An introduction to Bayesian statistical inference

S. Robin

INRA / AgroParisTech /univ. Paris-Saclay

JC(2)BIM, June 2018, Fréjus

Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

An example

Example:

- n patients: $i = 1 \dots n$
- Y_i = status (0 = healthy, 1 = sick) of patient i
- $\mathbf{x}_i = (x_{i1}, \dots x_{ip})$ = vector of gene expression for patient i (gene $j = 1 \dots p$)

Dataset: $n = 78$, $p = 15$

	AB033066	NM003056	NM000903	...	Status
1	0.178	0.116	0.22		0
2	0.065	-0.073	-0.014		0
3	-0.077	0.03	0.043		0
4	0.176	-0.041	0.362		0
5	-0.089	-0.164	-0.266		0

An example

Example:

- n patients: $i = 1 \dots n$
- Y_i = status (0 = healthy, 1 = sick) of patient i
- $\mathbf{x}_i = (x_{i1}, \dots x_{ip})$ = vector of gene expression for patient i (gene $j = 1 \dots p$)

Dataset: $n = 78$, $p = 15$

	AB033066	NM003056	NM000903	...	Status
1	0.178	0.116	0.22		0
2	0.065	-0.073	-0.014		0
3	-0.077	0.03	0.043		0
4	0.176	-0.041	0.362		0
5	-0.089	-0.164	-0.266		0

Similar question for genotyping data.

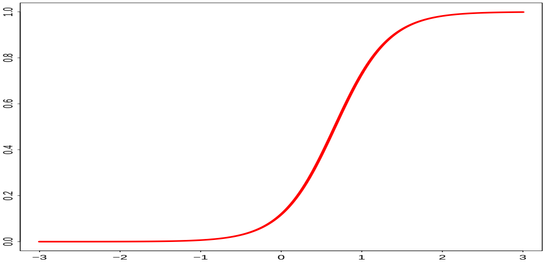
A statistical model

Logistic regression Logistic regression

- The patients are independent.
- The probability for patient i to be sick depends on $\mathbf{x}_i$:

$$\Pr\{Y_i = 1\} = \frac{e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}, \qquad \mathbf{x}_i^\intercal \boldsymbol{\theta} = \sum_{j=1}^{p} x_{ij}\theta_j$$

- $\boldsymbol{\theta} = (\theta_1, \dots \theta_p)$: unknown parameter (regression coefficients, incl. intercept)



Frequentist inference

$\boldsymbol{\theta}$ = fixed parameter:

- Statistical model:

$$\mathbf{Y} \sim p_{\boldsymbol{\theta}}$$

- Inference: get a (point) estimate $\widehat{\boldsymbol{\theta}}$ e.g.

$$\widehat{\boldsymbol{\theta}}: \qquad \log p_{\widehat{\boldsymbol{\theta}}}(\mathbf{Y}) = \max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y})$$

- The estimate $\widehat{\boldsymbol{\theta}}$ itself is random (depends on the data) $\rightarrow$ confidence interval, tests, ...

Frequentist inference

$\boldsymbol{\theta}$ = fixed parameter:

- Statistical model:

$$\mathbf{Y} \sim p_{\boldsymbol{\theta}}$$

- Inference: get a (point) estimate $\widehat{\boldsymbol{\theta}}$ e.g.

$$\widehat{\boldsymbol{\theta}}: \qquad \log p_{\widehat{\boldsymbol{\theta}}}(\mathbf{Y}) = \max_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y})$$

- The estimate $\widehat{\boldsymbol{\theta}}$ itself is random (depends on the data) $\rightarrow$ confidence interval, tests, ...

Output: `GLM = glm(Y ∼ X, family=binomial)`

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.7212697	0.6512707	-1.107481	0.2680861
XAB033066	7.23375	2.505118	2.887589	0.003882068
XNM003056	-0.6116423	1.854695	-0.3297806	0.7415658
XNM000903	1.732625	1.199888	1.443988	0.1487423
...				

