Bayesian data analysis for Ecologists

Christian P. Robert, Université Paris Dauphine

Parco Nazionale Gran Paradiso, 13-17 Luglio 2009

Outline

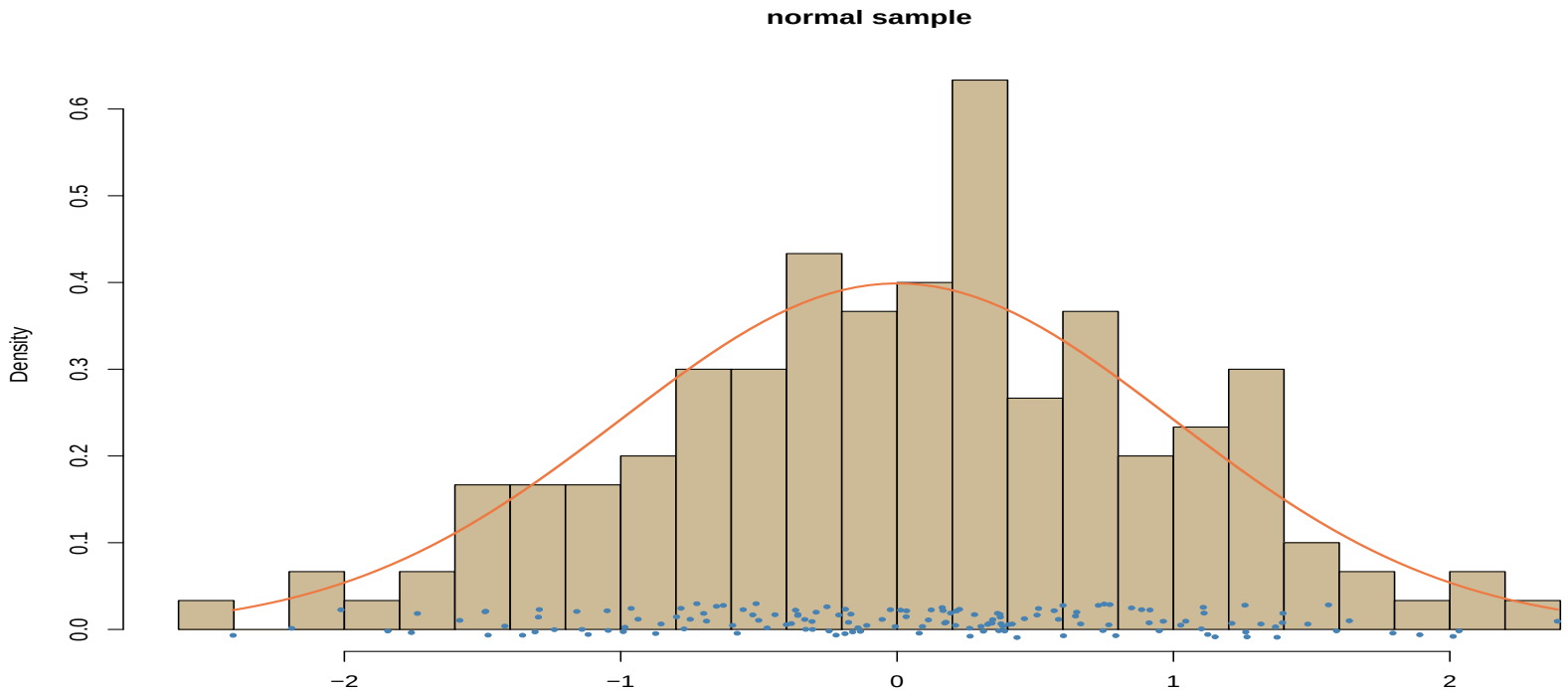
The normal model

Normal model

Sample

$$x_1, \ldots, x_n$$

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



Inference on (μ, σ) based on this sample

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

Inference on (μ, σ) based on this sample

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

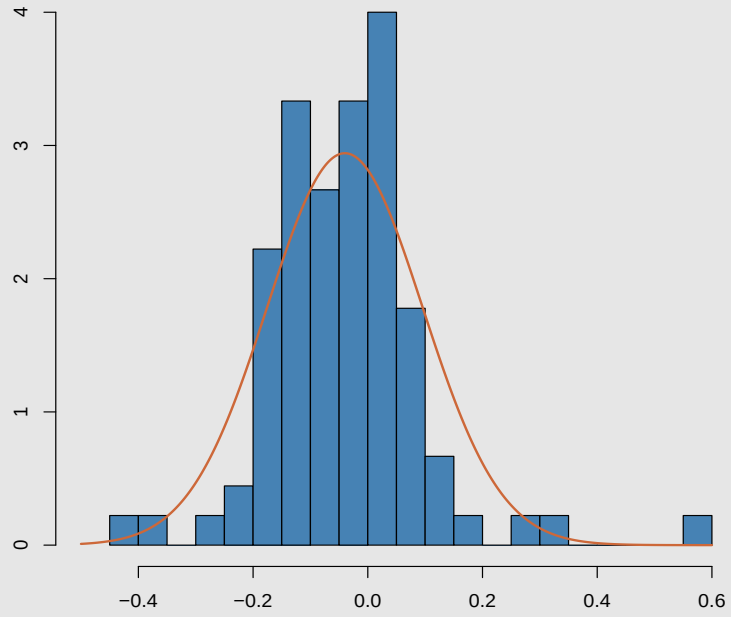
Inference on (μ, σ) based on this sample

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

Datasets

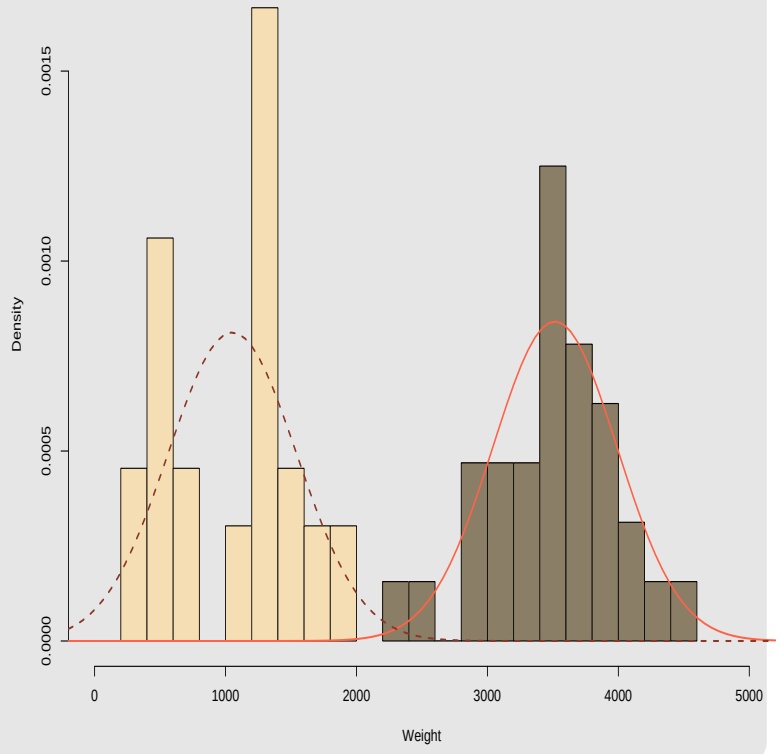
Larcenies =normaldata

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



Marmot weights

Marmots (72) separated by age between adults and others, with normal estimation



The Bayesian toolbox

Bayes theorem = Inversion of probabilities

The Bayesian toolbox

Bayes theorem = Inversion of probabilities

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

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

Who's Bayes?

Reverend Thomas Bayes (ca. 1702–1761)

Presbyterian minister in Tunbridge Wells (Kent) from 1731, son of Joshua Bayes, nonconformist minister. Election to the *Royal Society* based on a tract of 1736 where he defended the views and philosophy of Newton.

Who's Bayes?

Reverend Thomas Bayes (ca. 1702–1761)

Presbyterian minister in Tunbridge Wells (Kent) from 1731, son of Joshua Bayes, nonconformist minister. Election to the *Royal Society* based on a tract of 1736 where he defended the views and philosophy of Newton.

Sole probability paper, *"Essay Towards Solving a Problem in the Doctrine of Chances"*, published posthumously in 1763 by Pierce and containing the seeds of *Bayes' Theorem*.

New perspective

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

New perspective

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

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

Bayesian model

A Bayesian statistical model is made of

1. a likelihood

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

Bayesian model

A Bayesian statistical model is made of

1. a likelihood

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

and of

2. a prior distribution on the parameters,

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

Justifications

- Semantic drift from unknown θ to random θ

Justifications

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

Justifications

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

Justifications

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

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

Assume

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

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

Assume

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

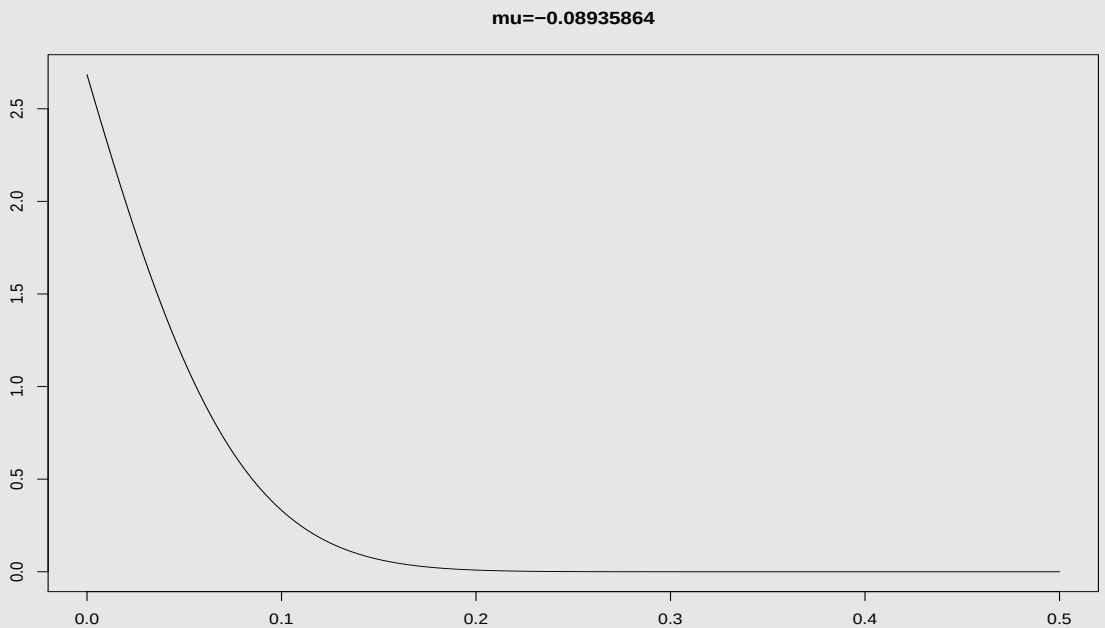
Then

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

Example (Normal illustration (2))

Truncated normal distribution

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



Prior and posterior distributions

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

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

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

Prior and posterior distributions

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

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

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

2. the *marginal distribution* of x,

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

3. the *posterior distribution* of θ,

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

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

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

Posterior distribution center of Bayesian inference

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

- Operates conditional upon the observations

Posterior distribution center of Bayesian inference

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

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

Posterior distribution center of Bayesian inference

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

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

Posterior distribution center of Bayesian inference

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

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

Posterior distribution center of Bayesian inference

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

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

Example (Normal-normal case)

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

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

Example (Normal-normal case)

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

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

and

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

Prior selection

The prior distribution is the key to Bayesian inference

Prior selection

The prior distribution is the key to Bayesian inference

But...

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

Prior selection

The prior distribution is the key to Bayesian inference

But...

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

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

Strategies for prior determination

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

Strategies for prior determination

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

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

Strategies for prior determination

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

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

Strategies for prior determination

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

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

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

Strategies for prior determination

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

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

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

- Empirical and *hierarchical* Bayes techniques

Conjugate priors

Specific parametric family with analytical properties

Conjugate prior

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

Conjugate priors

Specific parametric family with analytical properties

Conjugate prior

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

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

Justifications

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

Justifications

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

Justifications

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

Justifications

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

Justifications

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

Exponential families

Sampling models of interest

Exponential family

The family of distributions

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

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

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

the family is said to be *natural*.

Analytical properties of exponential families

- Sufficient statistics (Pitman–Koopman Lemma)

Analytical properties of exponential families

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

Analytical properties of exponential families

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

Analytical properties of exponential families

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

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

Standard exponential families

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

More...

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

Linearity of the posterior mean

If

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

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

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

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

Linearity of the posterior mean

If

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

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

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

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

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

Example (Normal-normal)

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

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

where

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

Example (Full normal)

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

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

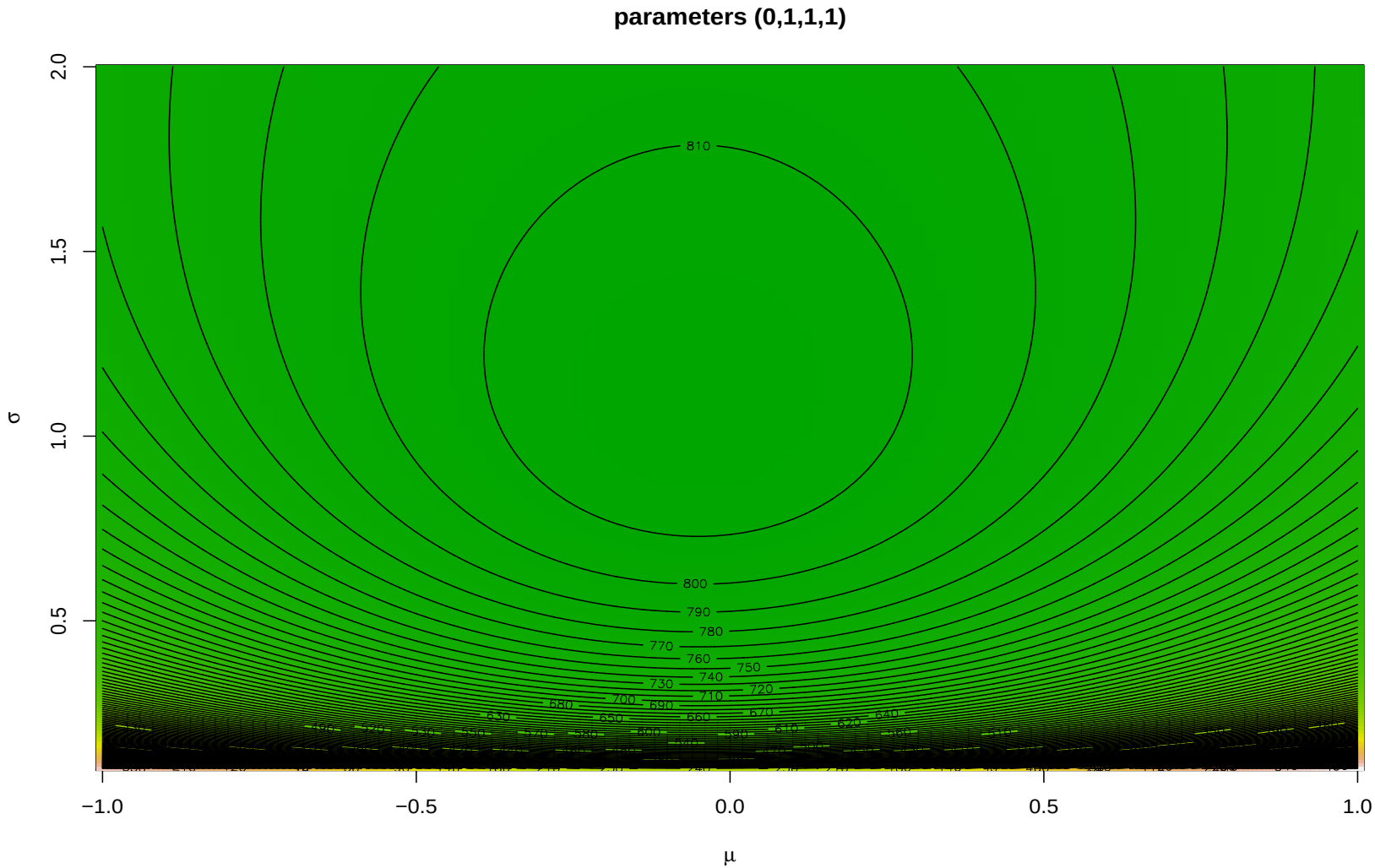
Example (Full normal)

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

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

since

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


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

Improper prior distribution

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

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

Improper prior distribution

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

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

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

Justifications

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

Justifications

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

Justifications

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

Justifications

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

Justifications

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

Validation

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

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

when

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

Example (Normal+improper)

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

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

and the posterior distribution of θ is

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

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

[independent of ϖ]

Meaningless as probability distribution

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

Meaningless as probability distribution

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

Example

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

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

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

Noninformative prior distributions

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

Noninformative prior distributions

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

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

Noninformative prior distributions

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

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

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

Laplace's prior

Principle of *Insufficient Reason* (Laplace)

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

Laplace's prior

Principle of *Insufficient Reason* (Laplace)

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

Extension to continuous spaces

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

[Lebesgue measure]

Who's Laplace?

Pierre Simon de Laplace (1749–1827)

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

Who's Laplace?

Pierre Simon de Laplace (1749–1827)

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

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

Laplace's problem

- Lack of reparameterization invariance/coherence

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

Laplace's problem

- Lack of reparameterization invariance/coherence

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

- Problems of properness

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

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

Jeffreys' prior

Based on Fisher information

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

Ron Fisher (1890–1962)

Jeffreys' prior

Based on Fisher information

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

Ron Fisher (1890–1962)

the Jeffreys prior distribution is

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

Who's Jeffreys?

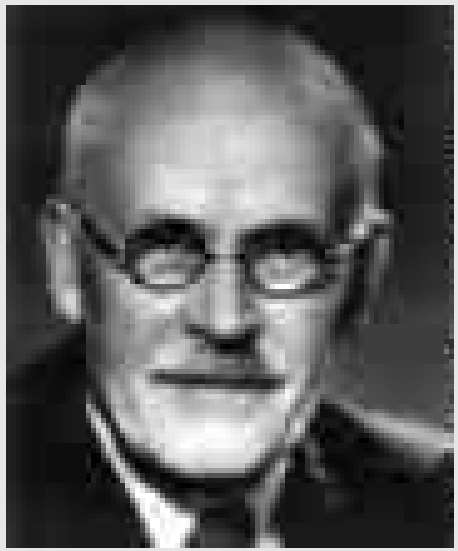
Sir Harold Jeffreys (1891–1989)

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

Who's Jeffreys?

