the likelihood to separate training sample from testing sample

$$B_{12}^F = B_{12}(x) \frac{\int L_2^b(\theta_2)\pi_2(\theta_2)d\theta_2}{\int L_1^b(\theta_1)\pi_1(\theta_1)d\theta_1}$$

[O'Hagan, 1995]

Proportion b of the sample used to gain proper-ness

Fractional Bayes factor (cont'd)

Example (Normal mean)

$$B_{12}^F = \frac{1}{\sqrt{b}}\, e^{n(b-1)\bar{x}_n^2/2}$$

corresponds to exact Bayes factor for the prior $\mathcal{N}\left(0, \frac{1-b}{nb}\right)$

- If b constant, prior variance goes to 0
- If $b = \dfrac{1}{n}$, prior variance stabilises around 1
- If $b = n^{-\alpha}$, $\alpha < 1$, prior variance goes to 0 too.

Comparison with classical tests

Standard answer

Definition (p-value)

The p*-value* $p(x)$ associated with a test is the largest significance level for which H_0 is rejected

Comparison with classical tests

Standard answer

Definition (p-value)

The *p-value* $p(x)$ associated with a test is the largest significance level for which H_0 is rejected

Note

An alternative definition is that a p-value is distributed uniformly under the null hypothesis.

p-value

Example (Normal mean)

Since the UUMP test is $\{|x| > k\}$, standard p-value

$$\begin{aligned} p(x) &= \inf\{\alpha;\ |x| > k_\alpha\} \\ &= P^X(|X| > |x|), \qquad X \sim \mathcal{N}(0,1) \\ &= 1 - \Phi(|x|) + \Phi(|x|) = 2[1 - \Phi(|x|)]. \end{aligned}$$

Thus, if $x = 1.68$, $p(x) = 0.10$ and, if $x = 1.96$, $p(x) = 0.05$.

Problems with p-values

- Evaluation of the **wrong** quantity, namely the probability to exceed the observed quantity.(wrong conditionin)
- No transfer of the UMP optimality
- No decisional support (occurences of inadmissibility)
- Evaluation only under the null hypothesis
- Huge numerical difference with the Bayesian range of answers

Bayesian lower bounds

For illustration purposes, consider a class $\mathscr{G}$ of prior distributions

$$
\begin{aligned}
B(x,\mathscr{G}) &= \inf_{g\in\mathscr{G}} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta)g(\theta)\,d\theta},\\
P(x,\mathscr{G}) &= \inf_{g\in\mathscr{G}} \frac{f(x|\theta_0)}{f(x|\theta_0)+\int_\Theta f(x|\theta)g(\theta)\,d\theta}
\end{aligned}
$$

when $\varrho_0 = 1/2$ or

$$
B(x,\mathscr{G}) = \frac{f(x|\theta_0)}{\sup_{g\in\mathscr{G}} \int_\Theta f(x|\theta)g(\theta)d\theta}, \quad P(x,\mathscr{G}) = \left[1+\frac{1}{(x,\mathscr{G})}\right]^{-1}.
$$

Resolution

Lemma

If there exists a mle for θ, $\hat{\theta}(x)$, the solutions to the Bayesian lower bounds are

$$B(x,\mathscr{G}) = \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}, \quad P(x,\mathscr{G}) = \left[1 + \frac{f(x|\hat{\theta}(x))}{f(x|\theta_0)}\right]^{-1}$$

respectively

Normal case

When $x \sim \mathscr{N}(\theta, 1)$ and $H_0 : \theta_0 = 0$, the lower bounds are

$$(x, G_A) = e^{-x^2/2} \quad \text{et} \quad (x, G_A) = \left(1 + e^{x^2/2}\right)^{-1},$$

Normal case

When $x \sim \mathscr{N}(\theta, 1)$ and $H_0 : \theta_0 = 0$, the lower bounds are

$$(x, G_A) = e^{-x^2/2} \quad \text{et} \quad (x, G_A) = \left(1 + e^{x^2/2}\right)^{-1},$$

i.e.

p-value	0.10	0.05	0.01	0.001
P	0.205	0.128	0.035	0.004
B	0.256	0.146	0.036	0.004

Normal case

When $x \sim \mathscr{N}(\theta, 1)$ and $H_0 : \theta_0 = 0$, the lower bounds are

$$(x, G_A) = e^{-x^2/2} \quad \text{et} \quad (x, G_A) = \left(1 + e^{x^2/2}\right)^{-1},$$

i.e.

p-value	0.10	0.05	0.01	0.001
P	0.205	0.128	0.035	0.004
B	0.256	0.146	0.036	0.004

[Quite different!]

Unilateral case

Different situation when $H_0 : \theta \leq 0$

Unilateral case

Different situation when $H_0 : \theta \leq 0$

- Single prior can be used both for H_0 and H_a

Unilateral case

Different situation when $H_0 : \theta \leq 0$

- Single prior can be used both for H_0 and H_a
- Improper priors are therefore acceptable

Unilateral case

Different situation when $H_0 : \theta \leq 0$

- Single prior can be used both for H_0 and H_a
- Improper priors are therefore acceptable
- Similar numerical values compared with p-values

Unilateral agreement

Theorem

When $x \sim f(x-\theta)$, with f symmetric around 0 and endowed with the monotone likelihood ratio property, if $H_0: \ \theta \le 0$, the p-value $p(x)$ is equal to the lower bound of the posterior probabilities, $P(x, \mathscr{G}_{SU})$, when $\mathscr{G}_{SU}$ is the set of symmetric unimodal priors and when $x > 0$.

Unilateral agreement

Theorem

When $x \sim f(x-\theta)$, with f symmetric around 0 and endowed with the monotone likelihood ratio property, if $H_0 : \ \theta \leq 0$, the p-value $p(x)$ is equal to the lower bound of the posterior probabilities, $P(x, \mathscr{G}_{SU})$, when $\mathscr{G}_{SU}$ is the set of symmetric unimodal priors and when $x > 0$.

Reason:

$$p(x) = P_{\theta=0}(X > x) = \int_x^{+\infty} f(t)\,\mathrm{d}t = \inf_K \frac{1}{1 + \left[\frac{\int_{-K}^{0} f(x-\theta)\,\mathrm{d}\theta}{\int_{-K}^{K} f(x-\theta)}\,\mathrm{d}\theta\right]^{-1}}$$

Cauchy example

When $x \sim \mathscr{C}(\theta, 1)$ and $H_0 : \theta \leq 0$, lower bound inferior to p-value:

p-value	0.437	0.102	0.063	0.013	0.004
$\underline{P}$	0.429	0.077	0.044	0.007	0.002

Model choice and model comparison

Choice of models

Several models available for the same observation

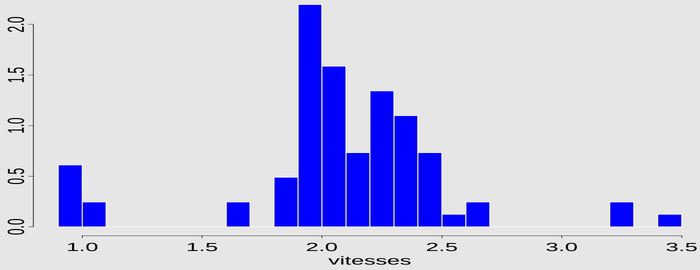
$$\mathfrak{M}_i : x \sim f_i(x|\theta_i), \qquad i \in \mathfrak{I}$$

where $\mathfrak{I}$ can be finite or infinite

Example (Galaxy normal mixture)

Set of observations of radial speeds of 82 galaxies possibly modelled as a mixture of normal distributions

$$\mathfrak{M}_i : x_j \sim \sum_{\ell=1}^{i} p_{\ell i} \mathcal{N}(\mu_{\ell i}, \sigma_{\ell i}^2)$$



Bayesian resolution

B Framework

Probabilises the entire model/parameter space

Bayesian resolution

B Framework

Probabilises the entire model/parameter space

This means:

