Calibration and Validation of Computer Models: a Bayesian Approach
Lecture 1: An outline of the Bayesian approach to Statistics

Rui Paulo

ISEG/CEMAPRE Technical University of Lisboa, Portugal

June 29 2011

The Bayesian approach to statistical inference is based on a particular interpretation of the content of the well-known Theorem of Bayes:

The Bayesian approach to statistical inference is based on a particular interpretation of the content of the well-known Theorem of Bayes:

Theorem

Let $\{A_i, i = 1, \ldots, n\}$ *form a partition of the sample space* Ω *such that* $P(A_i) > 0$ *for all* $i = 1, \ldots, n$. *Let* B *be an event such that* $P(B) > 0$. *Then, for all* $i = 1, \ldots, n$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)\ P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)\ P(A_j)}$$

The Bayesian approach to statistical inference is based on a particular interpretation of the content of the well-known Theorem of Bayes:

Theorem

Let $\{A_i, i = 1, \ldots, n\}$ form a partition of the sample space Ω such that $P(A_i) > 0$ for all $i = 1, \ldots, n$. Let B be an event such that $P(B) > 0$. Then, for all $i = 1, \ldots, n$,

$$P(A_i \mid B) = \frac{P(B \mid A_i)\ P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)\ P(A_j)}$$

The use of this theorem in a deductive context, that of Probability Theory, is not controversial; $P(B \mid A_i)$ and $P(A_i)$ are assumed known and we want merely to compute $P(A_i \mid B)$

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

- A_i denotes an hypothesis or a model that we use to explain a certain phenomenon; a theory to which the researcher attributes *a priori* a degree of credibility given by $P(A_i)$ — prior information

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

- A_i denotes an hypothesis or a model that we use to explain a certain phenomenon; a theory to which the researcher attributes *a priori* a degree of credibility given by $P(A_i)$ — prior information
- B represents the result of observing that phenomenon

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

- A_i denotes an hypothesis or a model that we use to explain a certain phenomenon; a theory to which the researcher attributes *a priori* a degree of credibility given by $P(A_i)$ — prior information
- B represents the result of observing that phenomenon
- $P(B \mid A_i)$ denotes the likelihood of observing B when explanation A_i is assumed correct — sampling information

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

- A_i denotes an hypothesis or a model that we use to explain a certain phenomenon; a theory to which the researcher attributes *a priori* a degree of credibility given by $P(A_i)$ — prior information
- B represents the result of observing that phenomenon
- $P(B \mid A_i)$ denotes the likelihood of observing B when explanation A_i is assumed correct — sampling information
- The prior probabilities $P(A_i)$ are then updated into posterior probabilities after B has been observed: $P(A_i \mid B)$

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

- A_i denotes an hypothesis or a model that we use to explain a certain phenomenon; a theory to which the researcher attributes *a priori* a degree of credibility given by $P(A_i)$ — prior information
- B represents the result of observing that phenomenon
- $P(B \mid A_i)$ denotes the likelihood of observing B when explanation A_i is assumed correct — sampling information
- The prior probabilities $P(A_i)$ are then updated into posterior probabilities after B has been observed: $P(A_i \mid B)$
- This use of Bayes' theorem raises questions regarding the interpretation of the concept of probability involved in $P(A_i)$ and therefore in $P(A_i \mid B)$

The possible controversy arises in an inductive context, that of Statistics, as an instrument of learning from an experiment:

- A_i denotes an hypothesis or a model that we use to explain a certain phenomenon; a theory to which the researcher attributes *a priori* a degree of credibility given by $P(A_i)$ — prior information
- B represents the result of observing that phenomenon
- $P(B \mid A_i)$ denotes the likelihood of observing B when explanation A_i is assumed correct — sampling information
- The prior probabilities $P(A_i)$ are then updated into posterior probabilities after B has been observed: $P(A_i \mid B)$
- This use of Bayes' theorem raises questions regarding the interpretation of the concept of probability involved in $P(A_i)$ and therefore in $P(A_i \mid B)$
- The frequentist interpretation is not flexible enough; we need to resort to its subjective interpretation

We need to extend the classical notion of statistical model in order to introduce Bayesian methodology. In (parametric) Statistics, we have $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ as a collection of possible probabilistic models for the observable data $\mathbf{X}$; however,

We need to extend the classical notion of statistical model in order to introduce Bayesian methodology. In (parametric) Statistics, we have $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ as a collection of possible probabilistic models for the observable data $\mathbf{X}$; however,

- in frequentist Statistics, θ is unknown but treated as fixed

We need to extend the classical notion of statistical model in order to introduce Bayesian methodology. In (parametric) Statistics, we have $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ as a collection of possible probabilistic models for the observable data $\mathbf{X}$; however,

- in frequentist Statistics, θ is unknown but treated as fixed
- in Bayesian statistics, all unknowns are regarded as random quantities because everything that is unknown is uncertain and all uncertainty must be quantified using the language of probability — probability distribution on the parameter space Θ denoted by $\pi(\theta)$ and referred to as prior distribution

$$\pi(\theta) - \text{prior distribution}$$
$$f(\mathbf{x} \mid \boldsymbol{\theta}) - \text{likelihood function}$$
$$\Downarrow$$
$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\ \pi(\theta)}{\int_\Theta f(\mathbf{x} \mid \theta)\ \pi(\theta)\ d\theta}, \quad \theta \in \Theta - \text{posterior distribution}$$

Notes:

- $\pi(\theta)\ f(\mathbf{x} \mid \theta) = \pi(\theta, \mathbf{x})$ defines a joint distribution on $(\mathcal{X}, \Theta)$
- $m(\mathbf{x}) = \int_\Theta f(\mathbf{x} \mid \theta)\ \pi(\theta)\ d\theta$ is the so-called prior predictive distribution of the data $\mathbf{x}$
- Another way of writing Bayes' theorem is $\pi(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta)\ \pi(\theta)$ where the normalization constant $m(\mathbf{x})$ is omitted

Example

Suppose $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathrm{Ber}(\theta)$ and that *a priori* $\theta \sim \mathrm{Be}(a, b)$, $a, b > 0$ known.