Sir Harold Jeffreys (1891–1989)

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



Pros & Cons

- Relates to information theory

Pros & Cons

- Relates to information theory
- Agrees with most invariant priors

Pros & Cons

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

Pros & Cons

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

Evaluating estimators

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

Evaluating estimators

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

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

Evaluating estimators

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

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

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

[DeGroot, 1970]

Loss functions

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

Loss functions

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

Impossible to uniformly minimize (in d) the loss function

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

when θ is unknown

Bayesian estimation

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

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

and minimise in d

Bayes estimates

Bayes estimator

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

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

The quadratic loss

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

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

The quadratic loss

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

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

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

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

The absolute error loss

Alternatives to the quadratic loss:

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

or

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

The absolute error loss

Alternatives to the quadratic loss:

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

or

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

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

MAP estimator

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

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

MAP estimator

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

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

- Penalized likelihood estimator

MAP estimator

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

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

- Penalized likelihood estimator
- Further appeal in restricted parameter spaces

Example (Binomial probability)

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

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

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

Example (Binomial probability)

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

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

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

Corresponding MAP estimators:

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

Not always appropriate

Example (Fixed MAP)

Consider

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

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

Not always appropriate

Example (Fixed MAP)

Consider

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

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

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

Credible regions

Natural confidence region: Highest posterior density (HPD) region

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

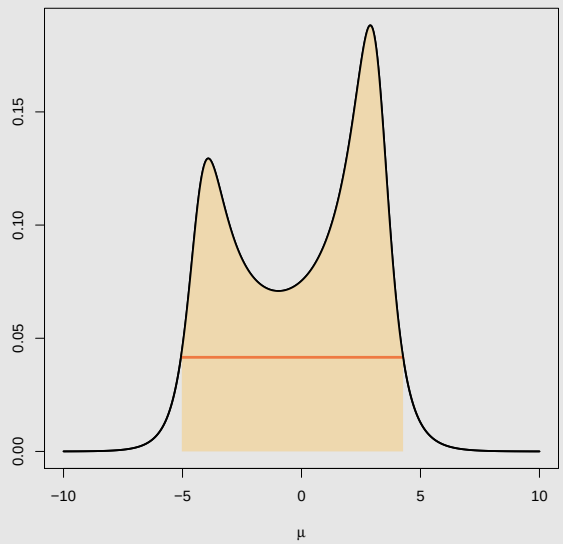
Credible regions

Natural confidence region: Highest posterior density (HPD) region

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

Optimality

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



Example

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

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

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

Example

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

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

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

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

the "usual" (classical) confidence interval

Full normal

Under [almost!] *Jeffreys prior* prior

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

posterior distribution of (μ, σ)

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

Full normal

Under [almost!] *Jeffreys prior* prior

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

posterior distribution of (μ, σ)

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

Then

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

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

Normal credible interval

Derived credible interval on μ

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

Normal credible interval

Derived credible interval on μ

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

marmotdata

Corresponding 95% confidence region for μ the mean for adults

$$[3484.435, 3546.815]$$

[Warning!] It is not because AUDREY has a weight of $2375g$ that she can be excluded from the adult group!

Testing hypotheses

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

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

Testing hypotheses

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

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

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

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

Testing hypotheses

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

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

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

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

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

The $0-1$ loss

Rudimentary loss function

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

Associated Bayes estimate

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

The $0-1$ loss

Rudimentary loss function

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

Associated Bayes estimate

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

Intuitive structure

Extension

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

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

Extension

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

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

Associated Bayes estimator

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

Example (Normal-normal)

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

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

Example (Normal-normal)

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

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

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

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

Example (Normal-normal (2))

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

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

H_0 is accepted when

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

the upper acceptance bound then being

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

Bayes factor

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

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

in the absence of loss function parameters a_0 and a_1

Associated reparameterisations

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

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

Associated reparameterisations

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

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

If we rewrite the prior as

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

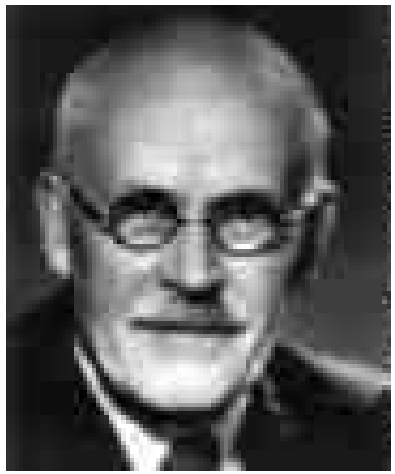
then

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

[Akin to likelihood ratio]

Jeffreys' scale

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



Point null difficulties

If π absolutely continuous,

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

Point null difficulties

If π absolutely continuous,

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

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

New prior for new hypothesis

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

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

(hidden information) or

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

New prior for new hypothesis

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

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

(hidden information) or

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

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

Posteriors with Dirac masses

If $H_0 : \theta = \theta_0 \ (= \Theta_0)$,

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

then

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

with

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

Example (Normal-normal)

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

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

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

Example (Normal-normal)

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

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

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

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

Example (cont'd)

and

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

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

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

Banning improper priors

Impossibility of using improper priors for testing!

Banning improper priors

Impossibility of using improper priors for testing!

Reason: When using the representation

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

π_1 and π_0 must be normalised

Example (Normal point null)

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

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

and

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

Example (Normal point null (2))

Consequence: probability of H_0 is bounded from above by

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

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

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

Example (Normal one-sided)

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

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

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

Example (Normal one-sided)

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

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

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

normaldata

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

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

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

Jeffreys–Lindley paradox

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

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

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

Jeffreys–Lindley paradox

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

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

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

Difference with the "noninformative" answer

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

[Invalid answer]

Normalisation difficulties

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

Normalisation difficulties

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

Example

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

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

Monte Carlo integration

Generic problem of evaluating an integral

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

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

Monte Carlo Principle

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

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

Monte Carlo Principle

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

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

Convergence of the average

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

by the **Strong Law of Large Numbers**

Bayes factor approximation

For the normal case

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

under prior

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

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

Example

CMBdata

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

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

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

Precision evaluation

Estimate the variance with

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

and for m large,

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

Precision evaluation

Estimate the variance with

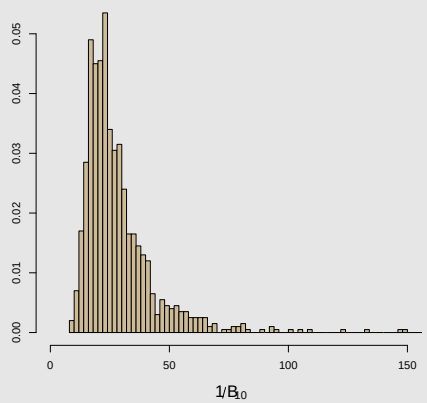
$$v_m = \frac{1}{m}\frac{1}{m-1}\sum_{j=1}^{m}[h(x_j) - \overline{h}_m]^2,$$

and for m large,

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

Note

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



Example (Cauchy-normal)

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

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

Under squared error loss, posterior mean

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

Example (Cauchy-normal (2))

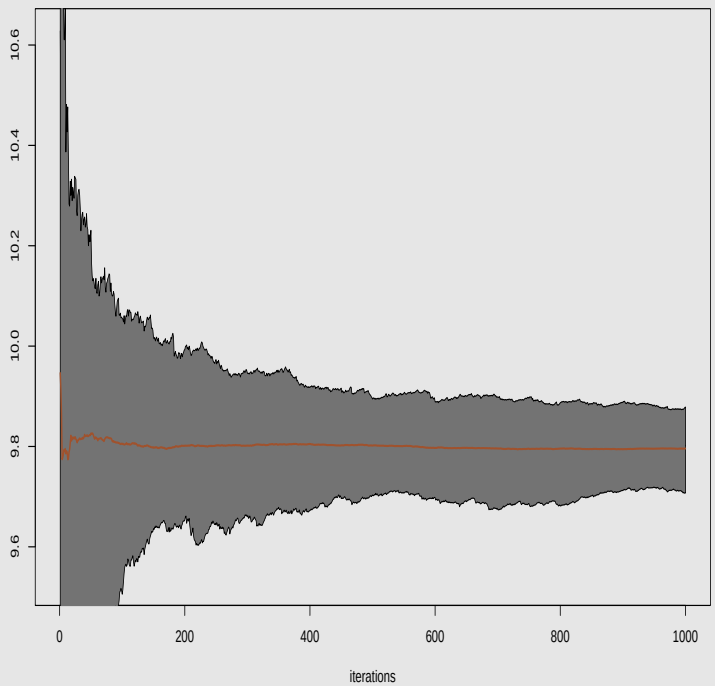
Form of δ^π suggests simulating iid variables $\theta_1, \cdots, \theta_m \sim \mathcal{N}(x, 1)$ and calculate

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

Example (Cauchy-normal (2))

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

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



LLN implies

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

Importance sampling

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

Importance sampling

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

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

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

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

Importance sampling (cont'd)

Importance sampling algorithm

Evaluation of

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

by

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

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

Justification

Convergence of the estimator

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

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

Justification

Convergence of the estimator

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

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

Justification

Convergence of the estimator

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

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

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

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

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

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

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

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

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

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

Choice of importance function

g can be any density but some choices better than others

1. Finite variance only when

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

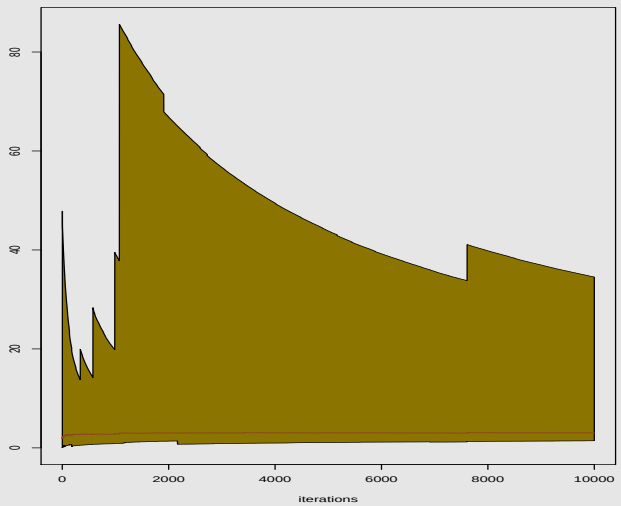
2. Instrumental distributions with tails lighter than those of f (that is, with $\sup f/g = \infty$) not appropriate, because weights $f(x_j)/g(x_j)$ vary widely, giving too much importance to a few values x_j.
3. If $\sup f/g = M < \infty$, the accept-reject algorithm can be used as well to simulate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

Example (Cauchy target)

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

Density ratio

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



very badly behaved: e.g.,

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

Poor performances of the associated importance sampling estimator

Practical alternative

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

where f and g are known up to constants.

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

Example (Student's t distribution)

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

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

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

Example (Student's t distribution)

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

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

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

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

Example (Student's t distribution (2))

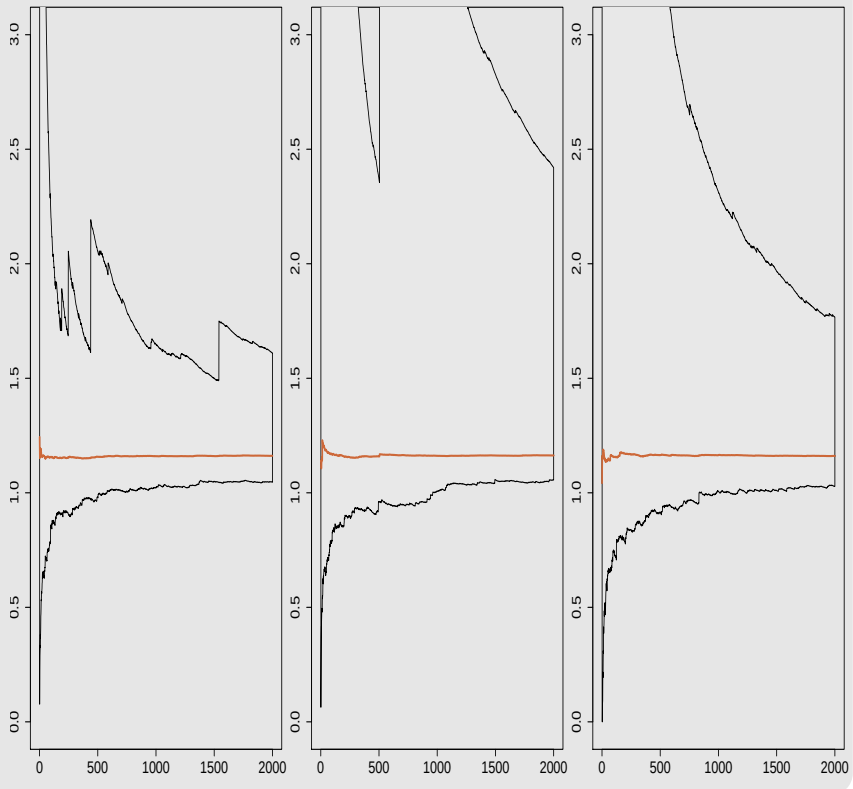
Choices of h:

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

Note: The ratio

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

does not have a finite integral



Explanation

Example (Student's t distribution (3))

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

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

Explanation

Example (Student's t distribution (3))

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

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

Consequence: the three estimators have infinite variance

Alternative

Example (Student's t distribution (4))

Choose a better behaved h:

Alternative

Example (Student's t distribution (4))

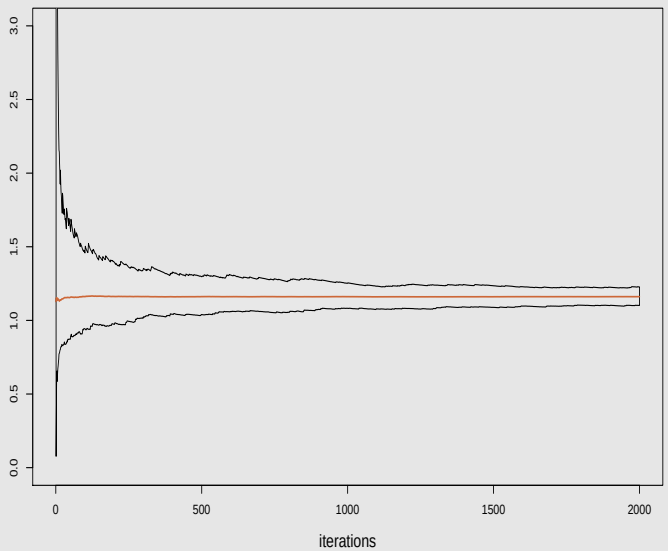
Choose a better behaved h: folded Gamma distribution, x symmetric around 1 with

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

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

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

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



Choice of importance function (termin'd)

The importance function may be π

Choice of importance function (termin'd)

The importance function may be π

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

Choice of importance function (termin'd)

The importance function may be π

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

Defensive sampling:

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

Example (Cauchy/Normal)

Consider

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

with known hyperparameters μ and σ^2.

Example (Cauchy/Normal)

Consider

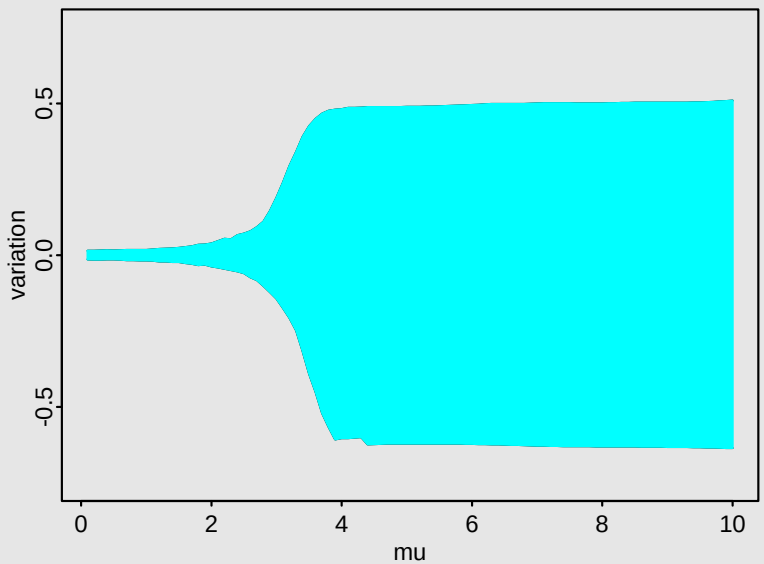
$$x_1, \ldots, x_n \sim \mathcal{C}(\theta, 1) \qquad \text{and} \qquad \theta \sim \mathcal{N}(\mu, \sigma^2),$$

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

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

Example (Cauchy/Normal (2))

Poor when the x_i's are all far from μ



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

Bridge sampling

Bayes factor

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

Bridge sampling

Bayes factor

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

If

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

then

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

Approximating evidence from a posterior sample

Use of the [harmonic mean] identity

$$\mathbb{E}^{\pi}\left[\left.\frac{\varphi(\theta)}{\pi(\theta)\ell(\theta)}\right| x\right] = \int \frac{\varphi(\theta)}{\pi(\theta)\ell(\theta)} \frac{\pi(\theta)\ell(\theta)}{m(x)}\,\mathrm{d}\theta = \frac{1}{m(x)}$$

no matter what the proposal $\varphi(\cdot)$ is.

[Gelfand & Dey, 1994; Bartolucci et al., 2006]

Approximating evidence from a posterior sample

Use of the [harmonic mean] identity

$$\mathbb{E}^{\pi}\left[\left.\frac{\varphi(\theta)}{\pi(\theta)\ell(\theta)}\right| x\right] = \int \frac{\varphi(\theta)}{\pi(\theta)\ell(\theta)} \frac{\pi(\theta)\ell(\theta)}{m(x)}\,\mathrm{d}\theta = \frac{1}{m(x)}$$

no matter what the proposal $\varphi(\cdot)$ is.

[Gelfand & Dey, 1994; Bartolucci et al., 2006]

Direct exploitation of the MC(MC) output from the posterior

Comparison with regular importance sampling

Harmonic mean: Constraint opposed to usual importance sampling constraints: $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi(\theta)\ell(\theta)$ for the approximation

$$\widehat{m(x)} = 1 \Bigg/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta^{(t)})}{\pi(\theta^{(t)})\ell(\theta^{(t)})}$$

to have a finite variance.