- allocating probabilities p_i to all models $\mathfrak{M}_i$
- defining priors $\pi_i(\theta_i)$ for each parameter space Θ_i

Formal solutions

Resolution

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \displaystyle\int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathtt{d}\theta_i}{\displaystyle\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathtt{d}\theta_j}$$

Formal solutions

Resolution

1. Compute

$$p(\mathfrak{M}_i|x) = \frac{p_i \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i)\mathrm{d}\theta_i}{\sum_j p_j \int_{\Theta_j} f_j(x|\theta_j)\pi_j(\theta_j)\mathrm{d}\theta_j}$$

2. Take largest $p(\mathfrak{M}_i|x)$ to determine ''best'' model, or use averaged predictive

$$\sum_j p(\mathfrak{M}_j|x) \int_{\Theta_j} f_j(x'|\theta_j)\pi_j(\theta_j|x)\mathrm{d}\theta_j$$

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences
  - representation of parsimony/sparcity (Ockham's rule)
  - how to fight overfitting for nested models

Several types of problems

- Concentrate on selection perspective:
  - averaging = estimation = non-parsimonious = no-decision
  - how to integrate loss function/decision/consequences
  - representation of parsimony/sparcity (Ockham's rule)
  - how to fight overfitting for nested models

Which loss ?

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$,

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)
    - $\pi_i(\theta_i)$ coherent (?) for nested models

Several types of problems (2)

- Choice of prior structures
  - adequate weights p_i:
    if $\mathfrak{M}_1 = \mathfrak{M}_2 \cup \mathfrak{M}_3$, $p(\mathfrak{M}_1) = p(\mathfrak{M}_2) + p(\mathfrak{M}_3)$?
  - priors distributions
    - $\pi_i(\theta_i)$ defined for every $i \in \mathfrak{I}$
    - $\pi_i(\theta_i)$ *proper* (Jeffreys)
    - $\pi_i(\theta_i)$ coherent (?) for nested models

Warning

Parameters common to several models must be treated as separate entities!

Several types of problems (3)

- Computation of predictives and marginals
  - infinite dimensional spaces
  - integration over parameter spaces
  - integration over different spaces
  - summation over many models (2^k)

Compatibility principle

Difficulty of finding simultaneously priors on a collection of models $\mathfrak{M}_i$ $(i \in \mathfrak{I})$

Compatibility principle

Difficulty of finding simultaneously priors on a collection of models $\mathfrak{M}_i$ $(i \in \mathfrak{I})$

Easier to start from a single prior on a “big” model and to derive the others from a coherence principle

[Dawid & Lauritzen, 2000]

Projection approach

For $\mathfrak{M}_2$ submodel of $\mathfrak{M}_1$, π_2 can be derived as the distribution of $\theta_2^{\perp}(\theta_1)$ when $\theta_1 \sim \pi_1(\theta_1)$ and $\theta_2^{\perp}(\theta_1)$ is a projection of θ_1 on $\mathfrak{M}_2$, e.g.

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^{\perp})) = \inf_{\theta_2 \in \Theta_2} d(f(\cdot\,|\theta_1)\,, f(\cdot\,|\theta_2))\,.$$

where d is a divergence measure

[McCulloch & Rossi, 1992]

Projection approach

For $\mathfrak{M}_2$ submodel of $\mathfrak{M}_1$, π_2 can be derived as the distribution of $\theta_2^{\perp}(\theta_1)$ when $\theta_1 \sim \pi_1(\theta_1)$ and $\theta_2^{\perp}(\theta_1)$ is a projection of θ_1 on $\mathfrak{M}_2$, e.g.

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^{\perp})) = \inf_{\theta_2\in\Theta_2} d(f(\cdot\,|\theta_1)\,, f(\cdot\,|\theta_2))\,.$$

where d is a divergence measure

[McCulloch & Rossi, 1992]

Or we can look instead at the posterior distribution of

$$d(f(\cdot\,|\theta_1), f(\cdot\,|{\theta_1}^{\perp}))$$

[Goutis & Robert, 1998]

Operational principle for variable selection

Selection rule

Among all subsets $\mathcal{A}$ of covariates such that

$$d(\mathfrak{M}_g, \mathfrak{M}_{\mathcal{A}}) = \mathbb{E}_x[d(f_g(\cdot|x, \alpha), f_{\mathcal{A}}(\cdot|x_{\mathcal{A}}, \alpha^{\perp}))] < \epsilon$$

select the submodel with the smallest number of variables.

[Dupuis & Robert, 2001]

Kullback proximity

Alternative to above

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1

Kullback proximity

Alternative to above

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1 when it achieves the minimum Kullback divergence between the corresponding marginals:
$m_1(x; \pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\mathrm{d}\theta$ and
$m_2(x); \pi_2 = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)\mathrm{d}\theta$,

Kullback proximity

Alternative to above

Definition (Compatible prior)

Given a prior π_1 on a model $\mathfrak{M}_1$ and a submodel $\mathfrak{M}_2$, a prior π_2 on $\mathfrak{M}_2$ is *compatible* with π_1 when it achieves the minimum Kullback divergence between the corresponding marginals:
$m_1(x;\pi_1) = \int_{\Theta_1} f_1(x|\theta)\pi_1(\theta)\mathrm{d}\theta$ and
$m_2(x);\pi_2 = \int_{\Theta_2} f_2(x|\theta)\pi_2(\theta)\mathrm{d}\theta$,

$$\pi_2 = \arg\min_{\pi_2} \int \log\left(\frac{m_1(x;\pi_1)}{m_2(x;\pi_2)}\right) m_1(x;\pi_1)\,\mathrm{d}x$$

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1
- Prohibits the use of improper priors

Difficulties

- Does not give a working principle when $\mathfrak{M}_2$ is not a submodel $\mathfrak{M}_1$
- Depends on the choice of π_1
- Prohibits the use of improper priors
- Worse: useless in unconstrained settings...

Case of exponential families

Models

$$\mathfrak{M}_1 : \{f_1(x|\theta), \theta \in \Theta\}$$

and

$$\mathfrak{M}_2 : \{f_2(x|\lambda), \lambda \in \Lambda\}$$

sub-model of $\mathcal{M}_1$,

$$\forall \lambda \in \Lambda, \exists\, \theta(\lambda) \in \Theta, f_2(x|\lambda) = f_1(x|\theta(\lambda))$$

Case of exponential families

Models

$$\mathfrak{M}_1 : \{f_1(x|\theta), \theta \in \Theta\}$$

and

$$\mathfrak{M}_2 : \{f_2(x|\lambda), \lambda \in \Lambda\}$$

sub-model of $\mathcal{M}_1$,

$$\forall \lambda \in \Lambda, \exists\, \theta(\lambda) \in \Theta, f_2(x|\lambda) = f_1(x|\theta(\lambda))$$

Both $\mathfrak{M}_1$ and $\mathfrak{M}_2$ are natural exponential families

$$\begin{aligned} f_1(x|\theta) &= h_1(x) \exp(\theta^{\mathsf{T}} t_1(x) - M_1(\theta)) \\ f_2(x|\lambda) &= h_2(x) \exp(\lambda^{\mathsf{T}} t_2(x) - M_2(\lambda)) \end{aligned}$$

Conjugate priors

Parameterised (conjugate) priors

$$\begin{aligned}\pi_1(\theta; s_1, n_1) &= C_1(s_1, n_1)\exp(s_1^{\mathsf{T}}\theta - n_1 M_1(\theta)) \\ \pi_2(\lambda; s_2, n_2) &= C_2(s_2, n_2)\exp(s_2^{\mathsf{T}}\lambda - n_2 M_2(\lambda))\end{aligned}$$

Conjugate priors

Parameterised (conjugate) priors

$$\begin{aligned} \pi_1(\theta; s_1, n_1) &= C_1(s_1, n_1) \exp(s_1^\mathsf{T}\theta - n_1 M_1(\theta)) \\ \pi_2(\lambda; s_2, n_2) &= C_2(s_2, n_2) \exp(s_2^\mathsf{T}\lambda - n_2 M_2(\lambda)) \end{aligned}$$

with closed form marginals ($i = 1, 2$)

$$m_i(x; s_i, n_i) = \int f_i(x|u)\pi_i(u)du = \frac{h_i(x)C_i(s_i, n_i)}{C_i(s_i + t_i(x), n_i + 1)}$$