Then, with $t = \sum_{i=1}^n x_i$,

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t (1-\theta)^{n-t}$$

and

$$\pi(\theta) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in (0,1) \ .$$

We can do the calculations to conclude that

$$m(\mathbf{x}) = \frac{B(t+a, n-t+b)}{B(a,b)}$$

Hence,

$$\pi(\theta \mid \mathbf{x}) = \frac{1}{B(t+a, n-t+b)} \theta^{t+a-1} (1-\theta)^{n-t+b-1}$$

that is,

$$\theta \mid \mathbf{x} \sim \mathrm{Be}\left(t+a, n-t+b\right)$$

Observations:

1. If two likelihood functions are proportional, they lead to the same posterior distribution. Implications:
   1.1 Bayesian inference only depends on observed data through the observed value of a sufficient statistic
   1.2 $\pi(\theta \mid \mathbf{x}) = \pi(\theta \mid \mathbf{T}(\mathbf{x}))$ if $\mathbf{T}$ is sufficient for θ
   1.3 Bayesian inference respects the sufficiency principle
   1.4 Bayesian inference only depends on the statistical model through the likelihood function $L(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta)$
   1.5 Bayesian inference respects the likelihood principle

Observations (ctd):

2. $\pi(\theta \mid \mathbf{x})$, $\theta \in \Theta$, contains all the available information about θ, combining the data (through $L(\theta \mid \mathbf{x})$) with the prior information (in $\pi(\theta)$)
3. The Bayesian operation of combining knowledge has a sequential nature: Suppose that $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ with $\mathbf{X}_1 \amalg \mathbf{X}_2 \mid \theta$. Then,

$$\pi(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\ \pi(\theta)}{\int f(\mathbf{x} \mid \theta)\ \pi(\theta)\ d\theta} = \frac{f(\mathbf{x}_2 \mid \theta)\ \pi(\theta \mid \mathbf{x}_1)}{\int f(\mathbf{x}_2 \mid \theta)\ \pi(\theta \mid \mathbf{x}_1)\ d\theta}$$

That is: $\pi(\theta \mid \mathbf{x})$ can also be viewed as resulting from updating the "prior" $\pi(\theta \mid \mathbf{x}_1)$ with the likelihood $f(\mathbf{x}_2 \mid \theta)$

Example

Suppose $X_1, \ldots, X_n \mid \lambda \overset{iid}{\sim} \mathrm{Po}(\lambda)$ and that *a priori* $\lambda \sim \mathrm{Ga}(a, b)$, $a, b > 0$ known, that is,

$$\pi(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}, \quad \lambda > 0.$$

Then, with $t = \sum x_i$, we have

$$L(\lambda \mid \mathbf{x}) \propto \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \propto e^{-n\lambda} \lambda^t$$

$$\begin{aligned} \pi(\lambda \mid \mathbf{x}) \propto f(\mathbf{x} \mid \lambda)\, \pi(\lambda) &\propto e^{-n\lambda} \lambda^t \times \lambda^{a-1} e^{-b\lambda} \\ &\propto \mathrm{Ga}(\lambda \mid t + a, n + b) \end{aligned}$$

and as a consequence we have that $\lambda \mid \mathbf{x} \sim \mathrm{Ga}(t + a, n + b)$.

Note that (Candidate's formula)

$$m(\mathbf{x}) = \frac{f(\mathbf{x} \mid \theta)\ \pi(\theta)}{\pi(\theta \mid \mathbf{x})} \quad \forall \theta \in \Theta$$

so that in this case we get the prior predictive of $\mathbf{X}$ is

$$m(\mathbf{x}) = b^a \frac{\Gamma(t+a)}{\Gamma(a)} \prod_{i=1}^{n} (x_i!)^{-1}\ (n+b)^{-(t+a)}$$

How do we go about addressing inferential questions within the Bayesian framework?

How do we go about addressing inferential questions within the Bayesian framework?

- The complete answer to this question requires the introduction of Statistical Decision Theory ideas: action space, state space, loss function, etc

How do we go about addressing inferential questions within the Bayesian framework?

- The complete answer to this question requires the introduction of Statistical Decision Theory ideas: action space, state space, loss function, etc
- However, the posterior distribution contains all the relevant information about θ, it's all a matter of finding its optimal summary

How do we go about addressing inferential questions within the Bayesian framework?

- The complete answer to this question requires the introduction of Statistical Decision Theory ideas: action space, state space, loss function, etc
- However, the posterior distribution contains all the relevant information about θ, it's all a matter of finding its optimal summary
- If the goal is to find a point estimate of θ, we can use as an estimate

How do we go about addressing inferential questions within the Bayesian framework?

- The complete answer to this question requires the introduction of Statistical Decision Theory ideas: action space, state space, loss function, etc
- However, the posterior distribution contains all the relevant information about θ, it's all a matter of finding its optimal summary
- If the goal is to find a point estimate of θ, we can use as an estimate
  - the mode of $\pi(\theta \mid \mathbf{x})$, the posterior mode
  - the posterior mean $E(\theta \mid \mathbf{x})$
  - the posterior median, etc

How do we go about addressing inferential questions within the Bayesian framework?