Comparison with regular importance sampling

Harmonic mean: Constraint opposed to usual importance sampling constraints: $\varphi(\theta)$ must have lighter (rather than fatter) tails than $\pi(\theta)\ell(\theta)$ for the approximation

$$\widehat{m(x)} = 1 \Big/ \frac{1}{T} \sum_{t=1}^{T} \frac{\varphi(\theta^{(t)})}{\pi(\theta^{(t)})\ell(\theta^{(t)})}$$

to have a finite variance.
E.g., use finite support kernels (like Epanechnikov's kernel) for φ

Prediction

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

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

Normal prediction

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

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

corresponding posterior

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

Normal prediction

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

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

corresponding posterior

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

Notation

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

Normal prediction (cont'd)

Predictive on x_{n+1}

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

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

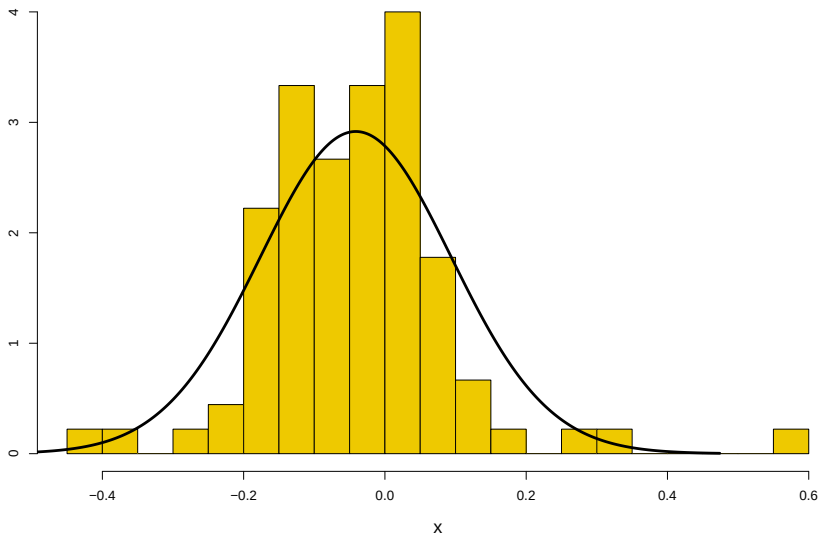
normaldata

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

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

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

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



Regression and variable selection

Regression

Large fraction of statistical analyses dealing with representation of dependences between several variables, rather than marginal distribution of each variable

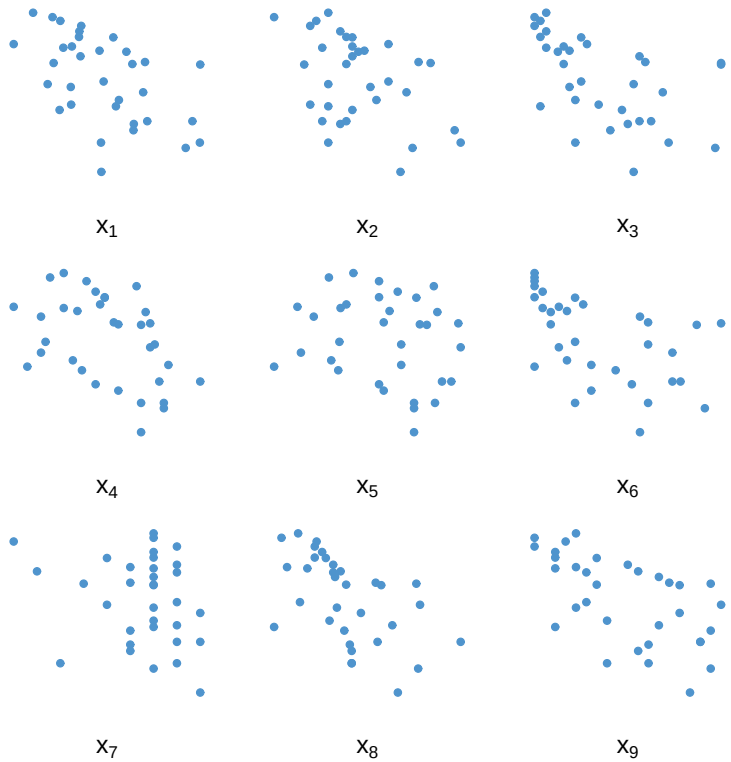
Pine processionary caterpillars

Pine processionary caterpillars

Pine processionary caterpillar colony size influenced by

- x_1 altitude
- x_2 slope (in degrees)
- x_3 number of pines in the area
- x_4 height of the central tree
- x_5 diameter of the central tree
- x_6 index of the settlement density
- x_7 orientation of the area (from 1 [southbound] to 2)
- x_8 height of the dominant tree
- x_9 number of vegetation strata
- x_{10} mix settlement index (from 1 if not mixed to 2 if mixed)

Pine processionary caterpillars



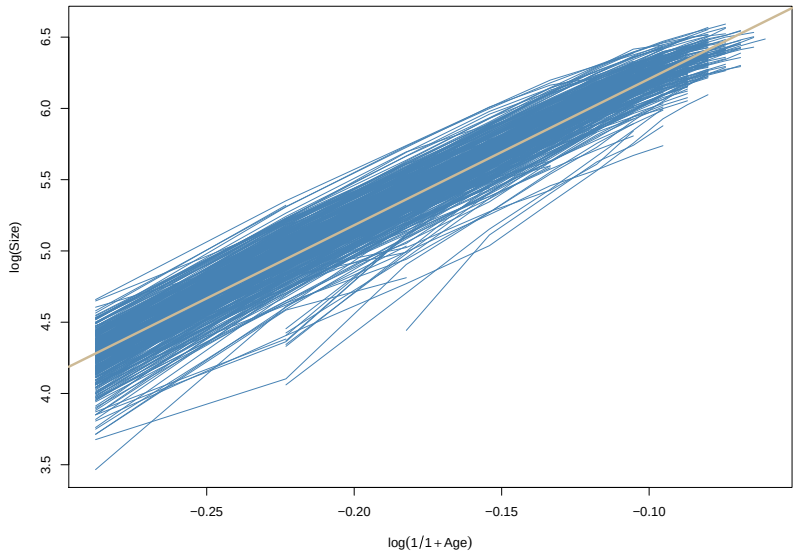
Ibex horn growth

Ibex horn growth

Growth of the Ibex horn size S against age A

$$\log(S) = \alpha + \beta \log \frac{A}{1+A} + \epsilon$$

Ibex horn growth



Goal of a regression model

From a statistical point of view, find a proper representation of the distribution, $f(y|\theta, x)$, of an observable variable y given a vector of observables x, based on a sample of $(x, y)_i$'s.

Linear regression

Linear regression: one of the most widespread tools of Statistics for analysing (linear) influence of some variables or some factors on others

Linear regression

Linear regression: one of the most widespread tools of Statistics for analysing (linear) influence of some variables or some factors on others

Aim

To uncover explanatory and predictive patterns

Regressors and response

Variable of primary interest, y, called the *response* or the *outcome* variable [assumed here to be continuous]

E.g., number of Pine processionary caterpillar colonies

Regressors and response

Variable of primary interest, y, called the *response* or the *outcome* variable [assumed here to be continuous]

E.g., number of Pine processionary caterpillar colonies

Covariates $x = (x_1, \ldots, x_k)$ called *explanatory variables* [may be discrete, continuous or both]

Regressors and response

Variable of primary interest, y, called the *response* or the *outcome* variable [assumed here to be continuous]

E.g., number of Pine processionary caterpillar colonies

Covariates $x = (x_1, \ldots, x_k)$ called *explanatory variables* [may be discrete, continuous or both]

Distribution of y given x typically studied in the context of a set of *units* or experimental *subjects*, $i = 1, \ldots, n$, for instance patients in an hospital ward, on which both y_i and $x_{i1}, \ldots, x_{ik}$ are measured.

Regressors and response cont'd

Dataset made of the conjunction of the vector of outcomes

$$y = (y_1, \ldots, y_n)$$

Regressors and response cont'd

Dataset made of the conjunction of the vector of outcomes

$$y = (y_1, \ldots, y_n)$$

and of the $n \times (k+1)$ matrix of explanatory variables

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ 1 & x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

Linear models

Ordinary normal linear regression model such that

$$y|\beta, \sigma^2, X \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

Linear models

Ordinary normal linear regression model such that

$$y|\beta, \sigma^2, X \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

and thus

$$\begin{aligned} \mathbb{E}[y_i|\beta, X] &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \\ \mathbb{V}(y_i|\sigma^2, X) &= \sigma^2 \end{aligned}$$

Categorical variables

- There is a difference between finite valued regressors like x_7 in caterpillar [orientation of the area] and *categorical* variables (or *factors*), which are also taking a finite number of values but whose range has no numerical meaning.

Categorical variables

- There is a difference between finite valued regressors like x_7 in caterpillar [orientation of the area] and *categorical* variables (or *factors*), which are also taking a finite number of values but whose range has no numerical meaning.

Example

If x is the socio-professional category of an employee, this variable ranges from 1 to 9 for a rough grid of socio-professional activities, and from 1 to 89 on a finer grid.

The numerical values are not comparable

Categorical variables (cont'd)

Makes little sense to involve x directly in the regression: replace the single regressor x [in $\{1, \ldots, m\}$, say] with m indicator (or *dummy*) variables

$$x_1 = \mathbb{I}_1(x), \ldots, x_m = \mathbb{I}_m(x)$$

Categorical variables (cont'd)

Makes little sense to involve x directly in the regression: replace the single regressor x [in $\{1, \ldots, m\}$, say] with m indicator (or *dummy*) variables

$$x_1 = \mathbb{I}_1(x), \ldots, x_m = \mathbb{I}_m(x)$$

Convention

Use of a different constant β_i for each class categorical variable value:

$$\mathbb{E}[y_i|\beta, X] = \ldots + \beta_1 \mathbb{I}_1(x) + \ldots + \beta_m \mathbb{I}_m(x) + \ldots$$

Identifiability

Identifiability issue: For dummy variables, sum of the indicators equal to one.

Convention

Assume that X is of full rank:

$$\text{rank}(X) = k + 1$$

[X is of full rank if and only if $X^{\mathrm{T}}X$ is invertible]

Identifiability

Identifiability issue: For dummy variables, sum of the indicators equal to one.

Convention

Assume that X is of full rank:

$$\text{rank}(X) = k + 1$$

[X is of full rank if and only if $X^{\mathrm{T}}X$ is invertible]

E.g., for dummy variables, this means eliminating one class

Likelihood function & estimator

The likelihood of the *ordinary normal linear model* is

$$\ell\left(\beta, \sigma^2 | y, X\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta)\right]$$

Likelihood function & estimator

The likelihood of the *ordinary normal linear model* is

$$\ell\left(\beta, \sigma^2 | y, X\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta)\right]$$

The MLE of β is solution of the least squares minimisation problem

$$\min_{\beta} (y - X\beta)^{\mathrm{T}}(y - X\beta) = \min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}\right)^2 ,$$

namely

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

Least square estimator

- $\hat{\beta}$ is an unbiased estimator of β.
- $\mathbb{V}(\hat{\beta}|\sigma^2, X) = \sigma^2(X^{\mathrm{T}}X)^{-1}$
- $\hat{\beta}$ is the *best* linear unbiased estimator of β: for all $a \in \mathbb{R}^{k+1}$,

$$\mathbb{V}(a^{\mathrm{T}}\hat{\beta}|\sigma^2, X) \leq \mathbb{V}(a^{\mathrm{T}}\tilde{\beta}|\sigma^2, X)$$

  for any unbiased linear estimator $\tilde{\beta}$ of β.
- Unbiased estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{n-k-1}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta}) = \frac{s^2}{n-k-1},$$

Pine processionary caterpillars

```
Residuals:
      Min       1Q   Median       3Q      Max
  -1.6989  -0.2731  -0.0003   0.3246   1.7305

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
intercept   10.998412   3.060272   3.594  0.00161 **
XV1         -0.004431   0.001557  -2.846  0.00939 **
XV2         -0.053830   0.021900  -2.458  0.02232 *
XV3          0.067939   0.099472   0.683  0.50174
XV4         -1.293636   0.563811  -2.294  0.03168 *
XV5          0.231637   0.104378   2.219  0.03709 *
XV6         -0.356800   1.566464  -0.228  0.82193
XV7         -0.237469   1.006006  -0.236  0.81558
XV8          0.181060   0.236724   0.765  0.45248
XV9         -1.285316   0.864847  -1.486  0.15142
XV10        -0.433106   0.734869  -0.589  0.56162
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ibex horn growth

```
> ibexX=cbind(log(ibex$accr/(1+ibex$accr)),ibex$sx,ibex$surv1,ibex$ril)
Residuals:
      Min         1Q     Median         3Q        Max
-1.020820 -0.060533  0.008867  0.075370  0.337181