Conjugate compatible priors

(Q.) Existence and unicity of Kullback-Leibler projection

$$
\begin{aligned}
(s_2^*, n_2^*) &= \arg\min_{(s_2,n_2)} \mathfrak{KL}(m_1(\cdot; s_1, n_1), m_2(\cdot; s_2, n_2)) \\
&= \arg\min_{(s_2,n_2)} \int \log\left(\frac{m_1(x; s_1, n_1)}{m_2(x; s_2, n_2)}\right) m_1(x; s_1, n_1)\mathrm{d}x
\end{aligned}
$$

A sufficient condition

Sufficient statistic $\psi = (\lambda, -M_2(\lambda))$

Theorem (Existence)

If, for all (s_2, n_2), *the matrix*

$$\mathbb{V}^{\pi_2}_{s_2,n_2}[\psi] - \mathbb{E}^{m_1}_{s_1,n_1}\left[\mathbb{V}^{\pi_2}_{s_2,n_2}(\psi|x)\right]$$

is semi-definite negative,

A sufficient condition

Sufficient statistic $\psi = (\lambda, -M_2(\lambda))$

Theorem (Existence)

If, for all (s_2, n_2)*, the matrix*

$$\mathbb{V}^{\pi_2}_{s_2,n_2}[\psi] - \mathbb{E}^{m_1}_{s_1,n_1}\left[\mathbb{V}^{\pi_2}_{s_2,n_2}(\psi|x)\right]$$

is semi-definite negative, the conjugate compatible prior exists, is unique and satisfies

$$\begin{aligned}
\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}[\lambda] - \mathbb{E}^{m_1}_{s_1,n_1}[\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(\lambda|x)] &= 0 \\
\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(M_2(\lambda)) - \mathbb{E}^{m_1}_{s_1,n_1}[\mathbb{E}^{\pi_2}_{s_2^*,n_2^*}(M_2(\lambda)|x)] &= 0.
\end{aligned}$$

An application to linear regression

$\mathfrak{M}_1$ and $\mathfrak{M}_2$ are two nested Gaussian linear regression models with Zellner's g-priors and the same variance $\sigma^2 \sim \pi(\sigma^2)$:

1. $\mathfrak{M}_1$:

$$y|\beta_1, \sigma^2 \sim \mathcal{N}(X_1\beta_1, \sigma^2), \quad \beta_1|\sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1^\mathsf{T} X_1)^{-1}\right)$$

where X_1 is a $(n \times k_1)$ matrix of rank $k_1 \leq n$

An application to linear regression

$\mathfrak{M}_1$ and $\mathfrak{M}_2$ are two nested Gaussian linear regression models with Zellner's g-priors and the same variance $\sigma^2 \sim \pi(\sigma^2)$:

1. $\mathfrak{M}_1$:

$$y|\beta_1, \sigma^2 \sim \mathcal{N}(X_1\beta_1, \sigma^2), \quad \beta_1|\sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1^\mathsf{T} X_1)^{-1}\right)$$

   where X_1 is a $(n \times k_1)$ matrix of rank $k_1 \leq n$

2. $\mathfrak{M}_2$:

$$y|\beta_2, \sigma^2 \sim \mathcal{N}(X_2\beta_2, \sigma^2), \quad \beta_2|\sigma^2 \sim \mathcal{N}\left(s_2, \sigma^2 n_2 (X_2^\mathsf{T} X_2)^{-1}\right),$$

   where X_2 is a $(n \times k_2)$ matrix with $\text{span}(X_2) \subseteq \text{span}(X_1)$

An application to linear regression

$\mathfrak{M}_1$ and $\mathfrak{M}_2$ are two nested Gaussian linear regression models with Zellner's g-priors and the same variance $\sigma^2 \sim \pi(\sigma^2)$:

1. $\mathfrak{M}_1$:

   $$y|\beta_1, \sigma^2 \sim \mathcal{N}(X_1\beta_1, \sigma^2), \quad \beta_1|\sigma^2 \sim \mathcal{N}\left(s_1, \sigma^2 n_1 (X_1^{\mathsf{T}} X_1)^{-1}\right)$$

   where X_1 is a $(n \times k_1)$ matrix of rank $k_1 \leq n$

2. $\mathfrak{M}_2$:

   $$y|\beta_2, \sigma^2 \sim \mathcal{N}(X_2\beta_2, \sigma^2), \quad \beta_2|\sigma^2 \sim \mathcal{N}\left(s_2, \sigma^2 n_2 (X_2^{\mathsf{T}} X_2)^{-1}\right),$$

   where X_2 is a $(n \times k_2)$ matrix with $\text{span}(X_2) \subseteq \text{span}(X_1)$

For a fixed (s_1, n_1), we need the projection $(s_2, n_2) = (s_1, n_1)^{\perp}$

Compatible g-priors

Since σ^2 is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on σ^2: $m_1(y|\sigma^2; s_1, n_1)$ and $m_2(y|\sigma^2; s_2, n_2)$

Compatible g-priors

Since σ^2 is a nuisance parameter, we can minimize the Kullback-Leibler divergence between the two marginal distributions conditional on σ^2: $m_1(y|\sigma^2; s_1, n_1)$ and $m_2(y|\sigma^2; s_2, n_2)$

Theorem

Conditional on σ^2, the conjugate compatible prior of $\mathfrak{M}_2$ wrt $\mathfrak{M}_1$ is

$$\beta_2|X_2, \sigma^2 \sim \mathcal{N}\left(s_2^*, \sigma^2 n_2^* (X_2^T X_2)^{-1}\right)$$

with

$$\begin{aligned} s_2^* &= (X_2^T X_2)^{-1} X_2^T X_1 s_1 \\ n_2^* &= n_1 \end{aligned}$$

Variable selection

Regression setup where y regressed on a set $\{x_1, \ldots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Variable selection

Regression setup where y regressed on a set $\{x_1, \ldots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Corresponding 2^p submodels $\mathfrak{M}_\gamma$, where $\gamma \in \Gamma = \{0,1\}^p$ indicates inclusion/exclusion of variables by a binary representation,

Variable selection

Regression setup where y regressed on a set $\{x_1, \ldots, x_p\}$ of p **potential explanatory** regressors (plus intercept)

Corresponding 2^p submodels $\mathfrak{M}_\gamma$, where $\gamma \in \Gamma = \{0, 1\}^p$ indicates inclusion/exclusion of variables by a binary representation,
e.g. $\gamma = 101001011$ means that x_1, x_3, x_5, x_7 and x_8 are included.

Notations

For model $\mathfrak{M}_\gamma$,

- q_γ variables included
- $t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,q_\gamma}(\gamma)\}$ indices of those variables and $t_0(\gamma)$ indices of the variables *not* included
- For $\beta \in \mathbb{R}^{p+1}$,

$$\begin{aligned} \beta_{t_1(\gamma)} &= \left[\beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q_\gamma}(\gamma)}\right] \\ X_{t_1(\gamma)} &= \left[1_n | x_{t_{1,1}(\gamma)} | \ldots | x_{t_{1,q_\gamma}(\gamma)}\right]. \end{aligned}$$

Notations

For model $\mathfrak{M}_\gamma$,

- q_γ variables included
- $t_1(\gamma) = \{t_{1,1}(\gamma), \ldots, t_{1,q_\gamma}(\gamma)\}$ indices of those variables and $t_0(\gamma)$ indices of the variables *not* included
- For $\beta \in \mathbb{R}^{p+1}$,

$$\begin{aligned}
\beta_{t_1(\gamma)} &= \left[\beta_0, \beta_{t_{1,1}(\gamma)}, \ldots, \beta_{t_{1,q_\gamma}(\gamma)}\right] \\
X_{t_1(\gamma)} &= \left[1_n | x_{t_{1,1}(\gamma)} | \ldots | x_{t_{1,q_\gamma}(\gamma)}\right].
\end{aligned}$$

Submodel $\mathfrak{M}_\gamma$ is thus

$$y|\beta, \gamma, \sigma^2 \sim \mathcal{N}\left(X_{t_1(\gamma)}\beta_{t_1(\gamma)}, \sigma^2 I_n\right)$$

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

Noninformative g

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}^\mathsf{T} X\tilde{\beta}, c\sigma^2 \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1}\right)$$

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}^\mathsf{T} X\tilde{\beta}, c\sigma^2 \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1}\right)$$

[Surprise!]