- The complete answer to this question requires the introduction of Statistical Decision Theory ideas: action space, state space, loss function, etc
- However, the posterior distribution contains all the relevant information about θ, it's all a matter of finding its optimal summary
- If the goal is to find a point estimate of θ, we can use as an estimate
  - the mode of $\pi(\theta \mid \mathbf{x})$, the posterior mode
  - the posterior mean $E(\theta \mid \mathbf{x})$
  - the posterior median, etc
- If the goal is to estimate θ by an interval, we can obtain $(a(\mathbf{x}), b(\mathbf{x}))$ such that $P(\theta \in (a(\mathbf{x}), b(\mathbf{x})) \mid \mathbf{x}) = 0.95$

Prediction

- We observe $X_1, \ldots, X_n$ a random sample from $\{f(\cdot \mid \theta) : \theta \in \Theta\}$
- a prior on θ is set and the posterior $\pi(\theta \mid \mathbf{x})$ is computed
- we wish to predict an outcome Y whose probability distribution depends on θ

Determine the probability distribution of $Y \mid \mathbf{x}$, the posterior predictive distribution of Y

$$
\begin{aligned}
f(y \mid \mathbf{x}) &= \int_\Theta f(y, \theta \mid \mathbf{x}) \; d\theta \\
&= \int_\Theta f(y \mid \theta, \mathbf{x}) \; \pi(\theta \mid \mathbf{x}) \; d\theta \\
&= \int_\Theta f(y \mid \theta) \; \pi(\theta \mid \mathbf{x}) \; d\theta \;\; \text{if } Y \amalg \mathbf{X} \mid \theta
\end{aligned}
$$

Example

$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathrm{Ber}(\theta)$; *a priori* $\theta \sim \mathrm{Be}(a, b)$, $a, b > 0$ known.

We know that $\theta \mid \mathbf{x} \sim \mathrm{Be}(a + t, b + n - t)$. Suppose we want to predict the outcome of the next observation, independent of the previous, X_{n+1}. Then,

$$
\begin{aligned}
f(x_{n+1} \mid \mathbf{x}) &= \int_0^1 f(x_{n+1} \mid \theta)\, \pi(\theta \mid \mathbf{x})\, d\theta \\
&= \frac{B(a + t + x_{n+1}, b + n - t + 1 - x_{n+1})}{B(a + t, b + n - t)}, \quad x_{n+1} = 0, 1 \ .
\end{aligned}
$$

Example

$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathrm{Ber}(\theta)$; *a priori* $\theta \sim \mathrm{Be}(a, b)$, $a, b > 0$ known.

We know that $\theta \mid \mathbf{x} \sim \mathrm{Be}(a + t, b + n - t)$. Suppose we want to predict the outcome of the next observation, independent of the previous, X_{n+1}. Then,

$$f(x_{n+1} \mid \mathbf{x}) = \int_0^1 f(x_{n+1} \mid \theta)\, \pi(\theta \mid \mathbf{x})\, d\theta$$
$$= \frac{B(a + t + x_{n+1}, b + n - t + 1 - x_{n+1})}{B(a + t, b + n - t)}, \quad x_{n+1} = 0, 1 \ .$$

It would be simpler to use the formula of the iterated expectation:

$$P(X_{n+1} = 1 \mid \mathbf{x}) = E[E_\theta[X_{n+1} \mid \theta, \mathbf{x}]\mathbf{x}] = E[\theta \mid \mathbf{x}] = \frac{a + t}{a + b + n}$$

Bayesian inference is conceptually very simple and particularly intuitive. However, its practical implementation is often considered difficult:

Bayesian inference is conceptually very simple and particularly intuitive. However, its practical implementation is often considered difficult:

- $\pi(\theta)$ should reflect information about θ available before the data $\mathbf{x}$ are observed. To summarize information that in general will exist in an non-organized fashion in a probability distribution is not trivial

Bayesian inference is conceptually very simple and particularly intuitive. However, its practical implementation is often considered difficult:

- $\pi(\theta)$ should reflect information about θ available before the data $\mathbf{x}$ are observed. To summarize information that in general will exist in an non-organized fashion in a probability distribution is not trivial
- What should one do when said information is vague or diffuse?

Bayesian inference is conceptually very simple and particularly intuitive. However, its practical implementation is often considered difficult:

- $\pi(\theta)$ should reflect information about θ available before the data $\mathbf{x}$ are observed. To summarize information that in general will exist in an non-organized fashion in a probability distribution is not trivial
- What should one do when said information is vague or diffuse?
- What if the goal is to produce a statistical analysis which is as "objective" as possible, e.g. one that uses little prior information about θ?

Bayesian inference is conceptually very simple and particularly intuitive. However, its practical implementation is often considered difficult:

- $\pi(\theta)$ should reflect information about θ available before the data $\mathbf{x}$ are observed. To summarize information that in general will exist in an non-organized fashion in a probability distribution is not trivial
- What should one do when said information is vague or diffuse?
- What if the goal is to produce a statistical analysis which is as "objective" as possible, e.g. one that uses little prior information about θ?
- Calculations: Very rarely will $\pi(\theta \mid \mathbf{x})$ exist in closed form, as $m(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\, \pi(\theta)\, d\theta$ will not be computable analytically

Bayesian inference is conceptually very simple and particularly intuitive. However, its practical implementation is often considered difficult:

- $\pi(\theta)$ should reflect information about θ available before the data $\mathbf{x}$ are observed. To summarize information that in general will exist in an non-organized fashion in a probability distribution is not trivial
- What should one do when said information is vague or diffuse?
- What if the goal is to produce a statistical analysis which is as "objective" as possible, e.g. one that uses little prior information about θ?
- Calculations: Very rarely will $\pi(\theta \mid \mathbf{x})$ exist in closed form, as $m(\mathbf{x}) = \int f(\mathbf{x} \mid \theta)\, \pi(\theta)\, d\theta$ will not be computable analytically
- The answer to many inferential questions will involve the calculation of $E[\psi(\theta) \mid \mathbf{x}]$ for different $\psi(\theta)$

"Solutions":

- prior distributions which allow analytical calculations
- "non-informative" prior distributions
- Simulation, analytic approximations, numerical calculations

Families of prior distributions which allow for analytical calculations.