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.0675133  0.0099228 712.250  < 2e-16 ***
ibexX1      11.0907326  0.0403489 274.870  < 2e-16 ***
ibexX2       0.0054787  0.0001487  36.840  < 2e-16 ***
ibexX3      -0.0418425  0.0073519  -5.691 1.38e-08 ***
ibexX4      -0.0003315  0.0018617  -0.178    0.859
---
Residual standard error: 0.1171 on 2954 degrees of freedom
  (83 observations deleted due to missingness)
Multiple R-Squared:  0.97,      Adjusted R-squared: 0.9699
F-statistic: 2.384e+04 on 4 and 2954 DF,  p-value: < 2.2e-16
```

Conjugate priors

If [conditional prior]

$$\beta|\sigma^2, X \sim \mathscr{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}) \,,$$

where M $(k+1, k+1)$ positive definite symmetric matrix, and

$$\sigma^2|X \sim \mathscr{IG}(a, b), \qquad a, b > 0,$$

Conjugate priors

If [conditional prior]

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, \sigma^2 M^{-1})\,,$$

where M $(k+1, k+1)$ positive definite symmetric matrix, and

$$\sigma^2|X \sim \mathcal{IG}(a, b), \qquad a, b > 0,$$

then

$$\beta|\sigma^2, \mathbf{y}, X \sim \mathcal{N}_{k+1}\left((M + X^{\mathrm{T}}X)^{-1}\{(X^{\mathrm{T}}X)\hat{\beta} + M\tilde{\beta}\}, \sigma^2(M + X^{\mathrm{T}}X)^{-1}\right)$$

and

$$\sigma^2|\mathbf{y}, X \sim \mathcal{IG}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \hat{\beta})^{\mathrm{T}}\left(M^{-1} + (X^{\mathrm{T}}X)^{-1}\right)^{-1}(\tilde{\beta} - \hat{\beta})}{2}\right)$$

Experimenter dilemma

Problem of the choice of M or of c if $M = I_{k+1}/c$

Experimenter dilemma

Problem of the choice of M or of c if $M = I_{k+1}/c$

Example (Processionary caterpillar)

No precise prior information about $\tilde{\beta}$, M, a and b. Take $a = 2.1$ and $b = 2$, i.e. prior mean and prior variance of σ^2 equal to 1.82 and 33.06, and $\tilde{\beta} = 0_{k+1}$.

Experimenter dilemma

Problem of the choice of M or of c if $M = I_{k+1}/c$

Example (Processionary caterpillar)

No precise prior information about $\tilde{\beta}$, M, a and b. Take $a = 2.1$ and $b = 2$, i.e. prior mean and prior variance of σ^2 equal to 1.82 and 33.06, and $\tilde{\beta} = 0_{k+1}$.
Lasting influence of c:

c	$\mathbb{E}^\pi(\sigma^2\|\mathbf{y}, X)$	$\mathbb{E}^\pi(\beta_0\|\mathbf{y}, X)$	$\mathbb{V}^\pi(\beta_0\|\mathbf{y}, X)$
.1	1.0044	0.1251	0.0988
1	0.8541	0.9031	0.7733
10	0.6976	4.7299	3.8991
100	0.5746	9.6626	6.8355
1000	0.5470	10.8476	7.3419

Zellner's informative G-prior

Constraint

Allow the experimenter to introduce information about the location parameter of the regression while bypassing the most difficult aspects of the prior specification, namely the derivation of the prior correlation structure.

Zellner's informative G-prior

Constraint

Allow the experimenter to introduce information about the location parameter of the regression while bypassing the most difficult aspects of the prior specification, namely the derivation of the prior correlation structure.

Zellner's prior corresponds to

$$\begin{aligned} \beta|\sigma^2, X &\sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^{\mathrm{T}}X)^{-1}) \\ \sigma^2 &\sim \pi(\sigma^2|X) \propto \sigma^{-2}\,. \end{aligned}$$

[Special conjugate]

Prior selection

Experimental prior determination restricted to the choices of $\tilde{\beta}$ and of the constant c.

Note

c can be interpreted as a measure of the amount of information available in the prior relative to the sample. For instance, setting $1/c = 0.5$ gives the prior the same weight as 50% of the sample.

Prior selection

Experimental prior determination restricted to the choices of $\tilde{\beta}$ and of the constant c.

Note

c can be interpreted as a measure of the amount of information available in the prior relative to the sample. For instance, setting $1/c = 0.5$ gives the prior the same weight as 50% of the sample.

↯ There still is a lasting influence of the factor c

Posterior structure

With this prior model, the posterior simplifies into

$$\begin{aligned}\pi(\beta,\sigma^2|y,X) &\propto f(y|\beta,\sigma^2,X)\pi(\beta,\sigma^2|X)\\ &\propto \left(\sigma^2\right)^{-(n/2+1)}\exp\left[-\frac{1}{2\sigma^2}(y-X\hat{\beta})^{\mathrm{T}}(y-X\hat{\beta})\right.\\ &\qquad\left.-\frac{1}{2\sigma^2}(\beta-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(\beta-\hat{\beta})\right]\left(\sigma^2\right)^{-k/2}\\ &\qquad\times\exp\left[-\frac{1}{2c\sigma^2}(\beta-\tilde{\beta})X^{\mathrm{T}}X(\beta-\tilde{\beta})\right],\end{aligned}$$

because $X^{\mathrm{T}}X$ used in both prior and likelihood

[G-prior trick]

Posterior structure (cont'd)

Therefore,

$$\begin{aligned}\beta|\sigma^2, y, X &\sim \mathcal{N}_{k+1}\left(\frac{c}{c+1}(\tilde{\beta}/c+\hat{\beta}), \frac{\sigma^2 c}{c+1}(X^{\mathrm{T}}X)^{-1}\right)\\ \sigma^2|y, X &\sim \mathcal{IG}\left(\frac{n}{2}, \frac{s^2}{2}+\frac{1}{2(c+1)}(\tilde{\beta}-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(\tilde{\beta}-\hat{\beta})\right)\end{aligned}$$

and

$$\begin{aligned}\beta|y, X \sim \mathscr{T}_{k+1}\Bigg(&n, \frac{c}{c+1}\left(\frac{\tilde{\beta}}{c}+\hat{\beta}\right),\\ &\frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}(X^{\mathrm{T}}X)^{-1}\Bigg).\end{aligned}$$

Bayes estimator

The Bayes estimators of β and σ^2 are given by

$$\mathbb{E}^{\pi}[\beta|y,X] = \frac{1}{c+1}(\tilde{\beta} + c\hat{\beta})$$

and

$$\mathbb{E}^{\pi}[\sigma^2|y,X] = \frac{s^2 + (\tilde{\beta}-\hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X (\tilde{\beta}-\hat{\beta})/(c+1)}{n-2}\,.$$

Bayes estimator

The Bayes estimators of β and σ^2 are given by

$$\mathbb{E}^{\pi}[\beta|y,X] = \frac{1}{c+1}(\tilde{\beta} + c\hat{\beta})$$

and

$$\mathbb{E}^{\pi}[\sigma^2|y,X] = \frac{s^2 + (\tilde{\beta}-\hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X (\tilde{\beta}-\hat{\beta})/(c+1)}{n-2}\,.$$

Note: Only when c goes to infinity does the influence of the prior vanish!

Pine processionary caterpillars

β_i	$\mathbb{E}^\pi(\beta_i\|\mathbf{y},X)$	$\mathbb{V}^\pi(\beta_i\|\mathbf{y},X)$
β_0	10.8895	6.4094
β_1	-0.0044	2e-06
β_2	-0.0533	0.0003
β_3	0.0673	0.0068
β_4	-1.2808	0.2175
β_5	0.2293	0.0075
β_6	-0.3532	1.6793
β_7	-0.2351	0.6926
β_8	0.1793	0.0383
β_9	-1.2726	0.5119
β_{10}	-0.4288	0.3696

$c = 100$

Pine processionary caterpillars (2)

β_i	$\mathbb{E}^\pi(\beta_i\|\mathbf{y},X)$	$\mathbb{V}^\pi(\beta_i\|\mathbf{y},X)$
β_0	10.9874	6.2604
β_1	-0.0044	2e-06
β_2	-0.0538	0.0003
β_3	0.0679	0.0066
β_4	-1.2923	0.2125
β_5	0.2314	0.0073
β_6	-0.3564	1.6403
β_7	-0.2372	0.6765
β_8	0.1809	0.0375
β_9	-1.2840	0.5100
β_{10}	-0.4327	0.3670

$c = 1,000$

Conjugacy

Moreover,

$$\mathbb{V}^{\pi}[\beta|y,X] = \frac{c(s^2 + (\tilde{\beta}-\hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X (\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)} (X^{\mathrm{T}}X)^{-1}\,.$$

Conjugacy

Moreover,

$$\mathbb{V}^{\pi}[\beta|y,X] = \frac{c(s^2 + (\tilde{\beta}-\hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)} (X^{\mathrm{T}}X)^{-1}\,.$$

Convenient tool for translating prior information on β: For instance, if $c = 1$, this is equivalent to putting the same weight on the prior information and on the sample:

$$\mathbb{E}^{\pi}(\beta|y,X) = \left(\frac{\tilde{\beta}+\hat{\beta}}{2}\right)$$

average between prior mean and maximum likelihood estimator.

Conjugacy

Moreover,

$$\mathbb{V}^{\pi}[\beta|y,X] = \frac{c(s^2 + (\tilde{\beta}-\hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)} (X^{\mathrm{T}}X)^{-1}\,.$$

Convenient tool for translating prior information on β: For instance, if $c = 1$, this is equivalent to putting the same weight on the prior information and on the sample:

$$\mathbb{E}^{\pi}(\beta|y,X) = \left(\frac{\tilde{\beta}+\hat{\beta}}{2}\right)$$

average between prior mean and maximum likelihood estimator. If, instead, $c = 100$, the prior gets a weight of 1% of the sample.

Predictive

Prediction of $m \geq 1$ future observations from units in which the explanatory variables $\tilde{X}$—but not the outcome variable

$$\tilde{y} \sim \mathcal{N}_m(\tilde{X}\beta, \sigma^2 I_m)$$

—have been observed

Predictive

Prediction of $m \geq 1$ future observations from units in which the explanatory variables $\tilde{X}$—but not the outcome variable

$$\tilde{y} \sim \mathcal{N}_m(\tilde{X}\beta, \sigma^2 I_m)$$

—have been observed

Predictive distribution on $\tilde{y}$ defined as marginal of the joint posterior distribution on $(\tilde{y}, \beta, \sigma^2)$. Can be computed analytically by

$$\int \pi(\tilde{y}|\sigma^2, y, X, \tilde{X})\pi(\sigma^2|y, X, \tilde{X})\,\mathrm{d}\sigma^2\,.$$

Gaussian predictive

Conditional on σ^2, the future vector of observations has a Gaussian distribution with

$$\begin{aligned}\mathbb{E}^\pi[\tilde{y}|\sigma^2,y,X,\tilde{X}] &= \mathbb{E}^\pi[\mathbb{E}^\pi(\tilde{y}|\beta,\sigma^2,y,X,\tilde{X})|\sigma^2,y,X,\tilde{X}]\\ &= \mathbb{E}^\pi[\tilde{X}\beta|\sigma^2,y,X,\tilde{X}]\\ &= \tilde{X}\frac{\tilde{\beta}+c\hat{\beta}}{c+1}\end{aligned}$$

independently of σ^2.

Gaussian predictive

Conditional on σ^2, the future vector of observations has a Gaussian distribution with

$$\begin{aligned}\mathbb{E}^\pi[\tilde{y}|\sigma^2,y,X,\tilde{X}] &= \mathbb{E}^\pi[\mathbb{E}^\pi(\tilde{y}|\beta,\sigma^2,y,X,\tilde{X})|\sigma^2,y,X,\tilde{X}]\\ &= \mathbb{E}^\pi[\tilde{X}\beta|\sigma^2,y,X,\tilde{X}]\\ &= \tilde{X}\frac{\tilde{\beta}+c\hat{\beta}}{c+1}\end{aligned}$$

independently of σ^2. Similarly,

$$\begin{aligned}\mathbb{V}^\pi(\tilde{y}|\sigma^2,y,X,\tilde{X}) &= \mathbb{E}^\pi[\mathbb{V}(\tilde{y}|\beta,\sigma^2,y,X,\tilde{X})|\sigma^2,y,X,\tilde{X}]\\ &\quad +\mathbb{V}^\pi[\mathbb{E}^\pi(\tilde{y}|\beta,\sigma^2,y,X,\tilde{X})|\sigma^2,y,X,\tilde{X}]\\ &= \mathbb{E}^\pi[\sigma^2 I_m|\sigma^2,y,X,\tilde{X}]+\mathbb{V}^\pi(\tilde{X}\beta|\sigma^2,y,X,\tilde{X})\\ &= \sigma^2\left(I_m+\frac{c}{c+1}\tilde{X}(X^{\mathrm{T}}X)^{-1}\tilde{X}^{\mathrm{T}}\right)\end{aligned}$$

Predictor

A predictor under squared error loss is the posterior predictive mean

$$\tilde{X}\frac{\tilde{\beta}+c\hat{\beta}}{c+1},$$

Representation quite intuitive, being the product of the matrix of explanatory variables $\tilde{X}$ by the Bayes estimate of β.

Credible regions

Highest posterior density (HPD) regions on subvectors of the parameter β derived from the marginal posterior distribution of β.

Credible regions

Highest posterior density (HPD) regions on subvectors of the parameter β derived from the marginal posterior distribution of β. For a single parameter,

$$\beta_i|y,X \sim \mathscr{T}_1\left(n, \frac{c}{c+1}\left(\frac{\tilde{\beta}_i}{c}+\hat{\beta}_i\right), \frac{c(s^2+(\tilde{\beta}-\hat{\beta})^{\mathrm{T}}X^{\mathrm{T}}X(\tilde{\beta}-\hat{\beta})/(c+1))}{n(c+1)}\omega_{(i,i)}\right),$$

where $\omega_{(i,i)}$ is the (i,i)-th element of the matrix $(X^{\mathrm{T}}X)^{-1}$.

T time

If

$$\tau = \frac{\tilde{\beta} + c\hat{\beta}}{c+1}$$

and

$$K = \frac{c(s^2 + (\tilde{\beta} - \hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X (\tilde{\beta} - \hat{\beta})/(c+1))}{n(c+1)} (X^{\mathrm{T}} X)^{-1} = \left(\kappa_{(i,j)}\right) ,$$

the transform

$$\mathfrak{T}_i = \frac{\beta_i - \tau_i}{\sqrt{\kappa_{(i,i)}}}$$

has a standard t distribution with n degrees of freedom.

T HPD

A $1-\alpha$ HPD interval on β_i is thus given by

$$\left[\tau_i - \sqrt{\kappa_{(i,i)}} F_n^{-1}(1-\alpha/2), \tau_i + \sqrt{\kappa_{(i,i)}} F_n^{-1}(1-\alpha/2)\right].$$

Pine processionary caterpillars

β_i	HPD interval
β_0	$[5.7435, 16.2533]$
β_1	$[-0.0071, -0.0018]$
β_2	$[-0.0914, -0.0162]$
β_3	$[-0.1029, 0.2387]$
β_4	$[-2.2618, -0.3255]$
β_5	$[0.0524, 0.4109]$
β_6	$[-3.0466, 2.3330]$
β_7	$[-1.9649, 1.4900]$
β_8	$[-0.2254, 0.5875]$
β_9	$[-2.7704, 0.1997]$
β_{10}	$[-1.6950, 0.8288]$

$$c = 100$$

T marginal

Marginal distribution of y is multivariate t distribution

Proof. Since $\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^{\mathrm{T}}X)^{-1})$,

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

which implies that

$$y|\sigma^2, X \sim \mathcal{N}_n(X\tilde{\beta}, \sigma^2(I_n + cX(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})).$$

Integrating in σ^2 yields

$$\begin{aligned} f(y|X) &= (c+1)^{-(k+1)/2}\pi^{-n/2}\Gamma(n/2) \\ &\quad \times \left[y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y - \frac{1}{c+1}\tilde{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\tilde{\beta}\right]^{-n/2}. \end{aligned}$$

Point null hypothesis

If a null hypothesis is $H_0 : R\beta = r$, the model under H_0 can be rewritten as

$$y|\beta^0, \sigma^2, X_0 \overset{H_0}{\sim} \mathcal{N}_n\left(X_0\beta^0, \sigma^2 I_n\right)$$

where β^0 is $(k+1-q)$ dimensional.

Point null marginal

Under the prior

$$\beta^0|X_0,\sigma^2 \sim \mathcal{N}_{k+1-q}\left(\tilde{\beta}^0, c_0\sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right) ,$$

the marginal distribution of y under H_0 is

$$\begin{aligned} f(y|X_0,H_0) &= (c_0+1)^{-(k+1-q)/2}\pi^{-n/2}\Gamma(n/2) \\ &\quad \times\left[y^{\mathrm{T}}y - \frac{c_0}{c_0+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y \right. \\ &\quad \left. -\frac{1}{c_0+1}\tilde{\beta}_0^{\mathrm{T}}X_0^{\mathrm{T}}X_0\tilde{\beta}_0\right]^{-n/2} . \end{aligned}$$

Bayes factor

Therefore the Bayes factor is closed form:

$$B^{\pi}_{10} = \frac{f(y|X,H_1)}{f(y|X_0,H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}} \left[\frac{y^{\mathrm{T}}y - \frac{c_0}{c_0+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y - \frac{1}{c_0+1}\tilde{\beta}_0^{\mathrm{T}}X_0^{\mathrm{T}}X_0\tilde{\beta}_0}{y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y - \frac{1}{c+1}\tilde{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\tilde{\beta}}\right]^{n/2}$$

Bayes factor

Therefore the Bayes factor is closed form:

$$B^{\pi}_{10} = \frac{f(y|X,H_1)}{f(y|X_0,H_0)} = \frac{(c_0+1)^{(k+1-q)/2}}{(c+1)^{(k+1)/2}} \left[\frac{y^{\mathrm{T}}y - \frac{c_0}{c_0+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y - \frac{1}{c_0+1}\tilde{\beta}_0^{\mathrm{T}}X_0^{\mathrm{T}}X_0\tilde{\beta}_0}{y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y - \frac{1}{c+1}\tilde{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\tilde{\beta}}\right]^{n/2}$$

- Means using the *same* σ^2 on both models
- Still depends on the choice of (c_0, c)

Zellner's noninformative G-prior

Difference with informative G-prior setup is that we now consider c as unknown (relief!)

Zellner's noninformative G-prior

Difference with informative G-prior setup is that we now consider c as unknown (relief!)

Solution

Use the same G-prior distribution with $\tilde{\beta} = 0_{k+1}$, conditional on c, and introduce a diffuse prior on c,

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

Posterior distribution

Corresponding marginal posterior on the parameters of interest

$$\begin{aligned}\pi(\beta,\sigma^2|y,X) &= \int \pi(\beta,\sigma^2|y,X,c)\pi(c|y,X)\,\mathrm{d}c \\ &\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c)f(y|X,c)\pi(c) \\ &\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c)f(y|X,c)\,c^{-1}\,.\end{aligned}$$

Posterior distribution

Corresponding marginal posterior on the parameters of interest

$$\begin{aligned}\pi(\beta,\sigma^2|y,X) &= \int \pi(\beta,\sigma^2|y,X,c)\pi(c|y,X)\,\mathrm{d}c \\ &\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c)f(y|X,c)\pi(c) \\ &\propto \sum_{c=1}^{\infty} \pi(\beta,\sigma^2|y,X,c)f(y|X,c)\,c^{-1}\,.\end{aligned}$$

and

$$f(y|X,c) \propto (c+1)^{-(k+1)/2}\left[y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\right]^{-n/2}.$$

Posterior means

The Bayes estimates of β and σ^2 are given by

$$\begin{aligned}\mathbb{E}^\pi[\beta|y,X] &= \mathbb{E}^\pi[\mathbb{E}^\pi(\beta|y,X,c)|y,X] = \mathbb{E}^\pi[c/(c+1)\hat{\beta})|y,X] \\ &= \left(\frac{\sum_{c=1}^{\infty} c/(c+1) f(y|X,c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X,c)c^{-1}}\right)\hat{\beta}\end{aligned}$$

and

$$\mathbb{E}^\pi[\sigma^2|y,X] = \frac{\sum_{c=1}^{\infty} \dfrac{s^2 + \hat{\beta}^{\mathrm{T}} X^{\mathrm{T}} X \hat{\beta}/(c+1)}{n-2} f(y|X,c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X,c)c^{-1}}\,.$$

Computational details

- Both terms involve infinite summations on c
- The denominator in both cases is the normalising constant of the posterior

$$\sum_{c=1}^{\infty} f(y|X,c)c^{-1}$$

Computational details (cont'd)

$$
\begin{aligned}
\mathbb{V}^\pi[\beta|y,X] &= \mathbb{E}^\pi[\mathbb{V}^\pi(\beta|y,X,c)|y,X] + \mathbb{V}^\pi[\mathbb{E}^\pi(\beta|y,X,c)|y,Xx] \\
&= \mathbb{E}^\pi\left[c/(n(c+1))(s^2+\hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}/(c+1))(X^{\mathrm{T}}X)^{-1}\right] \\
&\qquad +\mathbb{V}^\pi[c/(c+1)\hat{\beta}|y,X] \\
&= \left[\frac{\sum_{c=1}^{\infty} f(y|X,c)/(n(c+1))(s^2+\hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}/(c+1))}{\sum_{c=1}^{\infty} f(y|X,c)c^{-1}}\right](X^{\mathrm{T}}X)^{-1} \\
&\quad +\hat{\beta}\left(\frac{\sum_{c=1}^{\infty}(c/(c+1)-\mathbb{E}(c/(c+1)|y,X))^2 f(y|X,c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X,c)c^{-1}}\right)\hat{\beta}^{\mathrm{T}}\,.
\end{aligned}
$$

Marginal distribution

Important point: the marginal distribution of the dataset is available in closed form

$$f(y|X) \propto \sum_{i=1}^{\infty} c^{-1}(c+1)^{-(k+1)/2} \left[y^{\mathrm{T}} y - \frac{c}{c+1} y^{\mathrm{T}} X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \right]^{-n/2}$$

Marginal distribution

Important point: the marginal distribution of the dataset is available in closed form

$$f(y|X) \propto \sum_{i=1}^{\infty} c^{-1}(c+1)^{-(k+1)/2} \left[y^{\mathrm{T}} y - \frac{c}{c+1} y^{\mathrm{T}} X (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y \right]^{-n/2}$$

$\mathcal{T}$-shape means normalising constant can be computed too.

Point null hypothesis

For null hypothesis $H_0 : R\beta = r$, the model under H_0 can be rewritten as

$$y|\beta^0, \sigma^2, X_0 \overset{H_0}{\sim} \mathcal{N}_n\left(X_0\beta^0, \sigma^2 I_n\right)$$

where β^0 is $(k + 1 - q)$ dimensional.

Point null marginal

Under the prior

$$\beta^0|X_0,\sigma^2,c \sim \mathcal{N}_{k+1-q}\left(0_{k+1-q}, c\sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right)$$

and $\pi(c) = 1/c$, the marginal distribution of y under H_0 is

$$f(y|X_0,H_0) \propto \sum_{c=1}^{\infty}(c+1)^{-(k+1-q)/2}\left[y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y\right]^{-n/2}.$$

Point null marginal

Under the prior

$$\beta^0|X_0,\sigma^2,c \sim \mathcal{N}_{k+1-q}\left(0_{k+1-q}, c\sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right)$$

and $\pi(c) = 1/c$, the marginal distribution of y under H_0 is

$$f(y|X_0,H_0) \propto \sum_{c=1}^{\infty}(c+1)^{-(k+1-q)/2}\left[y^{\mathrm{T}}y - \frac{c}{c+1}y^{\mathrm{T}}X_0(X_0^{\mathrm{T}}X_0)^{-1}X_0^{\mathrm{T}}y\right]^{-n/2}.$$

Bayes factor $B_{10}^{\pi} = f(\mathbf{y}|X)/f(\mathbf{y}|X_0,H_0)$ can be computed

Processionary pine caterpillars

For $H_0 : \beta_8 = \beta_9 = 0$, $\log_{10}(B^{\pi}_{10}) = -0.7884$

Processionary pine caterpillars

For $H_0 : \beta_8 = \beta_9 = 0$, $\log_{10}(B^{\pi}_{10}) = -0.7884$

	Estimate	Post. Var.	log10(BF)	
(Intercept)	9.2714	9.1164	1.4205	(***)
X1	-0.0037	2e-06	0.8502	(**)
X2	-0.0454	0.0004	0.5664	(**)
X3	0.0573	0.0086	-0.3609	
X4	-1.0905	0.2901	0.4520	(*)
X5	0.1953	0.0099	0.4007	(*)
X6	-0.3008	2.1372	-0.4412	
X7	-0.2002	0.8815	-0.4404	
X8	0.1526	0.0490	-0.3383	
X9	-1.0835	0.6643	-0.0424	
X10	-0.3651	0.4716	-0.3838	

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

Markov Chain Monte Carlo Methods

Complexity of most models encountered in Bayesian modelling

Markov Chain Monte Carlo Methods

Complexity of most models encountered in Bayesian modelling

Standard simulation methods not good enough a solution

Markov Chain Monte Carlo Methods

Complexity of most models encountered in Bayesian modelling

Standard simulation methods not good enough a solution

New technique at the core of Bayesian computing, based on *Markov chains*

Markov chains

Markov chain

A process $(\theta^{(t)})_{t\in\mathbb{N}}$ is an *homogeneous Markov chain* if the distribution of $\theta^{(t)}$ given the past $(\theta^{(0)},\ldots,\theta^{(t-1)})$

1. only depends on $\theta^{(t-1)}$
2. is the same for all $t \in \mathbb{N}^*$.

Algorithms based on Markov chains

Idea: simulate from a posterior density $\pi(\cdot|x)$ [or any density] by producing a Markov chain

$$(\theta^{(t)})_{t\in\mathbb{N}}$$

whose stationary distribution is

$$\pi(\cdot|x)$$

Algorithms based on Markov chains

Idea: simulate from a posterior density $\pi(\cdot|x)$ [or any density] by producing a Markov chain

$$(\theta^{(t)})_{t\in\mathbb{N}}$$

whose stationary distribution is

$$\pi(\cdot|x)$$

Translation

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

Convergence

If an algorithm that generates such a chain can be constructed, the ergodic theorem guarantees that, in almost all settings, the average

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)})$$

converges to $\mathbb{E}^{\pi}[g(\theta)|x]$, for (almost) any starting value

More convergence

If the produced Markov chains are irreducible [can reach any region in a finite number of steps], then they are both positive recurrent with stationary distribution $\pi(\cdot|x)$ *and* ergodic [asymptotically independent from the starting value $\theta^{(0)}$]

- While, for t large enough, $\theta^{(t)}$ is approximately distributed from $\pi(\theta|x)$ and can thus be used like the output from a more standard simulation algorithm, one must take care of the correlations between the $\theta^{(t)}$'s

Demarginalising

Takes advantage of *hierarchical structures*: if

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

simulating from $\pi(\theta|x)$ comes from simulating from the joint distribution

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

Two-stage Gibbs sampler

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

Two-stage Gibbs sampler

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

More often, both *conditional posterior distributions*,

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

can be simulated.

Two-stage Gibbs sampler

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

More often, both *conditional posterior distributions*,

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

can be simulated.

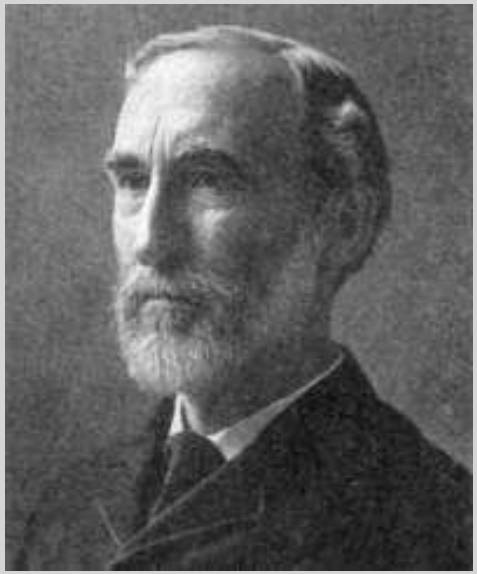
Idea: Create a Markov chain based on those conditionals

Two-stage Gibbs sampler (cont'd)

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 $\pi_2(\lambda|x, \theta^{(t)})$



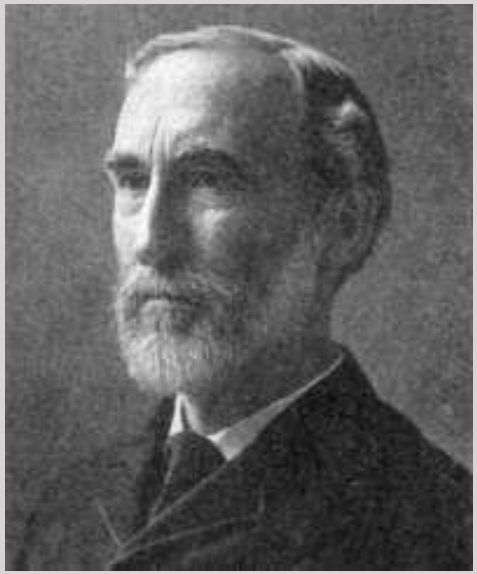
J.W. Gibbs (1839-1903)

Two-stage Gibbs sampler (cont'd)

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 $\pi_2(\lambda|x, \theta^{(t)})$



J.W. Gibbs (1839-1903)

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

Implementation

1. Derive efficient decomposition of the joint distribution into simulable conditionals (mixing behavior, acf(), blocking, &tc.)

Implementation

1. Derive efficient decomposition of the joint distribution into simulable conditionals (mixing behavior, acf(), blocking, &tc.)

2. Find when to stop the algorithm (mode chasing, missing mass, shortcuts, &tc.)

Simple Example: iid $\mathcal{N}(\mu, \sigma^2)$ Observations

When $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate outside a standard family

Simple Example: iid $\mathcal{N}(\mu, \sigma^2)$ Observations

When $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both μ and σ unknown, the posterior in (μ, σ^2) is conjugate outside a standard family

But...

$$\mu|\mathbf{y}, \sigma^2 \sim \mathcal{N}\left(\mu \left| \tfrac{1}{n}\sum_{i=1}^n y_i, \tfrac{\sigma^2}{n}\right.\right)$$
$$\sigma^2|\mathbf{y}, \mu \sim \mathcal{IG}\left(\sigma^2 \left| \tfrac{n}{2} - 1, \tfrac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right.\right)$$

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)

Gibbs output analysis

Example (Cauchy posterior)

$$\pi(\mu|\mathscr{D}) \propto \frac{e^{-\mu^2/20}}{(1+(x_1-\mu)^2)(1+(x_2-\mu)^2)}$$

is marginal of

$$\pi(\mu, \boldsymbol{\omega}|\mathscr{D}) \propto e^{-\mu^2/20} \times \prod_{i=1}^{2} e^{-\omega_i[1+(x_i-\mu)^2]}\,.$$

Gibbs output analysis

Example (Cauchy posterior)

$$\pi(\mu|\mathscr{D}) \propto \frac{e^{-\mu^2/20}}{(1+(x_1-\mu)^2)(1+(x_2-\mu)^2)}$$

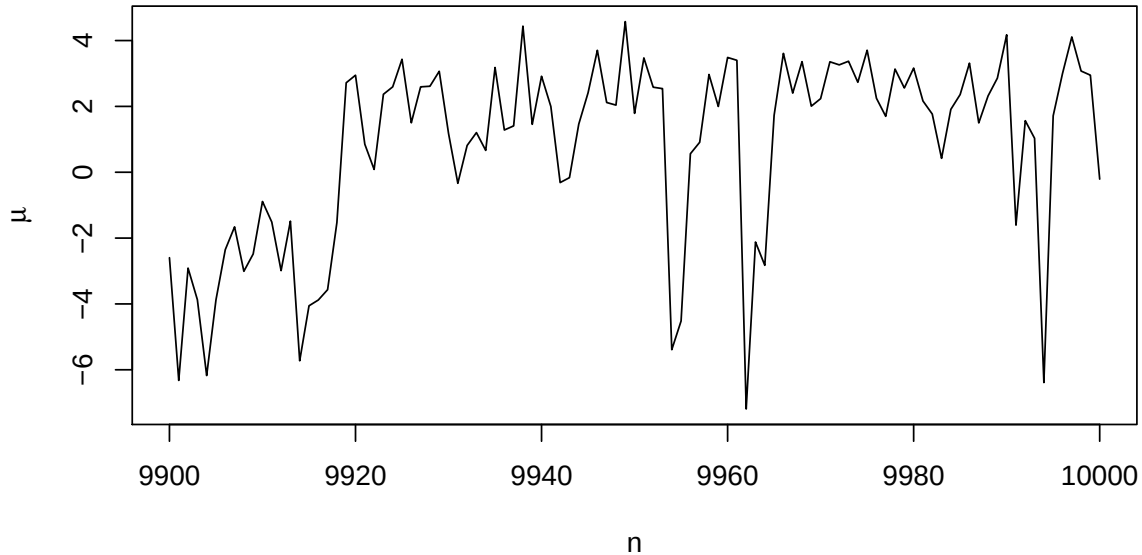
is marginal of

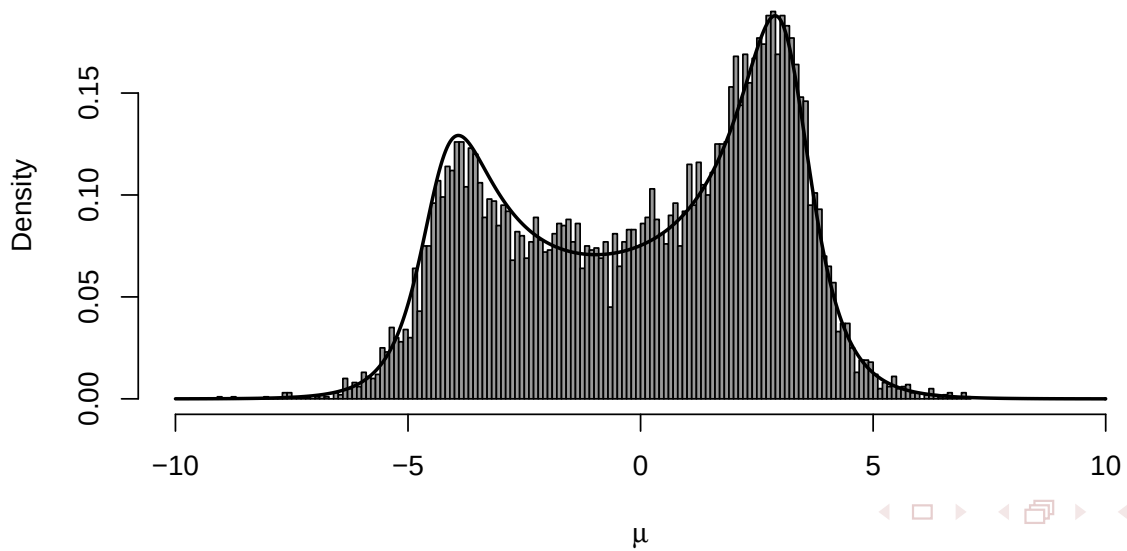
$$\pi(\mu,\boldsymbol{\omega}|\mathscr{D}) \propto e^{-\mu^2/20} \times \prod_{i=1}^{2} e^{-\omega_i[1+(x_i-\mu)^2]}\,.$$

Corresponding conditionals

$$\begin{aligned}(\omega_1,\omega_2)|\mu &\sim \mathscr{E}xp(1+(x_1-\mu)^2) \otimes \mathscr{E}xp(1+(x_2-\mu))^2)\\ \mu|\boldsymbol{\omega} &\sim \mathscr{N}\left(\sum_i \omega_i x_i/(\sum_i \omega_i + 1/20), 1/(2\sum_i \omega_i + 1/10)\right)\end{aligned}$$

Gibbs output analysis (cont'd)





Generalisation

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

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

The general Gibbs sampler

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

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

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

$\vdots$

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

Then $\theta^{(t)} \longrightarrow \theta \sim \pi$

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Variable selection

Back to regression: one dependent random variable y and a set $\{x_1, \ldots, x_k\}$ of k explanatory variables.

Question: Are all x_i's involved in the regression?

Assumption: every subset $\{i_1, \ldots, i_q\}$ of q $(0 \leq q \leq k)$ explanatory variables, $\{\mathbf{1}_n, x_{i_1}, \ldots, x_{i_q}\}$, is a proper set of explanatory variables for the regression of y [intercept included in every corresponding model]

Computational issue

2^k models in competition...

Model notations

1. $$X = \begin{bmatrix} \mathbf{1}_n & x_1 & \cdots & x_k \end{bmatrix}$$
   is the matrix containing $\mathbf{1}_n$ and all the k potential predictor variables
2. Each model $\mathfrak{M}_\gamma$ associated with binary indicator vector $\gamma \in \Gamma = \{0,1\}^k$ where $\gamma_i = 1$ means that the variable x_i is included in the model $\mathfrak{M}_\gamma$
3. $q_\gamma = \mathbf{1}_n^{\mathrm{T}}\gamma$ number of variables included in the model $\mathfrak{M}_\gamma$
4. $t_1(\gamma)$ and $t_0(\gamma)$ indices of variables included in the model and indices of variables not included in the model

Model indicators

For $\beta \in \mathbb{R}^{k+1}$ and X, we define β_γ as the subvector

$$\beta_\gamma = \left(\beta_0, (\beta_i)_{i \in t_1(\gamma)}\right)$$

and X_γ as the submatrix of X where only the column $\mathbf{1}_n$ and the columns in $t_1(\gamma)$ have been left.