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

Noninformative g

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T}X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}^\mathsf{T} X\tilde{\beta}, c\sigma^2 \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1}\right)$$

Global and compatible priors

Use Zellner's g-prior, i.e. a normal prior for β conditional on σ^2,

$$\beta|\sigma^2 \sim \mathcal{N}(\tilde{\beta}, c\sigma^2(X^\mathsf{T} X)^{-1})$$

and a Jeffreys prior for σ^2,

$$\pi(\sigma^2) \propto \sigma^{-2}$$

Noninformative g

Resulting compatible prior

$$\mathcal{N}\left(\left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1} X_{t_1(\gamma)}^\mathsf{T} X\tilde{\beta}, c\sigma^2 \left(X_{t_1(\gamma)}^\mathsf{T} X_{t_1(\gamma)}\right)^{-1}\right)$$

[Surprise!]

Model index

For the hierarchical parameter γ, we use

$$\pi(\gamma) = \prod_{i=1}^{p} \tau_i^{\gamma_i} (1 - \tau_i)^{1-\gamma_i},$$

where τ_i corresponds to the prior probability that variable i is present in the model (and a priori independence between the presence/absence of variables)

Model index

For the hierarchical parameter γ, we use

$$\pi(\gamma) = \prod_{i=1}^{p} \tau_i^{\gamma_i}(1-\tau_i)^{1-\gamma_i},$$

where τ_i corresponds to the prior probability that variable i is present in the model (and a priori independence between the presence/absence of variables)
Typically, when no prior information is available, $\tau_1 = \ldots = \tau_p = 1/2$, ie a uniform prior

$$\pi(\gamma) = 2^{-p}$$

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q_\gamma+1)/2}\left[y^\mathsf{T}y - \frac{cy^\mathsf{T}P_1y}{c+1} + \frac{\tilde\beta^\mathsf{T}X^\mathsf{T}P_1X\tilde\beta}{c+1} - \frac{2y^\mathsf{T}P_1X\tilde\beta}{c+1}\right]^{-n/2}.$$

Posterior model probability

Can be obtained in closed form:

$$\pi(\gamma|y) \propto (c+1)^{-(q_\gamma+1)/2}\left[y^\mathsf{T}y - \frac{cy^\mathsf{T}P_1y}{c+1} + \frac{\tilde\beta^\mathsf{T}X^\mathsf{T}P_1X\tilde\beta}{c+1} - \frac{2y^\mathsf{T}P_1X\tilde\beta}{c+1}\right]^{-n/2}.$$

Conditionally on γ, posterior distributions of β and σ^2:

$$\begin{aligned}
\beta_{t_1(\gamma)}|\sigma^2, y, \gamma &\sim \mathcal{N}\left[\frac{c}{c+1}(U_1y + U_1X\tilde\beta/c), \frac{\sigma^2 c}{c+1}\left(X^\mathsf{T}_{t_1(\gamma)}X_{t_1(\gamma)}\right)^{-1}\right],\\
\sigma^2|y,\gamma &\sim \mathcal{IG}\left[\frac{n}{2}, \frac{y^\mathsf{T}y}{2} - \frac{cy^\mathsf{T}P_1y}{2(c+1)} + \frac{\tilde\beta^\mathsf{T}X^\mathsf{T}P_1X\tilde\beta}{2(c+1)} - \frac{y^\mathsf{T}P_1X\tilde\beta}{c+1}\right].
\end{aligned}$$

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1} \mathbb{I}_{\mathbb{N}^*}(c)$$

Recall g-prior

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1} \mathbb{I}_{\mathbb{N}^*}(c)$$

Recall g-prior

The choice of this hierarchical diffuse prior distribution on c is due to the model posterior sensitivity to large values of c:

Noninformative case

Use the same compatible informative g-prior distribution with $\tilde{\beta} = 0_{p+1}$ and a hierarchical diffuse prior distribution on c,

$$\pi(c) \propto c^{-1} \mathbb{I}_{\mathbb{N}^*}(c)$$

Recall g-prior

The choice of this hierarchical diffuse prior distribution on c is due to the model posterior sensitivity to large values of c:

Taking $\tilde{\beta} = 0_{p+1}$ and c large does not work

Influence of c

Erase influence

Consider the 10-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{3}\beta_i x_i + \sum_{i=1}^{3}\beta_{i+3}x_i^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x3 + \beta_9 x_2 x_3 + \beta_{10} x_1 x_2 x_3, \sigma^2 I_n\right)$$

where the x_is are iid $\mathscr{U}(0,10)$

[Casella & Moreno, 2004]

Influence of c

Erase influence

Consider the 10-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{3}\beta_i x_i + \sum_{i=1}^{3}\beta_{i+3}x_i^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x3 + \beta_9 x_2 x_3 + \beta_{10} x_1 x_2 x_3, \sigma^2 I_n\right)$$

where the x_is are iid $\mathscr{U}(0,10)$

[Casella & Moreno, 2004]

True model: two predictors x_1 and x_2, i.e. $\gamma^* = 110\ldots0$, $(\beta_0,\beta_1,\beta_2) = (5,1,3)$, and $\sigma^2 = 4$.

Influence of c^2

$t_1(\gamma)$	$c = 10$	$c = 100$	$c = 10^3$	$c = 10^4$	$c = 10^6$
0,1,2	0.04062	0.35368	0.65858	0.85895	0.98222
0,1,2,7	0.01326	0.06142	0.08395	0.04434	0.00524
0,1,2,4	0.01299	0.05310	0.05805	0.02868	0.00336
0,2,4	0.02927	0.03962	0.00409	0.00246	0.00254
0,1,2,8	0.01240	0.03833	0.01100	0.00126	0.00126

Noninformative case (cont'd)

In the noninformative setting,

$$\pi(\gamma|y) \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-(q_\gamma+1)/2}\left[y^\mathsf{T}y - \frac{c}{c+1}y^\mathsf{T}P_1 y\right]^{-n/2}$$

converges for all y's

Casella & Moreno's example

$t_1(\gamma)$	$\sum_{i=1}^{10^6} \pi(\gamma\|y,c)\pi(c)$
0,1,2	0.78071
0,1,2,7	0.06201
0,1,2,4	0.04119
0,1,2,8	0.01676
0,1,2,5	0.01604

Gibbs approximation

When p large, impossible to compute the posterior probabilities of the 2^p models.

Gibbs approximation

When p large, impossible to compute the posterior probabilities of the 2^p models.
Use of a Monte Carlo approximation of $\pi(\gamma|y)$

Gibbs approximation

When p large, impossible to compute the posterior probabilities of the 2^p models.
Use of a Monte Carlo approximation of $\pi(\gamma|y)$

Gibbs sampling

- At $t = 0$, draw γ^0 from the uniform distribution on Γ
- At t, for $i = 1, \ldots, p$, draw $\gamma_i^t \sim \pi(\gamma_i|y, \gamma_1^t, \ldots, \gamma_{i-1}^t, \ldots, \gamma_{i+1}^{t-1}, \ldots, \gamma_p^{t-1})$

Gibbs approximation (cont'd)

Example (Simulated data)

Severe multicolinearities among predictors for a 20-predictor full model

$$y|\beta,\sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{20} \beta_i x_i, \sigma^2 I_n\right)$$

where $x_i = z_i + 3z$, the z_i's and z are iid $\mathcal{N}_n(0_n, I_n)$.