Example

Suppose $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathrm{Ber}(\theta)$; *a priori* $\theta \sim \mathrm{Be}(a, b)$, $a, b > 0$ known.

We saw that

$$\theta \mid \mathbf{x} \sim \mathrm{Be}\left(t + a, n - t + b\right)$$

that is, the updating is done within the same family of distributions:

$$(a, b) \longrightarrow (t + a, n - t + b)$$

Definition

The family $\Pi = \{\pi(\cdot \mid \tau) : \tau \in \Gamma\}$ is said to be natural conjugate of the statistical model $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$ if

1. $\forall \tau_0, \tau_1 \in \Gamma \; \exists \tau_2 \in \Gamma$:
$$\pi(\theta \mid \tau_0)\, \pi(\theta \mid \tau_1) \propto \pi(\theta \mid \tau_2)$$
2. $\exists \tau_0 \in \Gamma : f(\mathbf{x} \mid \theta) \propto \pi(\theta \mid \tau_0)$

Consequence:

$$\begin{aligned}\pi(\theta \mid \mathbf{x}) &\propto f(\mathbf{x} \mid \theta)\, \pi(\theta \mid \tau_1) \\ &\propto \pi(\theta \mid \tau_0)\, \pi(\theta \mid \tau_1) \\ &\propto \pi(\theta \mid \tau_2) \in \Pi\end{aligned}$$

Example

Suppose $X_i,\ i = 1, \ldots, n \overset{iid}{\sim} \mathrm{Po}(\lambda)$.

Example

Suppose $X_i, \; i = 1, \ldots, n \overset{iid}{\sim} \mathrm{Po}(\lambda)$.

Then, with $t = \sum x_i$,

$$
\begin{aligned}
f(\mathbf{x} \mid \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\
&\propto \lambda^{t} e^{-n\lambda} \\
&\propto \mathrm{Ga}(\lambda \mid t+1, n)
\end{aligned}
$$

Example

Suppose $X_i,\ i = 1, \dots, n \overset{iid}{\sim} \mathrm{Po}(\lambda)$.

Then, with $t = \sum x_i$,

$$
\begin{aligned}
f(\mathbf{x} \mid \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\
&\propto \lambda^t e^{-n\lambda} \\
&\propto \mathrm{Ga}(\lambda \mid t+1, n)
\end{aligned}
$$

Also, $\mathrm{Ga}(\lambda \mid a, b) \times \mathrm{Ga}(\lambda \mid c, d) \propto \mathrm{Ga}(\lambda \mid a + c - 1, b + d)$.

Example

Suppose $X_i, \; i = 1, \ldots, n \overset{iid}{\sim} \mathrm{Po}(\lambda)$.

Then, with $t = \sum x_i$,

$$\begin{aligned} f(\mathbf{x} \mid \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &\propto \lambda^{t} e^{-n\lambda} \\ &\propto \mathrm{Ga}(\lambda \mid t+1, n) \end{aligned}$$

Also, $\mathrm{Ga}(\lambda \mid a, b) \times \mathrm{Ga}(\lambda \mid c, d) \propto \mathrm{Ga}(\lambda \mid a + c - 1, b + d)$. Hence, the gamma family is the natural conjugate of the Poisson model. The prior-to-posterior update is $(a, b) \rightarrow (a + t, b + n)$.

Choosing (a, b):

- Set $E(\theta) = \mu_0$ and $\mathrm{Var}(\theta) = \sigma_0^2$ subjectively. Then solve $a/b = \mu_0$ and $a/b^2 = \sigma_0^2$.
- $\mathrm{Ga}(a, b)$ contains the same information as an imaginary sample of "size" b and sample total a:

$$(a, b) \longrightarrow (a + t, b + n)$$

- Treat a, b as unknown and place a prior on them, $\pi(a, b)$ - hierarchical prior

Choosing (a, b):

- Set $E(\theta) = \mu_0$ and $\mathrm{Var}(\theta) = \sigma_0^2$ subjectively. Then solve $a/b = \mu_0$ and $a/b^2 = \sigma_0^2$.
- $\mathrm{Ga}(a, b)$ contains the same information as an imaginary sample of "size" b and sample total a:

$$(a, b) \to (a + t, b + n)$$

- Treat a, b as unknown and place a prior on them, $\pi(a, b)$ - hierarchical prior

Drawbacks:

- Conjugate family does not always exist
- Functional form is chosen for convenience and it may have important consequences

Non-informative priors

- Situations where there is no considerable prior information
- Obtain posterior beliefs in situations where the sampling information should overwhelm the prior information
- Obtain a “reference” analysis, an “objective” analysis which can be compared with subjective ones as a way of ascertaining the influence of the prior information
- Research area called “Objective Bayes” — methods or strategies to obtain “objective” priors in various situations which are then evaluated

Bayes-Laplace method

Principle of insufficient reason of Bayes-Laplace: in the absence of any reason to consider that two probabilities are different, they should be considered equal.

Consequences:

Bayes-Laplace method

Principle of insufficient reason of Bayes-Laplace: in the absence of any reason to consider that two probabilities are different, they should be considered equal.

Consequences:

- Θ finite, $\Theta = \{\theta_1, \ldots, \theta_k\}$, then $\pi(\theta_i) = 1/k$, $i = 1, \ldots, k$

Bayes-Laplace method

Principle of insufficient reason of Bayes-Laplace: in the absence of any reason to consider that two probabilities are different, they should be considered equal.

Consequences:

- Θ finite, $\Theta = \{\theta_1, \ldots, \theta_k\}$, then $\pi(\theta_i) = 1/k$, $i = 1, \ldots, k$
- If Θ is countable, there is no probability distribution which is compatible with this principle: $\pi(\theta) = c$, $\theta \in \{\theta_1, \ldots, \theta_k, \ldots\}$ implies that $\sum_{\theta\in\Theta} \pi(\theta) = +\infty$: It's an **improper** distribution

Bayes-Laplace method

Principle of insufficient reason of Bayes-Laplace: in the absence of any reason to consider that two probabilities are different, they should be considered equal.