Models in competition

The model $\mathfrak{M}_\gamma$ is thus defined as

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

where $\beta_\gamma \in \mathbb{R}^{q_\gamma+1}$ and $\sigma^2 \in \mathbb{R}_+^*$ are the unknown parameters.

Models in competition

The model $\mathfrak{M}_\gamma$ is thus defined as

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

where $\beta_\gamma \in \mathbb{R}^{q_\gamma+1}$ and $\sigma^2 \in \mathbb{R}_+^*$ are the unknown parameters.

Warning

σ^2 is common to all models and thus uses the same prior for all models

Informative G-prior

Many (2^k) models in competition: we cannot expect a practitioner to specify a prior on every $\mathfrak{M}_\gamma$ in a completely subjective and autonomous manner.

Shortcut: We derive *all* priors from a single global prior associated with the so-called *full model* that corresponds to $\gamma = (1, \ldots, 1)$.

Prior definitions

(i) For the full model, Zellner's G-prior:

$$\beta|\sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^{\mathrm{T}}X)^{-1}) \quad \text{and} \quad \sigma^2 \sim \pi(\sigma^2|X) = \sigma^{-2}$$

(ii) For each model $\mathfrak{M}_\gamma$, the prior distribution of β_γ conditional on σ^2 is fixed as

$$\beta_\gamma|\gamma, \sigma^2 \sim \mathcal{N}_{q_\gamma+1}\left(\tilde{\beta}_\gamma, c\sigma^2\left(X_\gamma^{\mathrm{T}}X_\gamma\right)^{-1}\right),$$

where $\tilde{\beta}_\gamma = \left(X_\gamma^T X_\gamma\right)^{-1} X_\gamma^T \tilde{\beta}$ and same prior on σ^2.

Prior completion

The joint prior for model $\mathfrak{M}_\gamma$ is the improper prior

$$\pi(\beta_\gamma, \sigma^2|\gamma) \propto \left(\sigma^2\right)^{-(q_\gamma+1)/2-1} \exp\left[-\frac{1}{2(c\sigma^2)}\left(\beta_\gamma - \tilde{\beta}_\gamma\right)^{\mathrm{T}} (X_\gamma^{\mathrm{T}} X_\gamma)\left(\beta_\gamma - \tilde{\beta}_\gamma\right)\right].$$

Prior competition (2)

Infinitely many ways of defining a prior on the model index γ: choice of uniform prior $\pi(\gamma|X) = 2^{-k}$.

Posterior distribution of γ central to variable selection since it is proportional to marginal density of y on $\mathfrak{M}_\gamma$ (or *evidence* of $\mathfrak{M}_\gamma$)

$$\begin{aligned}\pi(\gamma|y,X) &\propto f(y|\gamma,X)\pi(\gamma|X) \propto f(y|\gamma,X)\\ &= \int\left(\int f(y|\gamma,\beta,\sigma^2,X)\pi(\beta|\gamma,\sigma^2,X)\,\mathrm{d}\beta\right)\pi(\sigma^2|X)\,\mathrm{d}\sigma^2\,.\end{aligned}$$

$$
\begin{aligned}
f(y|\gamma,\sigma^2,X) &= \int f(y|\gamma,\beta,\sigma^2)\pi(\beta|\gamma,\sigma^2)\,\mathrm{d}\beta \\
&= (c+1)^{-(q_\gamma+1)/2}(2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}y^{\mathrm{T}}y\right. \\
&\qquad \left.+\frac{1}{2\sigma^2(c+1)}\left\{cy^{\mathrm{T}}X_\gamma\left(X_\gamma^{\mathrm{T}}X_\gamma\right)^{-1}X_\gamma^{\mathrm{T}}y-\tilde{\beta}_\gamma^{\mathrm{T}}X_\gamma^{\mathrm{T}}X_\gamma\tilde{\beta}_\gamma\right\}\right)
\end{aligned}
$$

this posterior density satisfies

$$
\begin{aligned}
\pi(\gamma|y,X) &\propto (c+1)^{-(q_\gamma+1)/2}\left[y^{\mathrm{T}}y-\frac{c}{c+1}y^{\mathrm{T}}X_\gamma\left(X_\gamma^{\mathrm{T}}X_\gamma\right)^{-1}X_\gamma^{\mathrm{T}}y\right. \\
&\qquad \left.-\frac{1}{c+1}\tilde{\beta}_\gamma^{\mathrm{T}}X_\gamma^{\mathrm{T}}X_\gamma\tilde{\beta}_\gamma\right]^{-n/2}.
\end{aligned}
$$

Pine processionary caterpillars

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y}, X)$
0,1,2,4,5	0.2316
0,1,2,4,5,9	0.0374
0,1,9	0.0344
0,1,2,4,5,10	0.0328
0,1,4,5	0.0306
0,1,2,9	0.0250
0,1,2,4,5,7	0.0241
0,1,2,4,5,8	0.0238
0,1,2,4,5,6	0.0237
0,1,2,3,4,5	0.0232
0,1,6,9	0.0146
0,1,2,3,9	0.0145
0,9	0.0143
0,1,2,6,9	0.0135
0,1,4,5,9	0.0128
0,1,3,9	0.0117
0,1,2,8	0.0115

Pine processionary caterpillars (cont'd)

Interpretation

Model $\mathfrak{M}_\gamma$ with the highest posterior probability is $t_1(\gamma) = (1, 2, 4, 5)$, which corresponds to the variables

- altitude,
- slope,
- height of the tree sampled in the center of the area, and
- diameter of the tree sampled in the center of the area.

Pine processionary caterpillars (cont'd)

Interpretation

Model $\mathfrak{M}_\gamma$ with the highest posterior probability is $t_1(\gamma) = (1, 2, 4, 5)$, which corresponds to the variables

- altitude,
- slope,
- height of the tree sampled in the center of the area, and
- diameter of the tree sampled in the center of the area.

Corresponds to the five variables identified in the R regression output

Noninformative extension

For Zellner noninformative prior with $\pi(c) = 1/c$, we have

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

Pine processionary caterpillars

$t_1(\gamma)$	$\pi(\gamma\|\mathbf{y}, X)$
0,1,2,4,5	0.0929
0,1,2,4,5,9	0.0325
0,1,2,4,5,10	0.0295
0,1,2,4,5,7	0.0231
0,1,2,4,5,8	0.0228
0,1,2,4,5,6	0.0228
0,1,2,3,4,5	0.0224
0,1,2,3,4,5,9	0.0167
0,1,2,4,5,6,9	0.0167
0,1,2,4,5,8,9	0.0137
0,1,4,5	0.0110
0,1,2,4,5,9,10	0.0100
0,1,2,3,9	0.0097
0,1,2,9	0.0093
0,1,2,4,5,7,9	0.0092
0,1,2,6,9	0.0092

Stochastic search for the most likely model

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

Stochastic search for the most likely model

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

Need of a tailored algorithm that samples from $\pi(\gamma|y, X)$ and selects the most likely models.

Can be done by Gibbs sampling, given the availability of the full conditional posterior probabilities of the γ_i's.
If $\gamma_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k)$ $(1 \leq i \leq k)$

$$\pi(\gamma_i|y, \gamma_{-i}, X) \propto \pi(\gamma|y, X)$$

(to be evaluated in both $\gamma_i = 0$ and $\gamma_i = 1$)

Gibbs sampling for variable selection

Initialization: Draw γ^0 from the uniform distribution on Γ

Gibbs sampling for variable selection

Initialization: Draw γ^0 from the uniform distribution on Γ

Iteration t**:** Given $(\gamma_1^{(t-1)}, \ldots, \gamma_k^{(t-1)})$, generate

1. $\gamma_1^{(t)}$ according to $\pi(\gamma_1 | y, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$
2. $\gamma_2^{(t)}$ according to $\pi(\gamma_2 | y, \gamma_1^{(t)}, \gamma_3^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$

$\vdots$

p. $\gamma_k^{(t)}$ according to $\pi(\gamma_k | y, \gamma_1^{(t)}, \ldots, \gamma_{k-1}^{(t)}, X)$

MCMC interpretation

After $T \gg 1$ MCMC iterations, output used to approximate the posterior probabilities $\pi(\gamma|y, X)$ by empirical averages

$$\widehat{\pi}(\gamma|y, X) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \mathbb{I}_{\gamma^{(t)}=\gamma}.$$

where the T_0 first values are eliminated as *burnin*.

MCMC interpretation

After $T \gg 1$ MCMC iterations, output used to approximate the posterior probabilities $\pi(\gamma|y, X)$ by empirical averages

$$\widehat{\pi}(\gamma|y, X) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \mathbb{I}_{\gamma^{(t)}=\gamma}.$$

where the T_0 first values are eliminated as *burnin*.

And approximation of the probability to include i-th variable,

$$\widehat{P}^{\pi}(\gamma_i = 1|y, X) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \mathbb{I}_{\gamma_i^{(t)}=1}.$$

Pine processionary caterpillars

γ_i	$\widehat{P}^{\pi}(\gamma_i = 1\|\mathbf{y}, X)$	$\widehat{P}^{\pi}(\gamma_i = 1\|\mathbf{y}, X)$
γ_1	0.8624	0.8844
γ_2	0.7060	0.7716
γ_3	0.1482	0.2978
γ_4	0.6671	0.7261
γ_5	0.6515	0.7006
γ_6	0.1678	0.3115
γ_7	0.1371	0.2880
γ_8	0.1555	0.2876
γ_9	0.4039	0.5168
γ_{10}	0.1151	0.2609

Probabilities of inclusion with both informative $(\tilde{\beta} = 0_{11}, c = 100)$ and noninformative Zellner's priors

Generalized linear models

Generalisation of Linear Models

Linear models model connection between a response variable y and a set x of explanatory variables by a linear dependence relation with [approximately] normal perturbations.

Generalisation of Linear Models

Linear models model connection between a response variable y and a set x of explanatory variables by a linear dependence relation with [approximately] normal perturbations.

Many instances where either of these assumptions not appropriate, e.g. when the support of y restricted to $\mathbb{R}_+$ or to $\mathbb{N}$.

bank

Four measurements on 100 genuine Swiss banknotes and 100 counterfeit ones:

x_1 length of the bill (in mm),

x_2 width of the left edge (in mm),

x_3 width of the right edge (in mm),

x_4 bottom margin width (in mm).

Response variable y: status of the banknote [0 for genuine and 1 for counterfeit]

bank

Four measurements on 100 genuine Swiss banknotes and 100 counterfeit ones:

- x_1 length of the bill (in mm),
- x_2 width of the left edge (in mm),
- x_3 width of the right edge (in mm),
- x_4 bottom margin width (in mm).

Response variable y: status of the banknote [0 for genuine and 1 for counterfeit]

Probabilistic model that predicts counterfeiting based on the four measurements

The impossible linear model

Example of the influence of x_4 on y
Since y is binary,

$$y|x_4 \sim \mathscr{B}(p(x_4))\,,$$

© **Normal model is impossible**

The impossible linear model

Example of the influence of x_4 on y
Since y is binary,

$$y|x_4 \sim \mathscr{B}(p(x_4))\,,$$

© Normal model is impossible

Linear dependence in $p(x) = \mathbb{E}[y|x]$'s

$$p(x_{4i}) = \beta_0 + \beta_1 x_{4i}\,,$$

The impossible linear model

Example of the influence of x_4 on y
Since y is binary,

$$y|x_4 \sim \mathscr{B}(p(x_4))\,,$$

© Normal model is impossible

Linear dependence in $p(x) = \mathbb{E}[y|x]$'s

$$p(x_{4i}) = \beta_0 + \beta_1 x_{4i}\,,$$

estimated [by MLE] as

$$\hat{p}_i = -2.02 + 0.268\, x_{i4}$$

which gives $\hat{p}_i = .12$ for $x_{i4} = 8$ and ...

The impossible linear model

Example of the influence of x_4 on y
Since y is binary,

$$y|x_4 \sim \mathscr{B}(p(x_4))\,,$$

© Normal model is impossible

Linear dependence in $p(x) = \mathbb{E}[y|x]$'s

$$p(x_{4i}) = \beta_0 + \beta_1 x_{4i}\,,$$

estimated [by MLE] as

$$\hat{p}_i = -2.02 + 0.268\, x_{i4}$$

which gives $\hat{p}_i = .12$ for $x_{i4} = 8$ and ... $\hat{p}_i = 1.19$ for $x_{i4} = 12$!!!
© Linear dependence is impossible

Generalisation of the linear dependence

Broader class of models to cover various dependence structures.

Generalisation of the linear dependence

Broader class of models to cover various dependence structures.

Class of *generalised linear models* (GLM) where

$$y|\mathbf{x}, \beta \sim f(y|\mathbf{x}^{\mathrm{T}}\beta)\,.$$

i.e., dependence of y on $\mathbf{x}$ partly *linear*

Notations

Same as in linear regression chapter, with n–sample

$$\mathbf{y} = (y_1, \ldots, y_n)$$

and corresponding explanatory variables/covariates

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

Specifications of GLM's

Definition (GLM)

A GLM is a conditional model specified by two functions:

1. the density f of y given $\mathbf{x}$ parameterised by its expectation parameter $\mu = \mu(\mathbf{x})$ [and possibly its dispersion parameter $\varphi = \varphi(\mathbf{x})$]

Specifications of GLM's

Definition (GLM)

A GLM is a conditional model specified by two functions:

1. the density f of y given $\mathbf{x}$ parameterised by its expectation parameter $\mu = \mu(\mathbf{x})$ [and possibly its dispersion parameter $\varphi = \varphi(\mathbf{x})$]
2. the *link* g between the mean μ and the explanatory variables, written customarily as $g(\mu) = \mathbf{x}^{\mathrm{T}}\beta$ or, equivalently, $\mathbb{E}[y|\mathbf{x}, \beta] = g^{-1}(\mathbf{x}^{\mathrm{T}}\beta)$.

Specifications of GLM's

Definition (GLM)

A GLM is a conditional model specified by two functions:

1. the density f of y given $\mathbf{x}$ parameterised by its expectation parameter $\mu = \mu(\mathbf{x})$ [and possibly its dispersion parameter $\varphi = \varphi(\mathbf{x})$]
2. the *link* g between the mean μ and the explanatory variables, written customarily as $g(\mu) = \mathbf{x}^{\mathrm{T}}\beta$ or, equivalently, $\mathbb{E}[y|\mathbf{x}, \beta] = g^{-1}(\mathbf{x}^{\mathrm{T}}\beta)$.

For identifiability reasons, g needs to be bijective.

Likelihood

Obvious representation of the likelihood

$$\ell(\beta, \varphi|\mathbf{y}, X) = \prod_{i=1}^{n} f\left(y_i|\mathbf{x}^{i\mathrm{T}}\beta, \varphi\right)$$

with parameters $\beta \in \mathbb{R}^k$ and $\varphi > 0$.

Examples

- Ordinary linear regression
  Case of GLM where

$$g(x) = x, \ \varphi = \sigma^2, \quad \text{and} \quad \mathbf{y}|X, \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2).$$

Examples (2)

Case of binary and binomial data, when

$$y_i|\mathbf{x}^i \sim \mathscr{B}(n_i, p(\mathbf{x}^i))$$

with known n_i

- Logit [or logistic regression] model
  Link is *logit transform* on probability of success

$$g(p_i) = \log(p_i/(1-p_i))\,,$$

with likelihood

$$\prod_{i=1}^{n} \left(\begin{array}{c} n_i \\ y_i \end{array} \right) \left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)} \right)^{y_i} \left(\frac{1}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)} \right)^{n_i-y_i}$$

$$\propto \exp\left\{ \sum_{i=1}^{n} y_i \mathbf{x}^{i\mathrm{T}}\beta \right\} \Big/ \prod_{i=1}^{n} \left(1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right)^{n_i-y_i}$$

Canonical link

Special link function g that appears in the natural exponential family representation of the density

$$g^{\star}(\mu) = \theta \quad \text{if} \quad f(y|\mu) \propto \exp\{T(y) \cdot \theta - \Psi(\theta)\}$$

Canonical link

Special link function g that appears in the natural exponential family representation of the density

$$g^\star(\mu) = \theta \quad \text{if} \quad f(y|\mu) \propto \exp\{T(y) \cdot \theta - \Psi(\theta)\}$$

Example

Logit link is canonical for the binomial model, since

$$f(y_i|p_i) = \binom{n_i}{y_i} \exp\left\{ y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i) \right\},$$

and thus

$$\theta_i = \log p_i/(1-p_i)$$

Examples (3)

Customary to use the canonical link, but only customary ...

- Probit model
  Probit link function given by

$$g(\mu_i) = \Phi^{-1}(\mu_i)$$

  where Φ standard normal cdf
  Likelihood

$$\ell(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathrm{T}}\beta)^{y_i}(1 - \Phi(\mathbf{x}^{i\mathrm{T}}\beta))^{n_i - y_i} .$$

▸ Full processing

Log-linear models

Standard approach to describe associations between several *categorical* variables, i.e, variables with finite support

Sufficient statistic: *contingency table*, made of the cross-classified counts for the different categorical variables. ▸ Full entry to loglinear models

Log-linear models

Standard approach to describe associations between several *categorical* variables, i.e, variables with finite support

Sufficient statistic: *contingency table*, made of the cross-classified counts for the different categorical variables. ▸ Full entry to loglinear models

Example (Titanic survivors)

Survivor	Class	Child		Adult	
		Male	Female	Male	Female
No	1st	0	0	118	4
	2nd	0	0	154	13
	3rd	35	17	387	89
	Crew	0	0	670	3
Yes	1st	5	1	57	140
	2nd	11	13	14	80
	3rd	13	14	75	76
	Crew	0	0	192	20

Poisson regression model

1. Each count y_i is Poisson with mean $\mu_i = \mu(\mathbf{x}_i)$
2. Link function connecting $\mathbb{R}^+$ with $\mathbb{R}$, e.g. logarithm $g(\mu_i) = \log(\mu_i)$.

Poisson regression model

1. Each count y_i is Poisson with mean $\mu_i = \mu(\mathbf{x}_i)$
2. Link function connecting $\mathbb{R}^+$ with $\mathbb{R}$, e.g. logarithm $g(\mu_i) = \log(\mu_i)$.

Corresponding likelihood

$$\ell(\beta|y, X) = \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \exp\left\{y_i \mathbf{x}^{i\mathrm{T}} \beta - \exp(\mathbf{x}^{i\mathrm{T}} \beta)\right\} .$$

Metropolis–Hastings algorithms

Convergence assessment

Posterior inference in GLMs harder than for linear models

Metropolis–Hastings algorithms

Convergence assessment

Posterior inference in GLMs harder than for linear models

© Working with a GLM requires specific numerical or simulation tools [E.g., GLIM in classical analyses]

Metropolis–Hastings algorithms

Convergence assessment

Posterior inference in GLMs harder than for linear models

© Working with a GLM requires specific numerical or simulation tools [E.g., GLIM in classical analyses]

Opportunity to introduce universal MCMC method: *Metropolis–Hastings* algorithm

Generic MCMC sampler

- Metropolis–Hastings algorithms are generic/down-the-shelf MCMC algorithms
- Only require likelihood up to a constant [difference with Gibbs sampler]
- can be tuned with a wide range of possibilities [difference with Gibbs sampler & blocking]
- natural extensions of standard simulation algorithms: based on the choice of a *proposal* distribution [difference in Markov proposal $q(x, y)$ and acceptance]

Why Metropolis?

Originally introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller in a setup of optimization on a discrete state-space. All authors involved in Los Alamos during and after WWII:

Why Metropolis?

Originally introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller in a setup of optimization on a discrete state-space. All authors involved in Los Alamos during and after WWII:

- Physicist and mathematician, Nicholas Metropolis is considered (with Stanislaw Ulam) to be the father of Monte Carlo methods.
- Also a physicist, Marshall Rosenbluth worked on the development of the hydrogen (H) bomb
- Edward Teller was one of the first scientists to work on the Manhattan Project that led to the production of the A bomb. Also managed to design with Ulam the H bomb.

Generic Metropolis–Hastings sampler

For *target* π and proposal kernel $q(x, y)$

Initialization: Choose an arbitrary $x^{(0)}$

Generic Metropolis–Hastings sampler

For *target* π and proposal kernel $q(x, y)$

Initialization: Choose an arbitrary $x^{(0)}$

Iteration t**:**

1. Given $x^{(t-1)}$, generate $\tilde{x} \sim q(x^{(t-1)}, x)$
2. Calculate

$$\rho(x^{(t-1)}, \tilde{x}) = \min\left(\frac{\pi(\tilde{x})/q(x^{(t-1)}, \tilde{x})}{\pi(x^{(t-1)})/q(\tilde{x}, x^{(t-1)})}, 1\right)$$

3. With probability $\rho(x^{(t-1)}, \tilde{x})$ accept $\tilde{x}$ and set $x^{(t)} = \tilde{x}$; otherwise reject $\tilde{x}$ and set $x^{(t)} = x^{(t-1)}$.

Universality

Algorithm only needs to simulate from

$$q$$

which can be chosen [almost!] arbitrarily, i.e. unrelated with π [q also called *instrumental* distribution]

Note: π and q known up to proportionality terms ok since proportionality constants cancel in ρ.

Validation

Markov chain theory

Target π is stationary distribution of Markov chain $(x^{(t)})_t$ because probability $\rho(x, y)$ satisfies *detailed balance equation*

$$\pi(x)q(x, y)\rho(x, y) = \pi(y)q(y, x)\rho(y, x)$$

[Integrate out x to see that π is stationary]

Validation

Markov chain theory

Target π is stationary distribution of Markov chain $(x^{(t)})_t$ because probability $\rho(x, y)$ satisfies *detailed balance equation*

$$\pi(x)q(x, y)\rho(x, y) = \pi(y)q(y, x)\rho(y, x)$$

[Integrate out x to see that π is stationary]

For convergence/ergodicity, Markov chain must be *irreducible*: q has positive probability of reaching all areas with positive π probability in a finite number of steps.

Choice of proposal

Theoretical guarantees of convergence very high, but choice of q is crucial in practice.

Choice of proposal

Theoretical guarantees of convergence very high, but choice of q is crucial in practice. Poor choice of q may result in

- very high rejection rates, with very few moves of the Markov chain $(x^{(t)})_t$ hardly moves, or in

Choice of proposal

Theoretical guarantees of convergence very high, but choice of q is crucial in practice. Poor choice of q may result in

- very high rejection rates, with very few moves of the Markov chain $(x^{(t)})_t$ hardly moves, or in
- a myopic exploration of the support of π, that is, in a dependence on the starting value $x^{(0)}$, with the chain stuck in a neighbourhood mode to $x^{(0)}$.

Choice of proposal

Theoretical guarantees of convergence very high, but choice of q is crucial in practice. Poor choice of q may result in

- very high rejection rates, with very few moves of the Markov chain $(x^{(t)})_t$ hardly moves, or in
- a myopic exploration of the support of π, that is, in a dependence on the starting value $x^{(0)}$, with the chain stuck in a neighbourhood mode to $x^{(0)}$.

Note: hybrid MCMC

Simultaneous use of different kernels valid *and* recommended

The independence sampler

Pick proposal q that is independent of its first argument,

$$q(x, y) = q(y)$$

ρ simplifies into

$$\rho(x, y) = \min\left(1, \frac{\pi(y)/q(y)}{\pi(x)/q(x)}\right) .$$

Special case: $q \propto \pi$

Reduces to $\rho(x, y) = 1$ and iid sampling

The independence sampler

Pick proposal q that is independent of its first argument,

$$q(x, y) = q(y)$$

ρ simplifies into

$$\rho(x, y) = \min\left(1, \frac{\pi(y)/q(y)}{\pi(x)/q(x)}\right) .$$

Special case: $q \propto \pi$

Reduces to $\rho(x, y) = 1$ and iid sampling

Analogy with Accept-Reject algorithm where $\max \pi/q$ replaced with the current value $\pi(x^{(t-1)})/q(x^{(t-1)})$ but sequence of accepted $x^{(t)}$'s not i.i.d.

Choice of q

Convergence properties highly dependent on q.

- q needs to be positive everywhere on the support of π
- for a good exploration of this support, π/q needs to be bounded.

Choice of q

Convergence properties highly dependent on q.

- q needs to be positive everywhere on the support of π
- for a good exploration of this support, π/q needs to be bounded.

Otherwise, the chain takes too long to reach regions with low q/π values.

The random walk sampler

Independence sampler requires too much global information about π: opt for a local gathering of information

The random walk sampler

Independence sampler requires too much global information about π: opt for a local gathering of information

Means exploration of the neighbourhood of the current value $x^{(t)}$ in search of other points of interest.

The random walk sampler

Independence sampler requires too much global information about π: opt for a local gathering of information

Means exploration of the neighbourhood of the current value $x^{(t)}$ in search of other points of interest.

Simplest exploration device is based on random walk dynamics.

Random walks

Proposal is a symmetric transition density

$$q(x, y) = q_{RW}(y - x) = q_{RW}(x - y)$$

Random walks

Proposal is a symmetric transition density

$$q(x, y) = q_{RW}(y - x) = q_{RW}(x - y)$$

Acceptance probability $\rho(x, y)$ reduces to the simpler form

$$\rho(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right) .$$

Only depends on the target π *[accepts all proposed values that increase π]*

Choice of q_{RW}

Considerable flexibility in the choice of q_{RW},

- tails: Normal versus Student's t
- scale: size of the neighbourhood

Choice of q_{RW}

Considerable flexibility in the choice of q_{RW},

- tails: Normal versus Student's t
- scale: size of the neighbourhood

Can also be used for restricted support targets [with a waste of simulations near the boundary]

Choice of q_{RW}

Considerable flexibility in the choice of q_{RW},

- tails: Normal versus Student's t
- scale: size of the neighbourhood

Can also be used for restricted support targets [with a waste of simulations near the boundary]

Can be tuned towards an acceptance probability of 0.234 at the *burnin* stage *[Magic number!]*

Convergence assessment

Capital question: How many iterations do we need to run???

◂ MCMC debuts

Convergence assessment

Capital question: How many iterations do we need to run???

◂ MCMC debuts

- **Rule # 1** There is no absolute number of simulations, i.e. $1,000$ is neither large, nor small.
- **Rule # 2** It takes [much] longer to check for convergence than for the chain itself to converge.
- **Rule # 3** MCMC is a *"what-you-get-is-what-you-see"* algorithm: it fails to tell about unexplored parts of the space.
- **Rule # 4** When in doubt, run MCMC chains in parallel and check for consistency.

Convergence assessment

Capital question: How many iterations do we need to run???

◂ MCMC debuts

- **Rule # 1** There is no absolute number of simulations, i.e. $1,000$ is neither large, nor small.
- **Rule # 2** It takes [much] longer to check for convergence than for the chain itself to converge.
- **Rule # 3** MCMC is a *"what-you-get-is-what-you-see"* algorithm: it fails to tell about unexplored parts of the space.
- **Rule # 4** When in doubt, run MCMC chains in parallel and check for consistency.

Many "quick-&-dirty" solutions in the literature, but not necessarily trustworthy.

Prohibited dynamic updating

- Tuning the proposal in terms of its past performances can only be implemented at *burnin*, because otherwise this cancels Markovian convergence properties.

Prohibited dynamic updating

- Tuning the proposal in terms of its past performances can only be implemented at *burnin*, because otherwise this cancels Markovian convergence properties.

Use of several MCMC proposals together within a single algorithm using circular or random design is ok. It almost always brings an improvement compared with its individual components (at the cost of increased simulation time)

Effective sample size

How many iid simulations from π are equivalent to N simulations from the MCMC algorithm?

Effective sample size

How many iid simulations from π are equivalent to N simulations from the MCMC algorithm?

Based on estimated k-th order auto-correlation,

$$\rho_k = \text{corr}\left(x^{(t)}, x^{(t+k)}\right) ,$$

effective sample size

$$N^{\text{ess}} = n \left(1 + 2 \sum_{k=1}^{T_0} \hat{\rho}_k^2\right)^{-1/2} ,$$

- Only partial indicator that fails to signal chains stuck in one mode of the target

The Probit Model

Likelihood

◂ Recall Probit

$$\ell(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathrm{T}}\beta)^{y_i}(1 - \Phi(\mathbf{x}^{i\mathrm{T}}\beta))^{n_i - y_i} .$$

The Probit Model

Likelihood ◂ Recall Probit

$$\ell(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathrm{T}}\beta)^{y_i}(1 - \Phi(\mathbf{x}^{i\mathrm{T}}\beta))^{n_i - y_i} .$$

If no prior information available, resort to the flat prior $\pi(\beta) \propto 1$ and then obtain the posterior distribution

$$\pi(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi\left(\mathbf{x}^{i\mathrm{T}}\beta\right)^{y_i}\left(1 - \Phi(\mathbf{x}^{i\mathrm{T}}\beta)\right)^{n_i - y_i} ,$$

nonstandard and simulated using MCMC techniques.

MCMC resolution

Metropolis–Hastings random walk sampler works well for binary regression problems with small number of predictors

Uses the maximum likelihood estimate $\hat{\beta}$ as starting value and asymptotic (Fisher) covariance matrix of the MLE, $\hat{\Sigma}$, as scale

MLE proposal

R function glm very useful to get the maximum likelihood estimate of β and its asymptotic covariance matrix $\hat{\Sigma}$.

Terminology used in R program