Gibbs approximation (cont'd)

Example (Simulated data)

Severe multicolinearities among predictors for a 20-predictor full model

$$y|\beta, \sigma^2 \sim \mathcal{N}\left(\beta_0 + \sum_{i=1}^{20} \beta_i x_i, \sigma^2 I_n\right)$$

where $x_i = z_i + 3z$, the z_i's and z are iid $\mathcal{N}_n(0_n, I_n)$.
True model with $n = 180$, $\sigma^2 = 4$ and seven predictor variables

$$x_1, x_3, x_5, x_6, x_{12}, x_{18}, x_{20},$$
$$(\beta_0, \beta_1, \beta_3, \beta_5, \beta_6, \beta_{12}, \beta_{18}, \beta_{20}) = (3, 4, 1, -3, 12, -1, 5, -6)$$

Gibbs approximation (cont'd)

Example (Simulated data (2))

γ	$\pi(\gamma\|y)$	$\widehat{\pi(\gamma\|y)}^{GIBBS}$
0,1,3,5,6,12,18,20	0.1893	0.1822
0,1,3,5,6,18,20	0.0588	0.0598
0,1,3,5,6,9,12,18,20	0.0223	0.0236
0,1,3,5,6,12,14,18,20	0.0220	0.0193
0,1,2,3,5,6,12,18,20	0.0216	0.0222
0,1,3,5,6,7,12,18,20	0.0212	0.0233
0,1,3,5,6,10,12,18,20	0.0199	0.0222
0,1,3,4,5,6,12,18,20	0.0197	0.0182
0,1,3,5,6,12,15,18,20	0.0196	0.0196

Gibbs $(T = 100,000)$ results for $\tilde{\beta} = 0_{21}$ and $c = 100$

Processionary caterpillar

Influence of some forest settlement characteristics on the development of caterpillar colonies

Processionary caterpillar

Influence of some forest settlement characteristics on the development of caterpillar colonies

Processionary caterpillar

Influence of some forest settlement characteristics on the development of caterpillar colonies

Response y log-transform of the average number of nests of caterpillars per tree on an area of 500 square meters ($n = 33$ areas)

Processionary caterpillar (cont'd)

Potential explanatory variables

x_1 altitude (in meters), x_2 slope (in degrees),

x_3 number of pines in the square,

x_4 height (in meters) of the tree at the center of the square,

x_5 diameter of the tree at the center of the square,

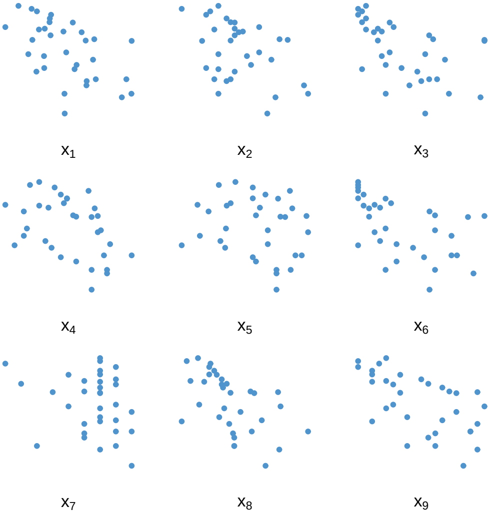
x_6 index of the settlement density,

x_7 orientation of the square (from 1 if southb'd to 2 ow),

x_8 height (in meters) of the dominant tree,

x_9 number of vegetation strata,

x_{10} mix settlement index (from 1 if not mixed to 2 if mixed).


x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9

Bayesian regression output

	Estimate	BF	log10(BF)
(Intercept)	9.2714	26.334	1.4205 (***)
X1	-0.0037	7.0839	0.8502 (**)
X2	-0.0454	3.6850	0.5664 (**)
X3	0.0573	0.4356	-0.3609
X4	-1.0905	2.8314	0.4520 (*)
X5	0.1953	2.5157	0.4007 (*)
X6	-0.3008	0.3621	-0.4412
X7	-0.2002	0.3627	-0.4404
X8	0.1526	0.4589	-0.3383
X9	-1.0835	0.9069	-0.0424
X10	-0.3651	0.4132	-0.3838

evidence against H0: (****) decisive, (***) strong, (**) subtantial, (*) poor

Bayesian variable selection

$t_1(\gamma)$	$\pi(\gamma\|y,X)$	$\widehat{\pi}(\gamma\|y,X)$
0,1,2,4,5	0.0929	0.0929
0,1,2,4,5,9	0.0325	0.0326
0,1,2,4,5,10	0.0295	0.0272
0,1,2,4,5,7	0.0231	0.0231
0,1,2,4,5,8	0.0228	0.0229
0,1,2,4,5,6	0.0228	0.0226
0,1,2,3,4,5	0.0224	0.0220
0,1,2,3,4,5,9	0.0167	0.0182
0,1,2,4,5,6,9	0.0167	0.0171
0,1,2,4,5,8,9	0.0137	0.0130

Noninformative G-prior model choice and Gibbs estimations

Postulate

Previous principle requires embedded models (or an encompassing model) and proper priors, while being hard to implement outside exponential families

Postulate

Previous principle requires embedded models (or an encompassing model) and proper priors, while being hard to implement outside exponential families
Now we determine prior measures on two models $\mathfrak{M}_1$ and $\mathfrak{M}_2$, π_1 and π_2, directly by a compatibility principle.

Generalised expected posterior priors

[Perez & Berger, 2000]

EPP Principle

Starting from reference priors π_1^N and π_2^N, substitute by prior distributions π_1 and π_2 that solve the system of integral equations

$$\pi_1(\theta_1) = \int_{\mathscr{X}} \pi_1^N(\theta_1 \,|\, x) m_2(x) \mathrm{d}x$$

and

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \,|\, x) m_1(x) \mathrm{d}x,$$

where x is an imaginary minimal training sample and m_1, m_2 are the marginals associated with π_1 and π_2 respectively.

Motivations

- Eliminates the "imaginary observation" device and proper-isation through part of the data by integration under the "truth"

Motivations

- Eliminates the "imaginary observation" device and proper-isation through part of the data by integration under the "truth"
- Assumes that both models are *equally* valid and equipped with ideal unknown priors

$$\pi_i, \quad i = 1, 2,$$

that yield "true" marginals balancing each model wrt the other

Motivations

- Eliminates the “imaginary observation” device and proper-isation through part of the data by integration under the “truth”
- Assumes that both models are *equally* valid and equipped with ideal unknown priors

$$\pi_i, \quad i = 1, 2,$$

  that yield “true” marginals balancing each model wrt the other
- For a *given* π_1, π_2 is an **expected posterior prior**
  Using both equations introduces symmetry into the game

Dual properness

Theorem (Proper distributions)

If π_1 is a probability density then π_2 solution to

$$\pi_2(\theta_2) = \int_{\mathscr{X}} \pi_2^N(\theta_2 \,|\, x) m_1(x) dx$$

is a probability density

© Both EPPs are either proper or improper

Bayesian coherence

Theorem (True Bayes factor)

If π_1 and π_2 are the EPPs and if their marginals are finite, then the corresponding Bayes factor

$$B_{1,2}(\mathbf{x})$$

is either a (true) Bayes factor or a limit of (true) Bayes factors.

Bayesian coherence

Theorem (True Bayes factor)

If π_1 and π_2 are the EPPs and if their marginals are finite, then the corresponding Bayes factor

$$B_{1,2}(\mathbf{x})$$

is either a (true) Bayes factor or a limit of (true) Bayes factors.

Obviously only interesting when both π_1 and π_2 are improper.