Consequences:

- Θ finite, $\Theta = \{\theta_1, \ldots, \theta_k\}$, then $\pi(\theta_i) = 1/k$, $i = 1, \ldots, k$
- If Θ is countable, there is no probability distribution which is compatible with this principle: $\pi(\theta) = c$, $\theta \in \{\theta_1, \ldots, \theta_k, \ldots\}$ implies that $\sum_{\theta\in\Theta} \pi(\theta) = +\infty$: It's an **improper** distribution
- The formal use of Bayes' theorem with an improper prior is controversial; however, it's often utilized as long as the resulting posterior is proper

Bayes-Laplace method

Principle of insufficient reason of Bayes-Laplace: in the absence of any reason to consider that two probabilities are different, they should be considered equal.

Consequences:

- Θ finite, $\Theta = \{\theta_1, \ldots, \theta_k\}$, then $\pi(\theta_i) = 1/k$, $i = 1, \ldots, k$
- If Θ is countable, there is no probability distribution which is compatible with this principle: $\pi(\theta) = c$, $\theta \in \{\theta_1, \ldots, \theta_k, \ldots\}$ implies that $\sum_{\theta \in \Theta} \pi(\theta) = +\infty$: It's an **improper** distribution
- The formal use of Bayes' theorem with an improper prior is controversial; however, it's often utilized as long as the resulting posterior is proper
- Θ not countable: $\pi(\theta) \propto c$, $\theta \in \Theta$ is improper unless Θ is bounded

Most important objection to uniform priors:

Example

$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathrm{Ber}(\theta)$.

The Bayes-Laplace prior would be $\pi(\theta) = 1$, $\theta \in (0, 1)$. An alternative parameterization of the Bernoulli model is in terms of $\psi = \ln[\theta/(1-\theta)]$. The induced distribution in ψ is

$$\pi(\psi) = \frac{e^{\psi}}{(1+e^{\psi})^2}, \; \psi \in \mathbb{R}$$

Ignorance about θ implies some information about ψ!
In general, with $\theta = g(\psi)$,

$$\pi(\psi) = |g'(\psi)| \pi(g(\psi))$$

Jeffreys Method

Idea: invariance with respect to reparametrizations.
Let $\theta = g(\psi)$ and denote by $I_X(\theta)$ the Fisher information about θ in X. Then, the Fisher information about ψ in X is

$$I^*_X(\psi) = [g'(\psi)]^2 \; I_X(g(\psi)) \; .$$

If *a priori*

$$\pi(\theta) \propto \sqrt{I_X(\theta)}$$

then the induce prior on ψ is

$$\begin{aligned}\pi(\psi) &= |g'(\psi)| \; \pi(g(\psi)) \\ &= |g'(\psi)| \sqrt{I_X(g(\psi))} \\ &= \sqrt{I^*_X(\psi)}\end{aligned}$$

It does not matter to which parameterization we apply the rule!

Example

$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} \mathrm{Ber}(\theta)$

Recall that $I_X(\theta) = E_\theta[-d^2 \ln f(X \mid \theta)/d\theta^2]$. Hence,

$$I_X(\theta) = E_\theta[X/\theta^2 - (1-X)/(1-\theta)^2] = \theta^{-1}(1-\theta)^{-1}$$

and so

$$\pi^J(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto \mathrm{Be}(\theta|1/2, 1/2)$$

Example

$X_1, \ldots, X_n \mid \mu \stackrel{iid}{\sim} \mathrm{N}(\mu, 1)$

Easy calculations show that $I_X(\mu) = 1$, so,

$$\pi^J(\mu) \propto c, \quad \mu \in \mathbb{R}$$

which is an improper distribution. However, the formal use of Bayes' Theorem leads to

$$\mu \mid x_1, \ldots, x_n \sim \mathrm{N}(\bar{x}, 1/n)$$

Ordinary Monte Carlo

Theorem

Law of Large Numbers: Suppose $\{X_i\}$ is a sequence of iid random variables with $E[X_i] = \mu$. Then, with $\bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i$

$$\bar{X}_M \xrightarrow{a.s.} \mu$$

Ordinary Monte Carlo

Theorem

Law of Large Numbers: Suppose $\{X_i\}$ is a sequence of iid random variables with $E[X_i] = \mu$. Then, with $\bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i$

$$\bar{X}_M \xrightarrow{a.s.} \mu$$

- One common application is to justify the approximation of $E[X]$ by $\bar{x}_M$ when $x_1, \ldots, x_M$ are observed data

Ordinary Monte Carlo

Theorem

Law of Large Numbers: Suppose $\{X_i\}$ is a sequence of iid random variables with $E[X_i] = \mu$. Then, with $\bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i$

$$\bar{X}_M \xrightarrow{a.s.} \mu$$

- One common application is to justify the approximation of $E[X]$ by $\bar{x}_M$ when $x_1, \ldots, x_M$ are observed data
- Another application: represent approximately one probability distribution by a computer-generated sample $x_1, \ldots, x_M$ simulated from this distribution

Ordinary Monte Carlo

Theorem

Law of Large Numbers: Suppose $\{X_i\}$ is a sequence of iid random variables with $E[X_i] = \mu$. Then, with $\bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i$

$$\bar{X}_M \xrightarrow{a.s.} \mu$$

- One common application is to justify the approximation of $E[X]$ by $\bar{x}_M$ when $x_1, \ldots, x_M$ are observed data
- Another application: represent approximately one probability distribution by a computer-generated sample $x_1, \ldots, x_M$ simulated from this distribution
- (Almost) all the aspects of this probability distribution can be arbitrarily approximated using exclusively $x_1, \ldots, x_M$ for large enough M