Bayesian inference

$\boldsymbol{\theta}$ = random parameter:

- Statistical model:

$$\ell(\mathbf{Y} \mid \boldsymbol{\theta}) := p(\mathbf{Y} \mid \boldsymbol{\theta}) \qquad (= \textit{likelihood})$$

- Inference: provide the conditional distribution of $\boldsymbol{\theta}$ given the observed data $\mathbf{Y}$:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \qquad (= \textit{posterior} \text{ distribution})$$

$\rightarrow$ credibility intervals

- Requires to define a marginal distribution:

$$\pi(\boldsymbol{\theta}) := p(\boldsymbol{\theta}) \qquad (= \textit{prior} \text{ distribution})$$

Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

Why 'Bayes'

Bayes formula:

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(A)}{P(B)} P(B \mid A)$$

- $P(B)$ = marginal probability of B
- $P(A, B)$ = joint probability of A and B
- $P(A \mid B)$ = conditional probability of A given B [# 62]

Why 'Bayes'

Bayes formula:

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(A)}{P(B)} P(B \mid A)$$

- $P(B) =$ marginal probability of B
- $P(A, B) =$ joint probability of A and B
- $P(A \mid B) =$ conditional probability of A given B [# 62]

Be careful. Many methods, e.g.

Bayesian network, Naive Bayes, ...

- use conditional probabilities
- but have nothing to do with Bayesian inference (in the statistical sense)

Bayes formula for Bayesian inference (1/2)

Posterior distribution.

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{p(\mathbf{Y}, \boldsymbol{\theta})}{p(\mathbf{Y})} = \frac{\overbrace{\pi(\boldsymbol{\theta})}^{\textit{prior}} \; \overbrace{\ell(\mathbf{Y} \mid \boldsymbol{\theta})}^{\textit{likelihood}}}{p(\mathbf{Y})}$$

→ Requires to evaluate the *integrated likelihood* (i.e. marginal)

$$p(\mathbf{Y}) = \int \pi(\boldsymbol{\theta}) \ell(\mathbf{Y} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta},$$

which act as the normalizing constant of the posterior $p(\boldsymbol{\theta} \mid \mathbf{Y})$.

Bayes formula for Bayesian inference (2/2)

Some remarks.

Bayes formula for Bayesian inference (2/2)

Some remarks.

1. $p(\cdot)$ is sometimes denoted $[\cdot]$:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{\pi(\boldsymbol{\theta}) \; \ell(\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y})} \qquad \Leftrightarrow \qquad [\boldsymbol{\theta} \mid \mathbf{Y}] = \frac{[\boldsymbol{\theta}] \; [\mathbf{Y} \mid \boldsymbol{\theta}]}{[\mathbf{Y}]}$$

Bayes formula for Bayesian inference (2/2)

Some remarks.

1. $p(\cdot)$ is sometimes denoted $[\cdot]$:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{\pi(\boldsymbol{\theta})\, \ell(\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y})} \qquad \Leftrightarrow \qquad [\boldsymbol{\theta} \mid \mathbf{Y}] = \frac{[\boldsymbol{\theta}]\, [\mathbf{Y} \mid \boldsymbol{\theta}]}{[\mathbf{Y}]}$$

2. Computing $p(\mathbf{Y})$ is generally (very) difficult: see Section 2

Bayes formula for Bayesian inference (2/2)

Some remarks.

1. $p(\cdot)$ is sometimes denoted $[\cdot]$:

$$p(\boldsymbol{\theta} \,|\, \mathbf{Y}) = \frac{\pi(\boldsymbol{\theta})\, \ell(\mathbf{Y} \,|\, \boldsymbol{\theta})}{p(\mathbf{Y})} \qquad \Leftrightarrow \qquad [\boldsymbol{\theta} \,|\, \mathbf{Y}] = \frac{[\boldsymbol{\theta}]\; [\mathbf{Y} \,|\, \boldsymbol{\theta}]}{[\mathbf{Y}]}$$

2. Computing $p(\mathbf{Y})$ is generally (very) difficult: see Section 2

3. Obviously

$$p(\boldsymbol{\theta} \,|\, \mathbf{Y}) \propto \pi(\boldsymbol{\theta})\, \ell(\mathbf{Y} \,|\, \boldsymbol{\theta}),$$

$\rightarrow$ $p(\boldsymbol{\theta} \,|\, \mathbf{Y})$ and $p(\boldsymbol{\theta}' \,|\, \mathbf{Y})$ can be compared, without computing $p(\mathbf{Y})$

Bayes formula for Bayesian inference (2/2)

Some remarks.