Existence/Unicity

Theorem (Recurrence condition)

When both the observations and the parameters in both models are continuous, if the Markov chain with transition

$$Q\left(\theta_1' \,|\, \theta_1\right) = \int g\left(\theta_1, \theta_1', \theta_2, x, x'\right) \mathrm{d}x \mathrm{d}x' \mathrm{d}\theta_2$$

where

$$g\left(\theta_1, \theta_1', \theta_2, x, x'\right) = \pi_1^N\left(\theta_1' \,|\, x\right) f_2\left(x \,|\, \theta_2\right) \pi_2^N\left(\theta_2 \,|\, x'\right) f_1\left(x' \,|\, \theta_1\right),$$

is recurrent, *then there exists a solution to the integral equations, unique up to a multiplicative constant.*

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPS.

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPS.
- The transition density $Q\left(\theta_1' \mid \theta_1\right)$ has a dual transition density on Θ_2.

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPS.
- The transition density $Q\left(\theta_1' \mid \theta_1\right)$ has a dual transition density on Θ_2.
- There exists a parallel M chain on Θ_2 with identical properties; if one is (Harris) recurrent, so is the other.

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPS.
- The transition density $Q(\theta_1' \mid \theta_1)$ has a dual transition density on Θ_2.
- There exists a parallel M chain on Θ_2 with identical properties; if one is (Harris) recurrent, so is the other.
- **Duality property** found both in the MCMC literature and in decision theory

[Diebolt & Robert, 1992; Eaton, 1992]

Consequences

- If the M chain is positive recurrent, there exists a unique pair of proper EPPS.
- The transition density $Q(\theta_1' \mid \theta_1)$ has a dual transition density on Θ_2.
- There exists a parallel M chain on Θ_2 with identical properties; if one is (Harris) recurrent, so is the other.
- **Duality property** found both in the MCMC literature and in decision theory

  [Diebolt & Robert, 1992; Eaton, 1992]
- When Harris recurrence holds but the EPPs cannot be found, the Bayes factor can be approximated by MCMC simulation

Point null hypothesis testing

Testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$, i.e.

$$\begin{aligned} \mathfrak{M}_1 &: \quad f\left(x \,|\, \theta^*\right), \\ \mathfrak{M}_2 &: \quad f\left(x \,|\, \theta\right), \theta \in \Theta. \end{aligned}$$

Point null hypothesis testing

Testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$, i.e.

$$\begin{aligned} \mathfrak{M}_1 &: \quad f\left(x \,|\, \theta^*\right), \\ \mathfrak{M}_2 &: \quad f\left(x \,|\, \theta\right), \theta \in \Theta. \end{aligned}$$

Default priors

$$\pi_1^N\left(\theta\right) = \delta_{\theta^*}\left(\theta\right) \text{ and } \pi_2^N\left(\theta\right) = \pi^N\left(\theta\right)$$

Point null hypothesis testing

Testing $H_0 : \theta = \theta^*$ versus $H_1 : \theta \neq \theta^*$, i.e.

$$\begin{aligned} \mathfrak{M}_1 &: \quad f(x \mid \theta^*), \\ \mathfrak{M}_2 &: \quad f(x \mid \theta), \theta \in \Theta. \end{aligned}$$

Default priors

$$\pi_1^N(\theta) = \delta_{\theta^*}(\theta) \text{ and } \pi_2^N(\theta) = \pi^N(\theta)$$

For x minimal training sample, consider the proper priors

$$\pi_1(\theta) = \delta_{\theta^*}(\theta) \text{ and } \pi_2(\theta) = \int \pi^N(\theta \mid x) f(x \mid \theta^*)\, \mathrm{d}x$$

Point null hypothesis testing (cont'd)

Then

$$\int \pi_1^N(\theta \,|\, x)\, m_2(x)\, \mathrm{d}x = \delta_{\theta^*}(\theta) \int m_2(x)\, \mathrm{d}x = \delta_{\theta^*}(\theta) = \pi_1(\theta)$$

and

$$\int \pi_2^N(\theta \,|\, x)\, m_1(x)\, \mathrm{d}x = \int \pi^N(\theta \,|\, x)\, f(x \,|\, \theta^*)\, \mathrm{d}x = \pi_2(\theta)$$

Point null hypothesis testing (cont'd)

Then

$$\int \pi_1^N(\theta \,|\, x)\, m_2(x)\, \mathrm{d}x = \delta_{\theta^*}(\theta) \int m_2(x)\, \mathrm{d}x = \delta_{\theta^*}(\theta) = \pi_1(\theta)$$

and

$$\int \pi_2^N(\theta \,|\, x)\, m_1(x)\, \mathrm{d}x = \int \pi^N(\theta \,|\, x)\, f(x \,|\, \theta^*)\, \mathrm{d}x = \pi_2(\theta)$$

©$\pi_1(\theta)$ and $\pi_2(\theta)$ are integral priors

Point null hypothesis testing (cont'd)

Then

$$\int \pi_1^N(\theta\,|\,x)\,m_2(x)\,\mathrm{d}x = \delta_{\theta^*}(\theta)\int m_2(x)\,\mathrm{d}x = \delta_{\theta^*}(\theta) = \pi_1(\theta)$$

and

$$\int \pi_2^N(\theta\,|\,x)\,m_1(x)\,\mathrm{d}x = \int \pi^N(\theta\,|\,x)\,f(x\,|\,\theta^*)\,\mathrm{d}x = \pi_2(\theta)$$

©$\pi_1(\theta)$ and $\pi_2(\theta)$ are integral priors

Note

Uniqueness of the Bayes factor
Integral priors and intrinsic priors coincide

[Moreno, Bertolino and Racugno, 1998]

Location models

Two location models

$$
\begin{aligned}
\mathfrak{M}_1 &: \quad f_1\left(x \,|\, \theta_1\right) = f_1\left(x - \theta_1\right) \\
\mathfrak{M}_2 &: \quad f_2\left(x \,|\, \theta_2\right) = f_2\left(x - \theta_2\right)
\end{aligned}
$$

Location models

Two location models

$$\begin{aligned} \mathfrak{M}_1 &: \quad f_1\left(x \,|\, \theta_1\right) = f_1\left(x - \theta_1\right) \\ \mathfrak{M}_2 &: \quad f_2\left(x \,|\, \theta_2\right) = f_2\left(x - \theta_2\right) \end{aligned}$$

Default priors

$$\pi_i^N\left(\theta_i\right) = c_i, \quad i = 1, 2$$

with minimal training sample size **one**

Location models

Two location models

$$\begin{aligned}\mathfrak{M}_1 &: \quad f_1(x \mid \theta_1) = f_1(x - \theta_1) \\ \mathfrak{M}_2 &: \quad f_2(x \mid \theta_2) = f_2(x - \theta_2)\end{aligned}$$

Default priors

$$\pi_i^N(\theta_i) = c_i, \quad i = 1, 2$$

with minimal training sample size **one**
Marginal densities

$$m_i^N(x) = c_i, \quad i = 1, 2$$

Location models (cont'd)

In that case, $\pi_1^N(\theta_1)$ and $\pi_2^N(\theta_2)$ are integral priors **when** $c_1 = c_2$:

$$
\begin{aligned}
\int \pi_1^N(\theta_1 \mid x)\, m_2^N(x)\, \mathrm{d}x &= \int c_2 f_1(x-\theta_1)\, \mathrm{d}x = c_2 \\
\int \pi_2^N(\theta_2 \mid x)\, m_1^N(x)\, \mathrm{d}x &= \int c_1 f_2(x-\theta_2)\, \mathrm{d}x = c_1 .
\end{aligned}
$$

Location models (cont'd)

In that case, $\pi_1^N(\theta_1)$ and $\pi_2^N(\theta_2)$ are integral priors **when** $c_1 = c_2$:

$$\begin{aligned}\int \pi_1^N(\theta_1 \mid x)\, m_2^N(x)\, \mathrm{d}x &= \int c_2 f_1(x-\theta_1)\, \mathrm{d}x = c_2\\ \int \pi_2^N(\theta_2 \mid x)\, m_1^N(x)\, \mathrm{d}x &= \int c_1 f_2(x-\theta_2)\, \mathrm{d}x = c_1.\end{aligned}$$