Facts

Facts

- Expectations: $E[\psi(X)] \approx \frac{1}{M}\sum_{i=1}^{M}\psi(x_i)$

Facts

- Expectations: $E[\psi(X)] \approx \frac{1}{M}\sum_{i=1}^{M}\psi(x_i)$
- Probabilities:

$$P(X \in A) \approx \frac{1}{M}\#\{i : x_i \in A\}$$

Facts

- Expectations: $E[\psi(X)] \approx \frac{1}{M}\sum_{i=1}^{M}\psi(x_i)$
- Probabilities:

$$P(X \in A) \approx \frac{1}{M}\#\{i : x_i \in A\}$$

- Densities: for small enough $\delta > 0$

$$f(a) \approx \frac{1}{\delta}\,\frac{1}{M}\#\{i : \delta < x_i \leq a + \delta\}$$

that is, the histogram is a good approximation to the density

Facts (ctd.)

- $\psi(x_1), \ldots, \psi(x_M)$ is a sample from the distribution of $\psi(X)$

Facts (ctd.)

- $\psi(x_1), \ldots, \psi(x_M)$ is a sample from the distribution of $\psi(X)$
- Suppose we can obtain a sample $(x_1, y_1), \ldots, (x_M, y_M)$ from the joint distribution $f(x, y)$ of (X, Y). Then, $x_1, \ldots, x_M$ is a sample from the marginal distribution of X

Facts (ctd.)

- $\psi(x_1), \ldots, \psi(x_M)$ is a sample from the distribution of $\psi(X)$
- Suppose we can obtain a sample $(x_1, y_1), \ldots, (x_M, y_M)$ from the joint distribution $f(x, y)$ of (X, Y). Then, $x_1, \ldots, x_M$ is a sample from the marginal distribution of X
- If y is a draw from the distribution of Y and x is a draw from the distribution of $X \mid y$, then (x, y) is a draw from the joint (X, Y)

Facts (ctd.)

- $\psi(x_1), \ldots, \psi(x_M)$ is a sample from the distribution of $\psi(X)$
- Suppose we can obtain a sample $(x_1, y_1), \ldots, (x_M, y_M)$ from the joint distribution $f(x, y)$ of (X, Y). Then, $x_1, \ldots, x_M$ is a sample from the marginal distribution of X
- If y is a draw from the distribution of Y and x is a draw from the distribution of $X \mid y$, then (x, y) is a draw from the joint (X, Y)
- Very important for prediction: if $Y \amalg \mathbf{X} \mid \theta$

  $$f(y \mid \mathbf{x}) = \int_\Theta f(y \mid \theta)\, \pi(\theta \mid \mathbf{x})\, d\theta$$

  To obtain a sample from $f(y \mid \mathbf{x})$ we need a sample from $\pi(\theta \mid \mathbf{x})$, $\theta_1, \ldots, \theta_M$, and to be able to simulate y_i from $f(y \mid \theta_i)$

Example

Generating from a t distribution with ν degrees of freedom: $X \sim t_\nu$ can be written as mixture:

$$X \mid Y = y \sim \mathrm{N}(0, \nu/y) \text{ and } Y \sim \chi^2_\nu$$

Algorithm: for $i = 1, \ldots, M$

- Generate y_i from χ^2_ν
- Generate x_i from $\mathrm{N}(0, \nu/y_i)$

$(x_1, \ldots, x_M)$ is a sample from t_ν

Statistical models are sometimes written in the form

$$f(x \mid \theta) = \int f(x, y \mid \theta)\, dy$$

either artificially (data augmentation) or as a natural consequence of the modeling strategy (eg, latent variable models) and that can be explored in order to facilitate sampling.

Statistical models are sometimes written in the form

$$f(x \mid \theta) = \int f(x, y \mid \theta)\ dy$$

either artificially (data augmentation) or as a natural consequence of the modeling strategy (eg, latent variable models) and that can be explored in order to facilitate sampling.

Example

Probit regression: $Y_i \mid \theta_i \sim \mathrm{Ber}(\theta_i)$ independently, with $\theta_i = \Phi(\mathbf{x}_i' \boldsymbol{\beta})$, where $\mathbf{x}_i$ corresponds to known covariate information.

If we let $Z_i \mid \beta \sim \mathrm{N}(\mathbf{x}_i' \boldsymbol{\beta}, 1)$ and $Y_i = I_{(0,+\infty)}(Z_i)$ it's easy to see that

$$P(Y_i = 1) = \Phi(\mathbf{x}_i' \boldsymbol{\beta})$$

so that if we obtain a sample from $\boldsymbol{\beta}, \mathbf{Z} \mid \mathbf{y}$ we obtain also a sample from $\boldsymbol{\beta} \mid \mathbf{y}$.

MCMC

- Problem: in most cases, it will be very difficult to obtain a sample of simulated iid observations from $\pi(\theta \mid \mathbf{x})$, especially if $m(x)$ is unknown
- MCMC methods allow us to construct (even in situations where $m(\mathbf{x})$ is unknown) a Markov chain $\{\theta_n\}$ whose stationary (limiting) distribution is $\pi(\theta \mid \mathbf{x})$
- Additionally it is still the case that

$$\frac{1}{M}\sum_{n=1}^{M}\psi(\theta_n) \stackrel{as}{\rightarrow} E[\psi(\theta) \mid \mathbf{x}]$$

- Robert and Casella (2004). *Monte Carlo Statistical Methods*. Springer.