1. $p(\cdot)$ is sometimes denoted $[\cdot]$:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{\pi(\boldsymbol{\theta})\ \ell(\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y})} \qquad \Leftrightarrow \qquad [\boldsymbol{\theta} \mid \mathbf{Y}] = \frac{[\boldsymbol{\theta}]\ [\mathbf{Y} \mid \boldsymbol{\theta}]}{[\mathbf{Y}]}$$

2. Computing $p(\mathbf{Y})$ is generally (very) difficult: see Section 2

3. Obviously

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto \pi(\boldsymbol{\theta})\ \ell(\mathbf{Y} \mid \boldsymbol{\theta}),$$

$\rightarrow$ $p(\boldsymbol{\theta} \mid \mathbf{Y})$ and $p(\boldsymbol{\theta}' \mid \mathbf{Y})$ can be compared, without computing $p(\mathbf{Y})$

4. Obviously, the posterior $p(\boldsymbol{\theta} \mid \mathbf{Y})$ depends on the prior $\pi(\boldsymbol{\theta})$ (see next slides).

The posterior depends on the prior

Data & Model:

- $Y_i = 1$ if sick, 0 otherwise
- $n = 10$ patients
- $\vdots$: number sicks $/n$

The posterior depends on the prior

Data & Model:

- $Y_i = 1$ if sick, 0 otherwise
- $n = 10$ patients
- ⋮ : number sicks $/n$

Param:

- $\theta =$ proba. sick
- — : prior $\pi(\theta)$

The posterior depends on the prior

Data & Model:

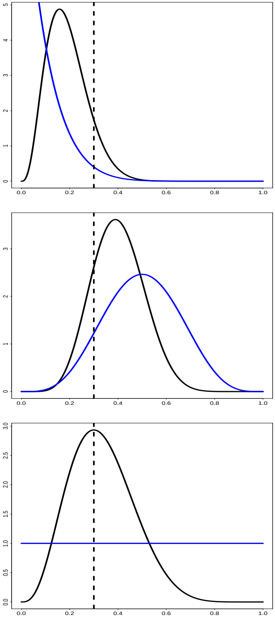
- $Y_i = 1$ if sick, 0 otherwise
- $n = 10$ patients
- $\vdots$: number sicks $/n$

Param:

- $\theta =$ proba. sick
- — : prior $\pi(\theta)$

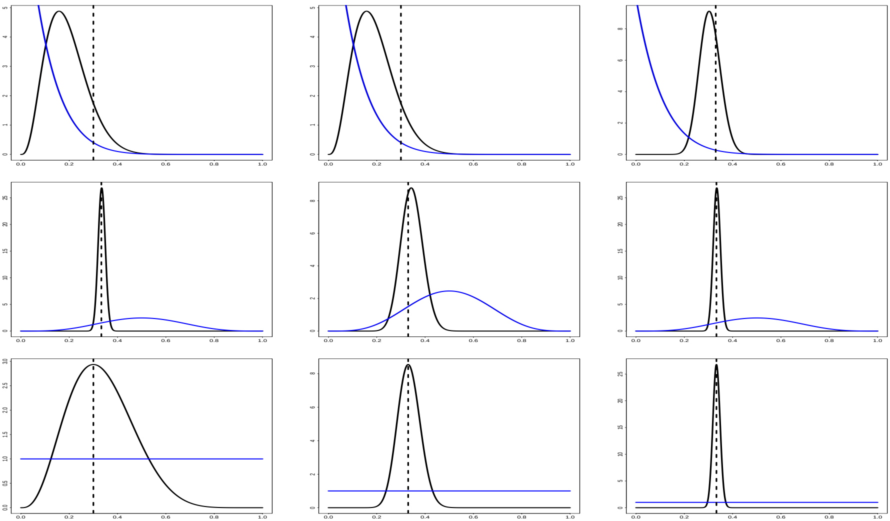
Output:

- — : posterior $p(\theta \mid \mathbf{Y})$



Dependency vanishes when n increases

$n = 10$ $n = 100$ $n = 1000$



Back to logistic regression

Model

- Prior: all coefficient θ_j independent:

$$\theta_j \sim \mathcal{N}(0, 100)$$

- Likelihood: all patients independent, *conditionally* on $\boldsymbol{\theta}$:

$$\Pr\{Y_i = 1 \,|\, \boldsymbol{\theta}\} = e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}} \Big/ \left(1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}\right)$$