```
mod=summary(glm(y~X-1,family=binomial(link="probit")))
```

with mod$coeff[,1] denoting $\hat{\beta}$ and mod$cov.unscaled $\hat{\Sigma}$.

MCMC algorithm

Probit random-walk Metropolis-Hastings

Initialization: Set $\beta^{(0)} = \hat{\beta}$ and compute $\hat{\Sigma}$

Iteration t**:**

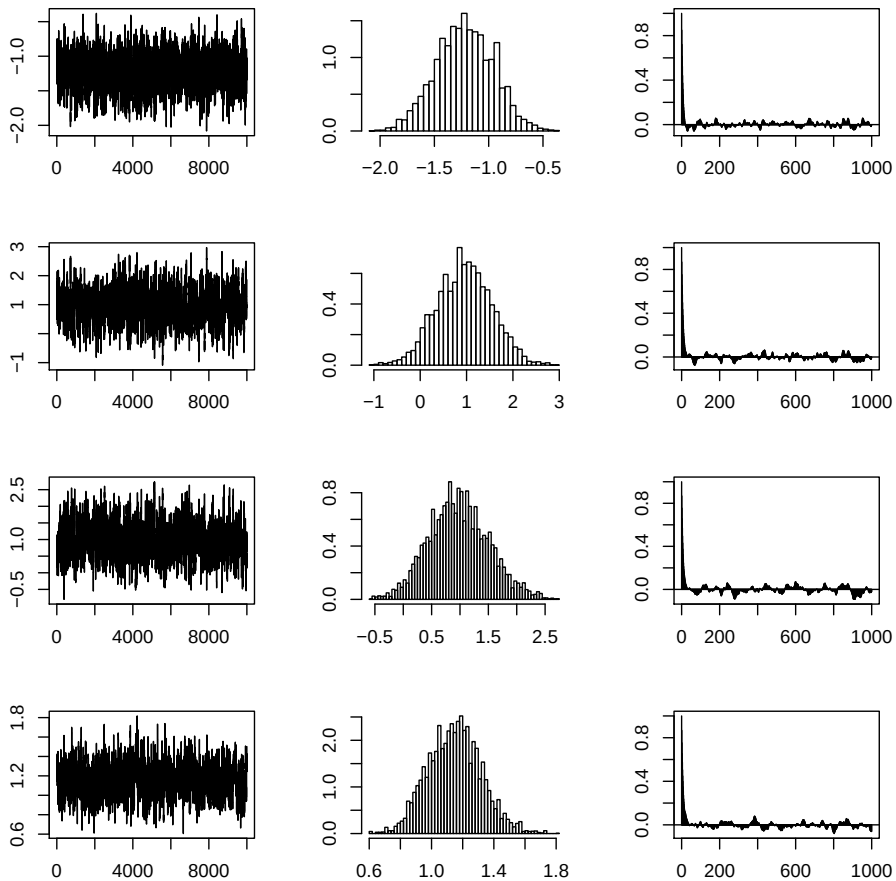
1. Generate $\tilde{\beta} \sim \mathcal{N}_{k+1}(\beta^{(t-1)}, \tau\hat{\Sigma})$
2. Compute

$$\rho(\beta^{(t-1)}, \tilde{\beta}) = \min\left(1, \frac{\pi(\tilde{\beta}|y)}{\pi(\beta^{(t-1)}|y)}\right)$$

3. With probability $\rho(\beta^{(t-1)}, \tilde{\beta})$ set $\beta^{(t)} = \tilde{\beta}$; otherwise set $\beta^{(t)} = \beta^{(t-1)}$.

bank

Probit modelling with no intercept over the four measurements. Three different scales $\tau = 1, 0.1, 10$: best mixing behavior is associated with $\tau = 1$. Average of the parameters over $9,000$ iterations gives plug-in estimate



$$\hat{p}_i = \Phi\left(-1.2193x_{i1} + 0.9540x_{i2} + 0.9795x_{i3} + 1.1481x_{i4}\right).$$

G-priors for probit models

Flat prior on β inappropriate for comparison purposes and Bayes factors.

G-priors for probit models

Flat prior on β inappropriate for comparison purposes and Bayes factors.
Replace the flat prior with a hierarchical prior,

$$\beta|\sigma^2, X \sim \mathcal{N}_k\left(0_k, \sigma^2(X^{\mathrm{T}}X)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X) \propto \sigma^{-3/2}\,,$$

(almost) as in normal linear regression

G-priors for probit models

Flat prior on β inappropriate for comparison purposes and Bayes factors.
Replace the flat prior with a hierarchical prior,

$$\beta|\sigma^2, X \sim \mathcal{N}_k\left(0_k, \sigma^2(X^{\mathrm{T}}X)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X) \propto \sigma^{-3/2}\,,$$

(almost) as in normal linear regression

Warning

The matrix $X^{\mathrm{T}}X$ is *not* the Fisher information matrix

G-priors for testing

Same argument as before: while π is improper, use of the *same* variance factor σ^2 in both models means the normalising constant cancels in the Bayes factor.

G-priors for testing

Same argument as before: while π is improper, use of the *same* variance factor σ^2 in both models means the normalising constant cancels in the Bayes factor.

Posterior distribution of β

$$\pi(\beta|\mathbf{y},X) \propto |X^{\mathrm{T}}X|^{1/2}\Gamma((2k-1)/4)\left(\beta^{\mathrm{T}}(X^{\mathrm{T}}X)\beta\right)^{-(2k-1)/4}\pi^{-k/2} \times\prod_{i=1}^{n}\Phi(\mathbf{x}^{i\mathrm{T}}\beta)^{y_i}\left[1-\Phi(\mathbf{x}^{i\mathrm{T}}\beta)\right]^{1-y_i}$$

[where k matters!]

Marginal approximation

Marginal

$$f(\mathbf{y}|X) \propto |X^{\mathrm{T}}X|^{1/2}\,\pi^{-k/2}\Gamma\{(2k-1)/4\}\int\left(\beta^{\mathrm{T}}(X^{\mathrm{T}}X)\beta\right)^{-(2k-1)/4} \times\prod_{i=1}^{n}\Phi(\mathbf{x}^{i\mathrm{T}}\beta)^{y_i}\left[1-(\Phi(\mathbf{x}^{i\mathrm{T}}\beta)\right]^{1-y_i}\,\mathrm{d}\,\beta\,,$$

approximated by

$$\frac{|X^{\mathrm{T}}X|^{1/2}}{\pi^{k/2}M}\sum_{m=1}^{M}\left|\left|X\beta^{(m)}\right|\right|^{-(2k-1)/2}\prod_{i=1}^{n}\Phi(\mathbf{x}^{i\mathrm{T}}\beta^{(m)})^{y_i}\left[1-\Phi(\mathbf{x}^{i\mathrm{T}}\beta^{(m)})\right]^{1-y_i} \times\Gamma\{(2k-1)/4\}\,|\widehat{V}|^{1/2}(4\pi)^{k/2}\,e^{+(\beta^{(m)}-\widehat{\beta})^{\mathrm{T}}\widehat{V}^{-1}(\beta^{(m)}-\widehat{\beta})/4}\,,$$

where

$$\beta^{(m)}\sim\mathcal{N}_k(\widehat{\beta},2\,\widehat{V})$$

with $\widehat{\beta}$ MCMC approximation of $\mathbb{E}^{\pi}[\beta|\mathbf{y},X]$ and $\widehat{V}$ MCMC approximation of $\mathbb{V}(\beta|\mathbf{y},X)$.

Linear hypothesis

Linear restriction on β

$$H_0 : R\beta = r$$

($r \in \mathbb{R}^q$, R $q \times k$ matrix) where β^0 is $(k-q)$ dimensional and X_0 and $\mathbf{x}_0$ are linear transforms of X and of $\mathbf{x}$ of dimensions $(n, k-q)$ and $(k-q)$.

Likelihood

$$\ell(\beta^0|\mathbf{y}, X_0) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)^{y_i} \left[1 - \Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)\right]^{1-y_i} ,$$

Linear test

Associated [projected] G-prior

$$\beta^0|\sigma^2, X_0 \sim \mathcal{N}_{k-q}\left(0_{k-q}, \sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X_0) \propto \sigma^{-3/2}\,,$$

Linear test

Associated [projected] G-prior

$$\beta^0|\sigma^2, X_0 \sim \mathcal{N}_{k-q}\left(0_{k-q}, \sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X_0) \propto \sigma^{-3/2}\,,$$

Marginal distribution of $\mathbf{y}$ of the same type

$$f(\mathbf{y}|X_0) \propto |X_0^{\mathrm{T}}X_0|^{1/2}\pi^{-(k-q)/2}\Gamma\left\{\frac{(2(k-q)-1)}{4}\right\}\int ||X\beta^0||^{-(2(k-q)-1)/2} \prod_{i=1}^{n}\Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)^{y_i}\left[1-(\Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)\right]^{1-y_i}\mathrm{d}\beta^0\,.$$

banknote

For $H_0 : \beta_1 = \beta_2 = 0$, $B_{10}^{\pi} = 157.73$ [against H_0]

Generic regression-like output:

```
            Estimate  Post. var.  log10(BF)

X1           -1.1552    0.0631       4.5844 (****)
X2            0.9200    0.3299      -0.2875
X3            0.9121    0.2595      -0.0972
X4            1.0820    0.0287      15.6765 (****)

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

Informative settings

If prior information available on $p(\mathbf{x})$, transform into prior distribution on β by technique of *imaginary observations:*

Informative settings

If prior information available on $p(\mathbf{x})$, transform into prior distribution on β by technique of *imaginary observations:*

Start with k different values of the covariate vector, $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^k$
For each of these values, the practitioner specifies

(i) a prior guess g_i at the probability p_i associated with $\mathbf{x}^i$;

(ii) an assessment of (un)certainty about that guess given by a number K_i of equivalent "prior observations".

Informative settings

If prior information available on $p(\mathbf{x})$, transform into prior distribution on β by technique of *imaginary observations:*

Start with k different values of the covariate vector, $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^k$
For each of these values, the practitioner specifies

(i) a prior guess g_i at the probability p_i associated with $\mathbf{x}^i$;

(ii) an assessment of (un)certainty about that guess given by a number K_i of equivalent "prior observations".

On how many imaginary observations did you build this guess?

Informative prior

$$\pi(p_1, \ldots, p_k) \propto \prod_{i=1}^{k} p_i^{K_i g_i - 1} (1 - p_i)^{K_i (1 - g_i) - 1}$$

translates into *[Jacobian rule]*

$$\pi(\beta) \propto \prod_{i=1}^{k} \Phi(\tilde{\mathbf{x}}^{i\mathrm{T}} \beta)^{K_i g_i - 1} \left[1 - \Phi(\tilde{\mathbf{x}}^{i\mathrm{T}} \beta)\right]^{K_i (1 - g_i) - 1} \phi(\tilde{\mathbf{x}}^{i\mathrm{T}} \beta)$$

Informative prior

$$\pi(p_1, \ldots, p_k) \propto \prod_{i=1}^{k} p_i^{K_i g_i - 1} (1 - p_i)^{K_i(1-g_i)-1}$$

translates into *[Jacobian rule]*

$$\pi(\beta) \propto \prod_{i=1}^{k} \Phi(\tilde{\mathbf{x}}^{i\mathrm{T}}\beta)^{K_i g_i - 1} \left[1 - \Phi(\tilde{\mathbf{x}}^{i\mathrm{T}}\beta)\right]^{K_i(1-g_i)-1} \phi(\tilde{\mathbf{x}}^{i\mathrm{T}}\beta)$$

[Almost] equivalent to using the G-prior

$$\beta \sim \mathcal{N}_k \left(0_k, \left[\sum_{j=1}^{k} \tilde{\mathbf{x}}^j \tilde{\mathbf{x}}^{j\mathrm{T}}\right]^{-1}\right)$$

The logit model

Recall that [for $n_i = 1$]

$$y_i|\mu_i \sim \mathscr{B}(1, \mu_i), \quad \varphi = 1 \quad \text{and} \quad g(\mu_i) = \left(\frac{\exp(\mu_i)}{1+\exp(\mu_i)}\right).$$

Thus

$$\mathbb{P}(y_i = 1|\beta) = \frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}$$

with likelihood

$$\ell(\beta|\mathbf{y}, X) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{1-y_i}$$

Links with probit

- usual vague prior for β, $\pi(\beta) \propto 1$

Links with probit

- usual vague prior for β, $\pi(\beta) \propto 1$
- Posterior given by

$$\pi(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)} \right)^{1-y_i}$$

[intractable]

Links with probit

- usual vague prior for β, $\pi(\beta) \propto 1$
- Posterior given by

$$\pi(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)} \right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)} \right)^{1-y_i}$$

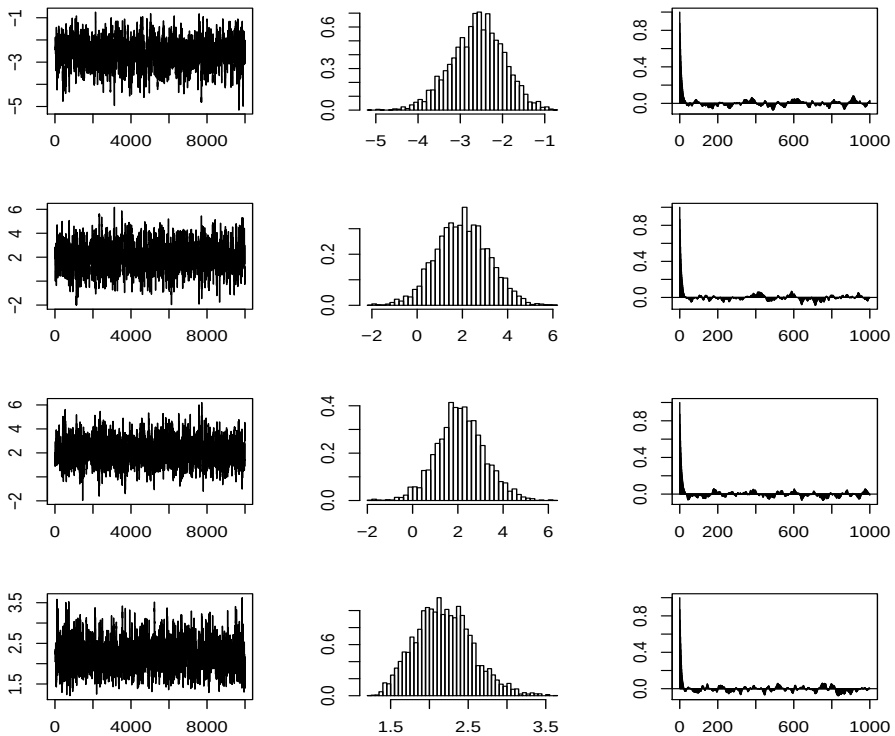
[intractable]

- Same Metropolis–Hastings sampler

bank

Same scale factor equal to $\tau = 1$: slight increase in the skewness of the histograms of the β_i's.