© If the associated Markov chain is recurrent,

$$\pi_1^N(\theta_1) = \pi_2^N(\theta_2) = c$$

are the unique integral priors and they are intrinsic priors

[Cano, Kessler & Moreno, 2004]

Location models (cont'd)

Example (Normal versus double exponential)

$$
\begin{aligned}
\mathfrak{M}_1 &: \mathcal{N}(\theta,1), \quad \pi_1^N(\theta) = c_1, \\
\mathfrak{M}_2 &: \mathcal{DE}(\lambda,1), \quad \pi_2^N(\lambda) = c_2.
\end{aligned}
$$

Minimal training sample size one and posterior densities

$$\pi_1^N(\theta \mid x) = \mathcal{N}(x,1) \text{ and } \pi_2^N(\lambda \mid x) = \mathcal{DE}(x,1)$$

Location models (cont'd)

Example (Normal versus double exponential (2))

Transition $\theta \to \theta'$ of the Markov chain made of steps :

1. $x' = \theta + \varepsilon_1$, $\varepsilon_1 \sim \mathcal{N}(0,1)$
2. $\lambda = x' + \varepsilon_2$, $\varepsilon_2 \sim \mathcal{DE}(0,1)$
3. $x = \lambda + \varepsilon_3$, $\varepsilon_3 \sim \mathcal{DE}(0,1)$
4. $\theta' = x + \varepsilon_4$, $\varepsilon_4 \sim \mathcal{N}(0,1)$

$$\text{i.e.} \qquad \theta' = \theta + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$$

Location models (cont'd)

Example (Normal versus double exponential (2))

Transition $\theta \to \theta'$ of the Markov chain made of steps :

1. $x' = \theta + \varepsilon_1$, $\varepsilon_1 \sim \mathcal{N}(0,1)$
2. $\lambda = x' + \varepsilon_2$, $\varepsilon_2 \sim \mathcal{DE}(0,1)$
3. $x = \lambda + \varepsilon_3$, $\varepsilon_3 \sim \mathcal{DE}(0,1)$
4. $\theta' = x + \varepsilon_4$, $\varepsilon_4 \sim \mathcal{N}(0,1)$

$$\text{i.e.} \qquad \theta' = \theta + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$$

random walk in θ with finite second moment, null recurrent
© **Resulting Lebesgue measures $\pi_1(\theta) = 1 = \pi_2(\lambda)$ invariant and unique solutions to integral equations**

Admissibility and Complete Classes

Introduction

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions

Bayesian Point Estimation

Bayesian Calculations

Tests and model choice

Admissibility and Complete Classes

Admissibility of Bayes estimators

Admissibility of Bayes estimators

Warning

Bayes estimators may be inadmissible when the Bayes risk is infinite

Example (Normal mean)

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

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

Example (Normal mean)

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

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

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

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

Example (Normal mean (2))

The corresponding Bayes risk is

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

Example (Normal mean (2))

The corresponding Bayes risk is

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

which is infinite for $\alpha \leq \sigma^2$.

Example (Normal mean (2))

The corresponding Bayes risk is

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

which is infinite for $\alpha \leq \sigma^2$. 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

Theorem (Existence of an admissible Bayes estimator)

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

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,

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

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,

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

Theorem (Conjugate admissibility)

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 (Binomial probability)

Consider $x \sim \mathscr{B}(n, p)$.

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

Example (Binomial probability)

Consider $x \sim \mathscr{B}(n, p)$.

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

Then the two integrals

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

cannot diverge simultaneously if $\lambda < 0$.

Example (Binomial probability (2))

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

Example (Binomial probability (2))

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

Admissible Bayes estimators of p

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

Differential representations

Setting of multidimensional exponential families

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

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

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

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

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

and

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

Consequence

Theorem

If

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

the estimator

$$\delta_0(x) = x$$

is admissible.

Consequence

Theorem

If

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

the estimator

$$\delta_0(x) = x$$

is admissible.

Example (Normal mean (3))

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

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

A generalised Bayes estimator of the form $\delta(x) = (1 - h(||x||))x$

1. is inadmissible if there exist $\epsilon > 0$ and $K < +\infty$ such that

$$||x||^2 h(||x||) < p - 2 - \epsilon \quad \text{for } ||x|| > K$$

2. is admissible if there exist K_1 and K_2 such that $h(||x||)||x|| \leq K_1$ for every x and

$$||x||^2 h(||x||) \geq p - 2 \quad \text{for } ||x|| > K_2$$

[Brown, 1971]

Recurrence conditions

General case

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

For a given prior π, Markovian transition kernel

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

Theorem (Recurrence)

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

[Eaton, 1994]

Recurrence conditions (cont.)

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

Recurrence conditions (cont.)

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

Theorem (Recurrence(2))

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

[Eaton, 1999]

Necessary and sufficient admissibility conditions

Formalisation of the statement that...

Necessary and sufficient admissibility conditions

Formalisation of the statement that...

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

Blyth's sufficient condition

Theorem (Blyth condition)

If, for an estimator δ_0, there exists a sequence (π_n) of generalised 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 (Normal mean (4))

Consider $x \sim \mathscr{N}(\theta, 1)$ 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

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

[condition (i) is satisfied]

Example (Normal mean (5))

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]

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,

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

Stein's necessary and sufficient condition (cont.)

Theorem (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

Stein's necessary and sufficient condition (cont.)

Theorem (Stein's n&s condition (cont.))

(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,* $\sup_n \pi_n(E) < +\infty$*;*

(iii) $\lim_n r(\pi_n, \delta) - r(\pi_n) = 0$*; and*

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

Complete classes

Definition (Complete class)

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

A special case

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

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

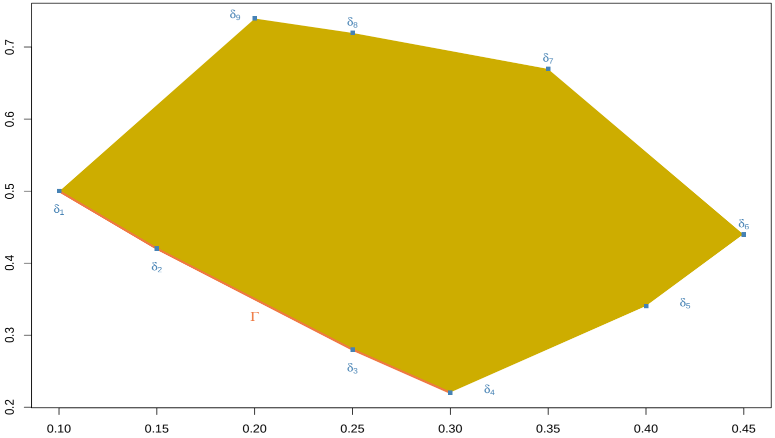
bounded and closed from below

A special case

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

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

bounded and closed from below
Then, the lower boundary, $\Gamma(\mathscr{R})$, provides the *admissible* points of $\mathscr{R}$.


δ9
δ8
δ7
δ6
δ5
δ4
δ3
δ2
δ1
Γ
0.2
0.3
0.4
0.5
0.6
0.7
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45

A special case (cont.)

Reason

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

$$p_1 r_1 + p_2 r_2 = k$$

A special case (cont.)

Reason

For every $r \in \Gamma(\mathscr{R})$, there exists a tangent line to $\mathscr{R}$ 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

Theorem

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

Wald's theorems

Theorem

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

Theorem

If Θ is compact and the risk set $\mathscr{R}$ 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 generalised Bayes estimators

Extensions

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

Example

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 generalised Bayes estimator.

Hierarchical and Empirical Bayes Extensions

Introduction

Decision-Theoretic Foundations of Statistical Inference

From Prior Information to Prior Distributions

Bayesian Point Estimation

Bayesian Calculations

Tests and model choice

Admissibility and Complete Classes

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

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.

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]

Hierarchical Bayes analysis

Decomposition of the prior distribution into several conditional levels of distributions

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

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, ...)