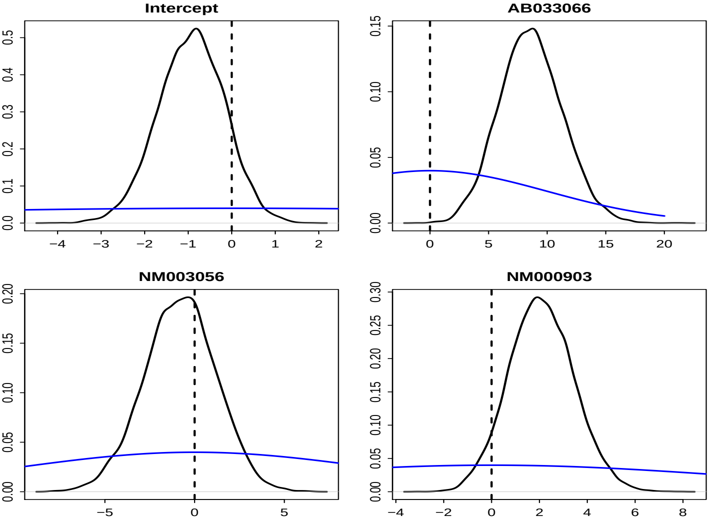
Inference:

$$\boldsymbol{\theta} \,|\, \mathbf{Y} \sim ?$$

(see later. For sure: $p(\boldsymbol{\theta} \,|\, \mathbf{Y}) \neq \mathcal{N}(\cdot, \cdot)$).

Bayesian inference

Output:



No test (and no estimator)

Frequentist hypothesis:

$$H_0 = \{\theta = 0\}$$

→ meaningless when θ is random: $P(H_0 \mid \mathbf{Y}) = 0$

No test (and no estimator)

Frequentist hypothesis:

$$H_0 = \{\theta = 0\}$$

→ meaningless when θ is random: $P(H_0 \,|\, \mathbf{Y}) = 0$

Bayesian assessment:

$$CI_{1-\alpha}(\theta \,|\, \mathbf{Y}) \ni 0 \text{ ?}$$

No test (and no estimator)

Frequentist hypothesis:

$$H_0 = \{\theta = 0\}$$

$\rightarrow$ meaningless when θ is random: $P(H_0 \,|\, \mathbf{Y}) = 0$

Bayesian assessment:

$$CI_{1-\alpha}(\theta \,|\, \mathbf{Y}) \ni 0 \text{ ?}$$

Parameter estimate. For the same reason:

$$\widehat{\theta} \text{ can no be an estimate of } \theta$$

(because θ is random).

Outline

Statistical inference: Bayesian point-of-view

Statistical inference: frequentist / Bayesian

Basics of Bayes inference

Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

Conjugate priors

Monte Carlo integration

Monte Carlo Markov chains (MCMC)

Extensions

Sequential Monte-Carlo (SMC)

Approximate Bayesian computation (ABC)

Posterior distribution and confidence intervals

Parameter 'estimate'.

$$\text{posterior mean:} \qquad \widehat{\theta}_j = \mathbb{E}(\theta_j \mid \mathbf{Y})$$

$$\text{posterior mode:} \qquad \widehat{\theta}_j = \arg\max_{\theta_j} \; p(\theta_j \mid \mathbf{Y})$$

Credibility interval (CI). With level $1 - \alpha$ (e.g. 95%):

$$CI_{1-\alpha}(\theta_j \mid \mathbf{Y}) = [\theta_j^\ell; \theta_j^u] : \qquad \Pr\{\theta_j^\ell < \theta_j < \theta_j^u \mid \mathbf{Y}\} = 1 - \alpha$$

Posterior distribution and confidence intervals

Parameter 'estimate'.

$$\text{posterior mean:} \qquad \widehat{\theta}_j = \mathbb{E}(\theta_j \mid \mathbf{Y})$$

$$\text{posterior mode:} \qquad \widehat{\theta}_j = \arg\max_{\theta_j} \; p(\theta_j \mid \mathbf{Y})$$

Credibility interval (CI).

With level $1 - \alpha$ (e.g. 95%):

$$CI_{1-\alpha}(\theta_j \mid \mathbf{Y}) = [\theta_j^\ell ; \theta_j^u] : \qquad \Pr\{\theta_j^\ell < \theta_j < \theta_j^u \mid \mathbf{Y}\} = 1 - \alpha$$

Example. [# 4]

	post.mean	post.mode	lower.CI	upper.CI
Intercept	-0.9298079	-0.8838218	-2.457669	0.5376564
AB033066	8.656539	8.497985	3.646142	13.98029
NM003056	-0.8669479	-0.5323168	-4.919099	3.084982
NM000903	2.088584	1.852784	-0.4736828	4.838164

Accounting for uncertainty

Question: What is the probability for patient 0 (with profile $\mathbf{x}_0$) to be sick?