Plug-in estimate of predictive probability of a counterfeit



$$\hat{p}_i = \frac{\exp\left(-2.5888x_{i1} + 1.9967x_{i2} + 2.1260x_{i3} + 2.1879x_{i4}\right)}{1 + \exp\left(-2.5888x_{i1} + 1.9967x_{i2} + 2.1260x_{i3} + 2.1879x_{i4}\right)}.$$

G-priors for logit models

Same story: Flat prior on β inappropriate for Bayes factors, to be replaced with hierarchical prior,

$$\beta|\sigma^2, X \sim \mathcal{N}_k\left(0_k, \sigma^2(X^{\mathrm{T}}X)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X) \propto \sigma^{-3/2}$$

Example (bank)

```
          Estimate  Post. var.  log10(BF)

X1         -2.3970    0.3286       4.8084 (****)
X2          1.6978    1.2220      -0.2453
X3          2.1197    1.0094      -0.1529
X4          2.0230    0.1132      15.9530 (****)

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

Loglinear models

Introduction to loglinear models

Example (airquality)

Benchmark in R

```
> air=data(airquality)
```

Repeated measurements over 111 consecutive days of ozone u (in parts per billion) and maximum daily temperature v discretized into dichotomous variables

```
                month  5  6  7  8  9
ozone     temp
[1,31]    [57,79] 17  4  2  5 18
          (79,97]  0  2  3  3  2
(31,168] [57,79]  6  1  0  3  1
          (79,97]  1  2 21 12  8
```

Contingency table with $5 \times 2 \times 2 = 20$ entries

Poisson regression

Observations/counts $\mathbf{y} = (y_1, \ldots, y_n)$ are integers, so we can choose

$$y_i \sim \mathscr{P}(\mu_i)$$

Saturated likelihood

$$\ell(\mu|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\mu_i!} \mu_i^{y_i} \exp(-\mu_i)$$

Poisson regression

Observations/counts $\mathbf{y} = (y_1, \ldots, y_n)$ are integers, so we can choose

$$y_i \sim \mathscr{P}(\mu_i)$$

Saturated likelihood

$$\ell(\mu|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\mu_i!} \mu_i^{y_i} \exp(-\mu_i)$$

GLM constraint via log-linear link

$$\log(\mu_i) = \mathbf{x}^{i\mathrm{T}}\beta\,, \quad y_i|\mathbf{x}^i \sim \mathscr{P}\left(e^{\mathbf{x}^{i\mathrm{T}}\beta}\right)$$

Categorical variables

Special feature

Incidence matrix $X = (\mathbf{x}^i)$ such that its elements are all zeros or ones, i.e. covariates are all indicators/dummy variables!

Categorical variables

Special feature

Incidence matrix $X = (\mathbf{x}^i)$ such that its elements are all zeros or ones, i.e. covariates are all indicators/dummy variables!

Several types of (sub)models are possible depending on relations between categorical variables.

Categorical variables

Special feature

Incidence matrix $X = (\mathbf{x}^i)$ such that its elements are all zeros or ones, i.e. covariates are all indicators/dummy variables!

Several types of (sub)models are possible depending on relations between categorical variables.

Re-special feature

Variable selection problem of a specific kind, in the sense that all indicators related with the *same* association must either remain or vanish at once. Thus much fewer submodels than in a regular variable selection problem.

Parameterisations

Example of three variables $1 \leq u \leq I$, $1 \leq v \leq j$ and $1 \leq w \leq K$.

Parameterisations

Example of three variables $1 \leq u \leq I$, $1 \leq v \leq j$ and $1 \leq w \leq K$.

Simplest non-constant model is

$$\log(\mu_\tau) = \sum_{b=1}^{I} \beta_b^u \mathbb{I}_b(u_\tau) + \sum_{b=1}^{J} \beta_b^v \mathbb{I}_b(v_\tau) + \sum_{b=1}^{K} \beta_b^w \mathbb{I}_b(w_\tau)\,,$$

that is,

$$\log(\mu_{l(i,j,k)}) = \beta_i^u + \beta_j^v + \beta_k^w\,,$$

where index $l(i,j,k)$ corresponds to $u = i$, $v = j$ and $w = k$.

Parameterisations

Example of three variables $1 \leq u \leq I$, $1 \leq v \leq j$ and $1 \leq w \leq K$.

Simplest non-constant model is

$$\log(\mu_\tau) = \sum_{b=1}^{I} \beta_b^u \mathbb{I}_b(u_\tau) + \sum_{b=1}^{J} \beta_b^v \mathbb{I}_b(v_\tau) + \sum_{b=1}^{K} \beta_b^w \mathbb{I}_b(w_\tau)\,,$$

that is,

$$\log(\mu_{l(i,j,k)}) = \beta_i^u + \beta_j^v + \beta_k^w\,,$$

where index $l(i,j,k)$ corresponds to $u = i$, $v = j$ and $w = k$.
Saturated model is

$$\log(\mu_{l(i,j,k)}) = \beta_{ijk}^{uvw}$$

Log-linear model (over-)parameterisation

Representation

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw} ,$$

as in Anova models.

Log-linear model (over-)parameterisation

Representation

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw} ,$$

as in Anova models.

- λ appears as the overall or reference average effect
- λ_i^u appears as the marginal discrepancy (against the reference effect λ) when $u = i$,
- λ_{ij}^{uv} as the interaction discrepancy (against the added effects $\lambda + \lambda_i^u + \lambda_j^v$) when $(u, v) = (i, j)$

and so on...

Example of submodels

1. if both v and w are irrelevant, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u \,,$$
2. if all three categorical variables are mutually independent, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w \,,$$
3. if u and v are associated but are both independent of w, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} \,,$$
4. if u and v are conditionally independent given w, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} \,,$$
5. if there is no three-factor interaction, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw}$$
[the most complete submodel]

Identifiability

Representation

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw} ,$$

not identifiable but Bayesian approach handles non-identifiable settings and still estimate properly identifiable quantities.

Identifiability

Representation

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw} \,,$$

not identifiable but Bayesian approach handles non-identifiable settings and still estimate properly identifiable quantities. Customary to impose identifiability constraints on the parameters: set to 0 parameters corresponding to the first category of each variable, i.e. remove the indicator of the first category.

E.g., if $u \in \{1, 2\}$ and $v \in \{1, 2\}$, constraint could be

$$\lambda_1^u = \lambda_1^v = \lambda_{11}^{uv} = \lambda_{12}^{uv} = \lambda_{21}^{uv} = 0 \,.$$

Inference under a flat prior

Noninformative prior $\pi(\beta) \propto 1$ gives posterior distribution

$$\begin{aligned}\pi(\beta|\mathbf{y}, X) &\propto \prod_{i=1}^{n} \left\{\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}^{y_i} \exp\{-\exp(\mathbf{x}^{i\mathrm{T}}\beta)\} \\ &= \exp\left\{\sum_{i=1}^{n} y_i\, \mathbf{x}^{i\mathrm{T}}\beta - \sum_{i=1}^{n} \exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}\end{aligned}$$

Inference under a flat prior

Noninformative prior $\pi(\beta) \propto 1$ gives posterior distribution

$$\begin{aligned}\pi(\beta|\mathbf{y}, X) &\propto \prod_{i=1}^{n} \{\exp(\mathbf{x}^{i\mathrm{T}}\beta)\}^{y_i} \exp\{-\exp(\mathbf{x}^{i\mathrm{T}}\beta)\} \\ &= \exp\left\{\sum_{i=1}^{n} y_i\, \mathbf{x}^{i\mathrm{T}}\beta - \sum_{i=1}^{n} \exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}\end{aligned}$$

Use of same random walk M-H algorithm as in probit and logit cases, starting with MLE evaluation

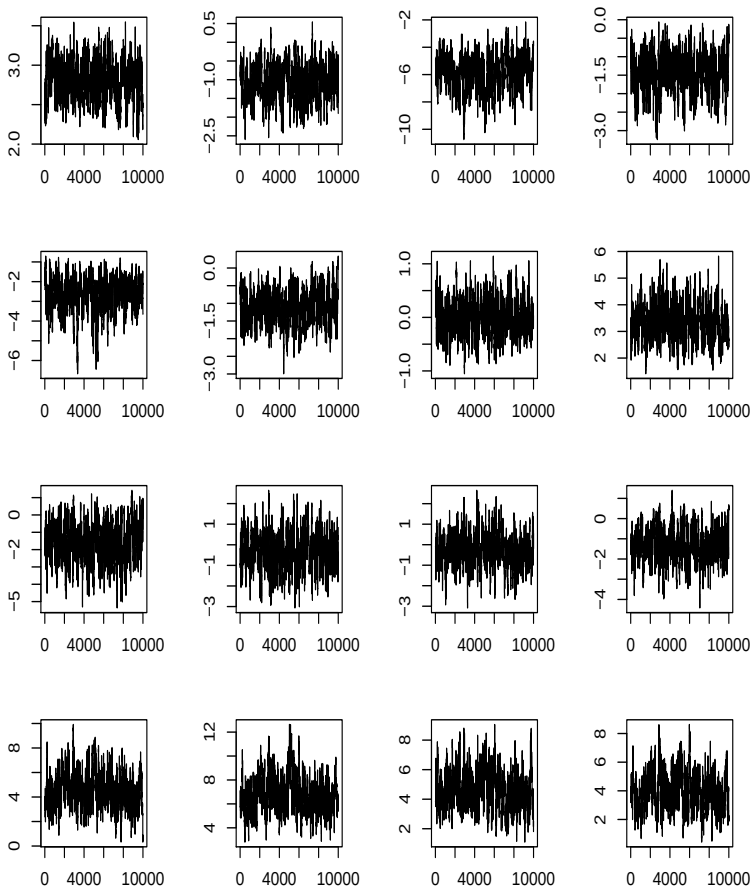
```
> mod=summary(glm(y~-1+X,family=poisson()))
```

airquality

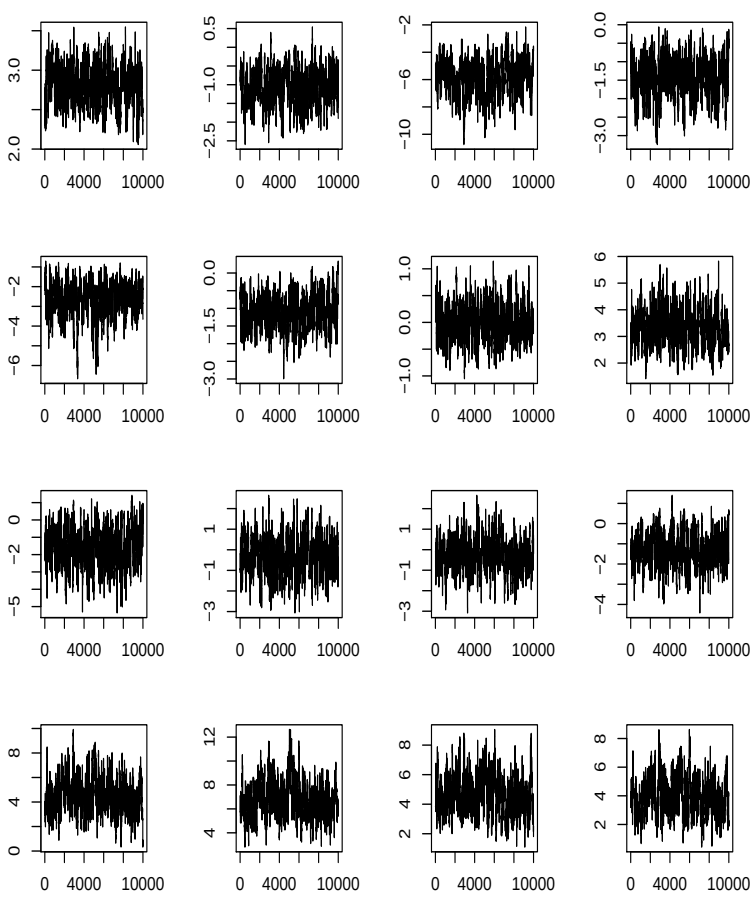
Identifiable non-saturated model involves 16 parameters
Obtained with $10,000$ MCMC iterations with scale factor $\tau^2 = 0.5$

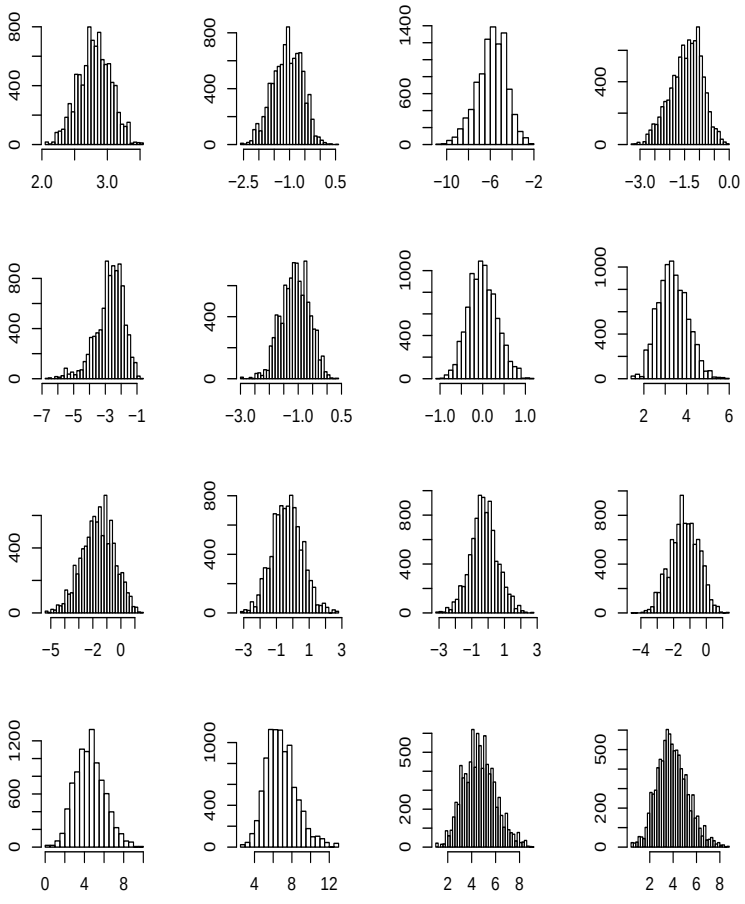
Effect	Post. mean	Post. var.
λ	2.8041	0.0612
λ_2^u	-1.0684	0.2176
λ_2^v	-5.8652	1.7141
λ_2^w	-1.4401	0.2735
λ_3^w	-2.7178	0.7915
λ_4^w	-1.1031	0.2295
λ_5^w	-0.0036	0.1127
λ_{22}^{uv}	3.3559	0.4490
λ_{22}^{uw}	-1.6242	1.2869
λ_{23}^{uw}	- 0.3456	0.8432
λ_{24}^{uw}	-0.2473	0.6658
λ_{25}^{uw}	-1.3335	0.7115
λ_{22}^{vw}	4.5493	2.1997
λ_{23}^{vw}	6.8479	2.5881
λ_{24}^{vw}	4.6557	1.7201
λ_{25}^{vw}	3.9558	1.7128

airquality: MCMC output



airquality: MCMC output





Model choice with G-prior

G-prior alternative used for probit and logit models still available:

$$\pi(\beta|\mathbf{y},X) \propto |X^{\mathrm{T}}X|^{1/2}\Gamma\left\{\frac{(2k-1)}{4}\right\}||X\beta||^{-(2k-1)/2}\,\pi^{-k/2} \times \exp\left\{\left(\sum_{i=1}^{n} y_i\,\mathbf{x}^i\right)^{\mathrm{T}}\beta - \sum_{i=1}^{n}\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}$$

Model choice with G-prior

G-prior alternative used for probit and logit models still available:

$$\pi(\beta|\mathbf{y},X) \propto |X^{\mathrm{T}}X|^{1/2}\Gamma\left\{\frac{(2k-1)}{4}\right\}||X\beta||^{-(2k-1)/2}\,\pi^{-k/2} \times \exp\left\{\left(\sum_{i=1}^{n} y_i\,\mathbf{x}^i\right)^{\mathrm{T}}\beta - \sum_{i=1}^{n}\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}$$

Same MCMC implementation and similar estimates for airquality

airquality

Bayes factors once more approximated by importance sampling based on normal importance functions

airquality

Bayes factors once more approximated by importance sampling based on normal importance functions

Anova-like output

```
Effect log10(BF)

u:v      6.0983 (****)
u:w     -0.5732
v:w      6.0802 (****)

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