Hierarchical models

Definition (Hierarchical model)

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

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

Hierarchical models

Definition (Hierarchical model)

A *hierarchical Bayes model* is a Bayesian statistic 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 (Rats (1))

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

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

Example (Rats (1))

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

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

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

Example (Rats (2))

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.

Example (Rats (2))

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

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

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.

Justifications

1. **Objective reasons based on prior information**

Justifications

1. **Objective reasons based on prior information**

Example (Rats (3))

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.

2. **Separation of structural information from subjective information**

2. **Separation of structural information from subjective information**

Example (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

2. **Separation of structural information from subjective information**

Example (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

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

3. **In noninformative settings, compromise between the Jeffreys noninformative distributions, and the conjugate distributions.**

3. **In noninformative settings, compromise between the Jeffreys noninformative distributions, and the conjugate distributions.**
4. **Robustification of the usual Bayesian analysis from a frequentist point of view**

3. **In noninformative settings, compromise between the Jeffreys noninformative distributions, and the conjugate distributions.**
4. **Robustification of the usual Bayesian analysis from a frequentist point of view**
5. **Often simplifies Bayesian calculations**

Conditional decompositions

Easy decomposition of the posterior distribution

Conditional decompositions

Easy decomposition of the posterior distribution
For instance, if

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

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

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

Conditional decompositions (cont.)

where

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

Conditional decompositions (cont.)

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

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

where

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

Example (Posterior distribution of the complete parameter vector)

Posterior distribution of the complete parameter vector

$$
\begin{aligned}
&\pi((\theta_i,\delta_i,\xi_i)_i,\mu_\theta,\ldots,\sigma_c,\ldots|\mathscr{D}) \propto \\
&\prod_{i=1}^{I}\Big\{\exp-\{(\theta_i-\mu_\theta)^2/2\sigma_\theta^2+(\delta_i-\mu_\delta)^2/2\sigma_\delta^2\} \\
&\prod_{j=1}^{J_i^c}\exp-\{(x_{ij}-\theta_i)^2/2\sigma_c^2\}\prod_{j=1}^{J_i^a}\exp-\{(y_{ij}-\theta_i-\delta_i)^2/2\sigma_a^2\} \\
&\prod_{j=1}^{J_i^t}\exp-\{(z_{ij}-\theta_i-\delta_i-\xi_i)^2/2\sigma_t^2\}\Big\} \\
&\prod_{\ell_i=0}\exp-\{(\xi_i-\mu_P)^2/2\sigma_P^2\}\prod_{\ell_i=1}\exp-\{(\xi_i-\mu_D)^2/2\sigma_D^2\}
\end{aligned}
$$

Local conditioning property

Theorem (Decomposition)

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* $\theta_{n+1} = 0$.

Computational issues

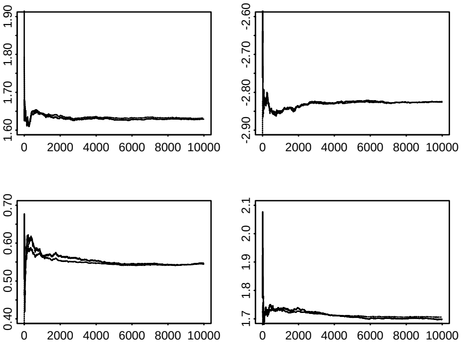
Rarely an explicit derivation of the corresponding Bayes estimators
Natural solution in hierarchical settings: use a simulation-based approach exploiting the hierarchical conditional structure

Computational issues

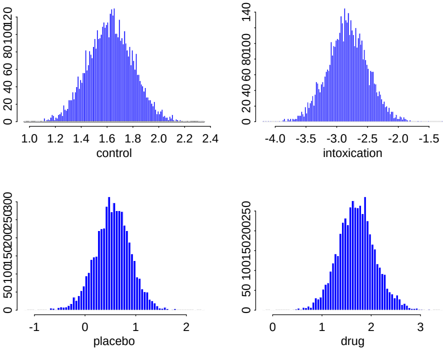
Rarely an explicit derivation of the corresponding Bayes estimators
Natural solution in hierarchical settings: use a simulation-based approach exploiting the hierarchical conditional structure

Example (Rats (4))

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

Hierarchical extensions for the normal model

For

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

the hierarchical Bayes estimator is

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

Hierarchical extensions for the normal model

For

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

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 (Exchangeable normal)

Consider the *exchangeable* hierarchical model

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

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

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

Example (Exchangeable normal (2))

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

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

Example (Exchangeable normal (2))

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

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

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

Example (Exchangeable normal (3))

Notice the particular form of the hierarchical Bayes estimator

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

[Double shrinkage]

The Stein effect

If a minimax estimator is unique, it is admissible

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 $p \geq p_0$,

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,

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

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

Universality

- Other distributions than the normal distribution
- Other losses other than the quadratic loss

Universality

- Other distributions than the normal distribution
- Other losses other than the quadratic loss
- Connections with admissibility

Universality

- Other distributions than the normal distribution
- Other losses other than the quadratic loss
- Connections with admissibility
- George's multiple shrinkage

Universality

- Other distributions than the normal distribution
- Other losses other than the quadratic loss
- Connections with admissibility
- George's multiple shrinkage
- Robustess against distribution

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

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

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

A general Stein-type domination result (cont.)

Theorem (Stein domination of δ_0)

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

dominates δ_0 under quadratic loss if there exist α, $\beta > 0$ 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) $0 \leq (t/u)h(t, u) \leq \dfrac{2(q-2)\alpha}{p - q - 2 + 4\beta}$.

Optimality of hierarchical Bayes estimators

Consider

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

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

Optimality of hierarchical Bayes estimators

Consider

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

where Σ is known.
Prior distribution on θ is $\theta \sim \mathscr{N}_p(\mu, \Sigma_\pi)$.
The prior distribution π_2 of the hyperparameters

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

is decomposed as

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

Optimality of hierarchical Bayes estimators

In this case,

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

with

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

Optimality of hierarchical Bayes estimators

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

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

if it satisfies

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

A sufficient condition (contd.)

Theorem (Minimaxity)

If $m(x)$ satisfies the three conditions

(1) $\mathbb{E}_\theta \|\nabla \log m(x)\|^2 < +\infty$; (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

$$\mathscr{D}\delta^\pi(x) = \mathrm{tr}(Q\Sigma) \\ + \frac{2}{m(x)}\mathrm{tr}(H_m(x)\tilde{Q}) - (\nabla \log m(x))^t \tilde{Q} (\nabla \log m(x))$$

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

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

δ^π is minimax if

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

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

Theorem (Superharmonicity)

δ^π **is minimax if**

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

Superharmonicity condition

Theorem (Superharmonicity)

δ^π **is minimax if**

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

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

Empirical Bayes alternative

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

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.

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 (Poisson estimation)

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

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

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

Example (Poisson estimation)

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

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

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

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

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

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

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 (Normal mean)

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

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

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

$$\delta^{\mathsf{EB}}(x) = \frac{\hat{\tau}^2 x}{1+\hat{\tau}^2} = \left(1 - \frac{p}{||x||^2}\right)^+ x.$$

Normal model

Take

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

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

Normal model

Take

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

with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and Z a $(p \times q)$ full rank matrix. The marginal distribution of x is

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

and the posterior distribution of θ is

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

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

Normal model (cont.)

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

Normal model (cont.)

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

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

which is of the form of the general Stein estimator

Normal model (cont.)

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

Normal model (cont.)

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

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

b. Variance evaluation

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

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,

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

induces an underestimation of this variance

Morris' correction

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

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

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

Unlimited range of applications

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

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

c@enumi). **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 c@enumi). **Conditioning on the data**

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

c@enumi). **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.

c@enumi). **Using priors as tools and summaries**

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

c@enumi). **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

c@enumi). **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 , then derive the best solution satisfying these properties.

c@enumi). **Looking for optimal procedures**

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

Optimality is easier to attain through the Bayes “filter”

c@enumi). **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.

c@enumi). **Providing a universal system of inference**

Given the three factors

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

the Bayesian approach validates one and only one inferential procedure

c@enumi). **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*method!universal 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.