Gibbs Sampler

- Suppose $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$
- Let $\boldsymbol{\theta}_{(-i)} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_p)$
- Let $\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{(-i)}, \mathbf{x} \sim f_i(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{(-i)})$
- the density f_i is called the full-conditional of $\boldsymbol{\theta}_i$
- the Gibbs sampler proceeds by iteratively sampling from each of these full-conditionals to transition from the current state $\boldsymbol{\theta}^{(t)}$ to state $\boldsymbol{\theta}^{(t+1)}$

The Gibbs sampler algorithm:

Start at $\theta^{(0)}$. For $t = 1, 2, \ldots,$ generate

1- $\theta_1^{(t+1)} \sim f_1(\theta_1 \mid \theta_2^{(t)}, \ldots, \theta_p^{(t)})$

2- $\theta_2^{(t+1)} \sim f_2(\theta_2 \mid \theta_1^{(t+1)}, \theta_2^{(t)}, \ldots, \theta_p^{(t)})$

3- $\theta_3^{(t+1)} \sim f_3(\theta_3 \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)} \ldots, \theta_p^{(t)})$

$\ldots$

p- $\theta_p^{(t+1)} \sim f_p(\theta_p \mid \theta_1^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)})$

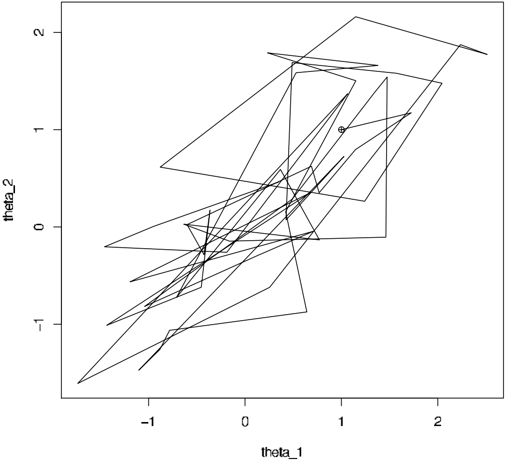
Example

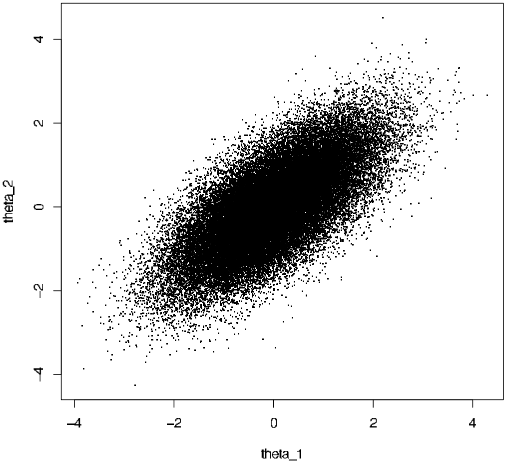
Let $\boldsymbol{\theta} = (\theta_1, \theta_2) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

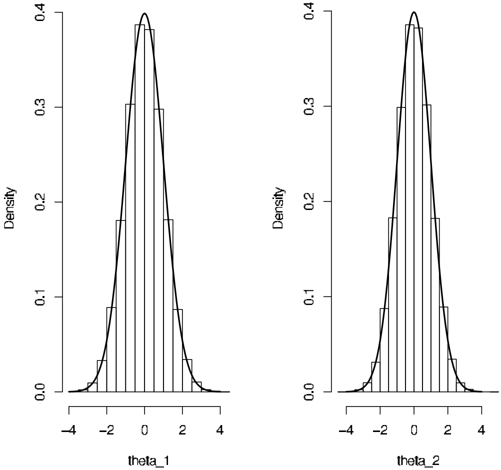
Gibbs sampler to obtain a sample from this probability distribution: if the current state is $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)})$ to obtain the next state generate

$$\theta_1^{(t+1)} \sim \mathrm{N}(\rho\theta_2^{(t)}, 1 - \rho^2)$$
$$\theta_2^{(t+1)} \sim \mathrm{N}(\rho\theta_1^{(t+1)}, 1 - \rho^2)$$


theta_2
theta_1
2
1
0
−1
−1
0
1
2


4
2
0
-2
-4
theta_2
-4
-2
0
2
4
theta_1

Markov chain Monte Carlo



The Metropolis-Hastings Algorithm

We need a conditional density $q(\theta \mid \theta')$ called the instrumental or proposal density. The target is the posterior $\pi(\theta \mid \mathbf{x})$.

Start at $\theta^{(0)}$. For $t = 1, 2, \ldots,$

1. Generate $\theta^* \sim q(\theta \mid \theta^{(t)})$
2. Take

$$\theta^{(t+1)} = \begin{cases} \theta^* \text{ with probability } \rho(\theta^{(t)}, \theta^*) \\ \theta^{(t)} \text{ with probability } 1 - \rho(\theta^{(t)}, \theta^*) \end{cases}$$

   where

$$\rho(\theta^{(t)}, \theta^*) = \min\left\{ \frac{\pi(\theta^* \mid \mathbf{x})}{\pi(\theta^{(t)} \mid \mathbf{x})} \; \frac{q(\theta^{(t)} \mid \theta^*)}{q(\theta^* \mid \theta^{(t)})}, 1 \right\}$$

Observations:

Observations:

- To compute the acceptance ratio ρ we do not need to know $m(\mathbf{x})$

Observations:

- To compute the acceptance ratio ρ we do not need to know $m(\mathbf{x})$
- The algorithm is implementable in practice if $q(\cdot \mid \theta')$ is easy to simulate from and is either available explicitly (up to a constant independent of θ') or symmetric, ie $q(\theta \mid \theta') = q(\theta' \mid \theta)$

Observations:

- To compute the acceptance ratio ρ we do not need to know $m(\mathbf{x})$
- The algorithm is implementable in practice if $q(\cdot \mid \theta')$ is easy to simulate from and is either available explicitly (up to a constant independent of θ') or symmetric, ie $q(\theta \mid \theta') = q(\theta' \mid \theta)$
- with very minor restrictions on the support of the proposal, the algorithm works in *theory*

Independent Metropolis-Hastings:

- $q(\theta \mid \theta') = q(\theta)$
- close connections to the accept-reject method
- $q(\theta)$ is typically designed to closely approximate the target (eg, analytic approximations to the posterior)

Independent Metropolis-Hastings:

- $q(\theta \mid \theta') = q(\theta)$
- close connections to the accept-reject method
- $q(\theta)$ is typically designed to closely approximate the target (eg, analytic approximations to the posterior)

Random walk Metropolis-Hastings:

- $q(\theta \mid \theta') = q(\theta - \theta')$, ie $\theta^* = \theta^{(t)} + \varepsilon_t$ with ε_t a random perturbation with density q independent of $\theta^{(t)}$
- Typical choices for q are uniform, normal or t centered at the origin and appropriately scaled

Metropolis-within-Gibbs or Hybrid MCMC:

- The Gibbs sampler as described can only be implemented if we can directly generate from all the full-conditionals $f_i(\theta_i \mid \theta_{(-i)})$

Metropolis-within-Gibbs or Hybrid MCMC:

- The Gibbs sampler as described can only be implemented if we can directly generate from all the full-conditionals $f_i(\theta_i \mid \theta_{(-i)})$
- However, the algorithm is still valid if simulation from the ith full conditional is replaced by a Metropolis-Hastings step, that is, a simulation from a proposal which is accepted according to a M-H ratio

Metropolis-within-Gibbs or Hybrid MCMC:

- The Gibbs sampler as described can only be implemented if we can directly generate from all the full-conditionals $f_i(\theta_i \mid \theta_{(-i)})$
- However, the algorithm is still valid if simulation from the ith full conditional is replaced by a Metropolis-Hastings step, that is, a simulation from a proposal which is accepted according to a M-H ratio
- Typically, a number of M-H steps are done and only the last is retained (to reduce auto-correlation)

Practical considerations:

Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs

Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs
- Parameters whose full-conditionals do not have standard form: M-H step with the scale of the proposal tuned so that the acceptance rate is about 20% (vector of parameters) or 40% (scalar parameter)

Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs
- Parameters whose full-conditionals do not have standard form: M-H step with the scale of the proposal tuned so that the acceptance rate is about 20% (vector of parameters) or 40% (scalar parameter)
- run multiple chains starting at different values

Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs
- Parameters whose full-conditionals do not have standard form: M-H step with the scale of the proposal tuned so that the acceptance rate is about 20% (vector of parameters) or 40% (scalar parameter)
- run multiple chains starting at different values
- Look at traceplots to empirically ascertain convergence and decide about the length of burn-in

Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs
- Parameters whose full-conditionals do not have standard form: M-H step with the scale of the proposal tuned so that the acceptance rate is about 20% (vector of parameters) or 40% (scalar parameter)
- run multiple chains starting at different values
- Look at traceplots to empirically ascertain convergence and decide about the length of burn-in
- Thinning: retaining only the mth iteration

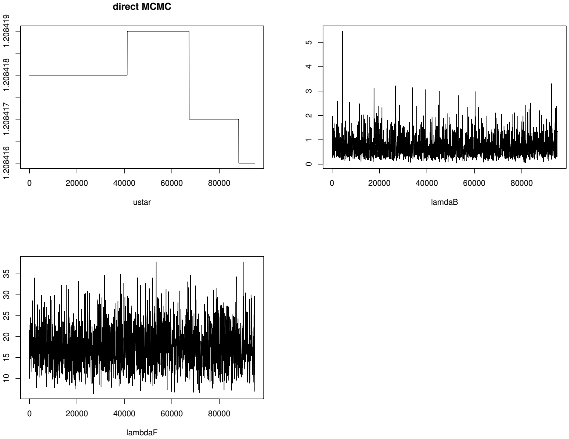
Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs
- Parameters whose full-conditionals do not have standard form: M-H step with the scale of the proposal tuned so that the acceptance rate is about 20% (vector of parameters) or 40% (scalar parameter)
- run multiple chains starting at different values
- Look at traceplots to empirically ascertain convergence and decide about the length of burn-in
- Thinning: retaining only the mth iteration
- Plots of autocorrelation functions to identify highly correlated chains

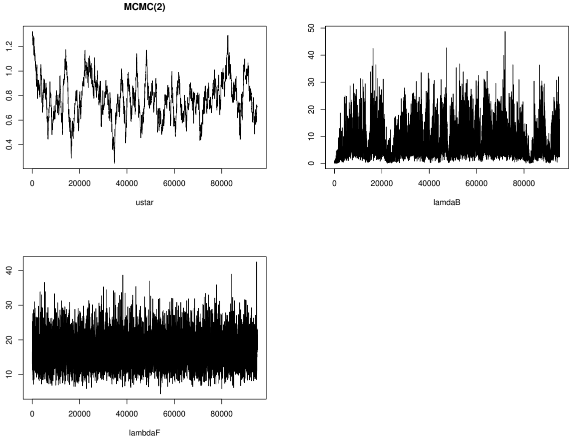
Practical considerations:

- Look at full-conditionals; for the parameters whose full-conditionals have standard form, use Gibbs
- Parameters whose full-conditionals do not have standard form: M-H step with the scale of the proposal tuned so that the acceptance rate is about 20% (vector of parameters) or 40% (scalar parameter)
- run multiple chains starting at different values
- Look at traceplots to empirically ascertain convergence and decide about the length of burn-in
- Thinning: retaining only the mth iteration
- Plots of autocorrelation functions to identify highly correlated chains
- WinBUGS is a popular software which automatically implements Bayesian analysis via MCMC

Markov chain Monte Carlo



Markov chain Monte Carlo



A few books

- Robert, C. (2001). *The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation*. Spinger
- Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. Springer
- Gelman et al. (2004). *Bayesian Data Analysis*. Chapman & Hall.
- Marin, JM and Robert, C. (2007). *Bayesian Core: a Practical Approach to Computational Bayesian Statistics*. Springer.
- Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.