Model answer:

$$\Pr\{Y_0 = 1 \,|\, \boldsymbol{\theta}\} = e^{\mathbf{x}_0^\intercal \boldsymbol{\theta}} \Big/ \Big(1 + e^{\mathbf{x}_0^\intercal \boldsymbol{\theta}}\Big)$$

but $\boldsymbol{\theta}$ is unknown (and random).

Bayesian answer: *posterior predictive* probability

$$\Pr\{Y_0 = 1 \,|\, \mathbf{Y}\} = \int \Pr\{Y_0 = 1 \,|\, \boldsymbol{\theta}\} p(\boldsymbol{\theta} \,|\, \mathbf{Y}) \, \mathrm{d}\boldsymbol{\theta}$$

Model comparison (1/2)

Problem. Which model fits the data better:

$$\begin{aligned}
M_0 &: \text{none of the genes has an effect, i.e. } \boldsymbol{\theta} = (\theta_0, 0, \dots, 0) \\
M_1 &: \text{only the fist gene has an effect, i.e. } \boldsymbol{\theta} = (\theta_0, \theta_1, 0, \dots, 0) \\
&\dots \\
M_p &: \text{all genes have an effect, i.e. } \boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_p)
\end{aligned}$$

Bayesian model comparison. For each model $M \in \mathcal{M} = \{M_0, \dots, M_p\}$, evaluate

$$p(M \mid \mathbf{Y})$$

Model comparison (2/2)

Ingredients:

- Prior on the models: $p(M)$, e.g.

$$p(M) = \text{cst} \qquad \text{(uniform prior)}$$

- Conditional prior on the parameters: $\pi(\boldsymbol{\theta} \mid M)$, e.g.

$$\theta_j \mid M_k \left\{ \begin{array}{lll} \sim & \mathcal{N}(0, 100) & \text{if } j \leq k \\ = & 0 & \text{otherwise} \end{array} \right.$$

Model comparison (2/2)

Ingredients:

- Prior on the models: $p(M)$, e.g.

$$p(M) = \text{cst} \qquad \text{(uniform prior)}$$

- Conditional prior on the parameters: $\pi(\boldsymbol{\theta} \mid M)$, e.g.

$$\theta_j \mid M_k \left\{ \begin{array}{lll} \sim & \mathcal{N}(0, 100) & \text{if } j \leq k \\ = & 0 & \text{otherwise} \end{array} \right.$$

Recipe:

- Evaluate the marginal likelihood of the data for each model M:

$$p(\mathbf{Y} \mid M) = \int \ell(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid M) \, \mathrm{d}\boldsymbol{\theta}$$

Model comparison (2/2)

Ingredients:

- Prior on the models: $p(M)$, e.g.

$$p(M) = \text{cst} \qquad \text{(uniform prior)}$$

- Conditional prior on the parameters: $\pi(\boldsymbol{\theta} \mid M)$, e.g.

$$\theta_j \mid M_k \left\{ \begin{array}{lll} \sim & \mathcal{N}(0, 100) & \text{if } j \leq k \\ = & 0 & \text{otherwise} \end{array} \right.$$

Recipe:

- Evaluate the marginal likelihood of the data for each model M:

$$p(\mathbf{Y} \mid M) = \int \ell(\mathbf{Y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid M) \, \mathrm{d}\boldsymbol{\theta}$$

- Evaluate the $p(M_k \mid \mathbf{Y})$ using Bayes rule

$$p(M_k \mid \mathbf{Y}) = \frac{p(M_k) p(\mathbf{Y} \mid M_k)}{p(\mathbf{Y})} = \frac{p(M_k) p(\mathbf{Y} \mid M_k)}{\sum_{k'} p(M_{k'}) p(\mathbf{Y} \mid M_{k'})}$$

Model averaging (uncertainty on models)

Question: Probability for patient 0 to be sick?

Model averaging (uncertainty on models)

Question: Probability for patient 0 to be sick?

Model selection.

- Select the 'best' model $\widehat{M}$, i.e. with largest posterior $p(M \mid \mathbf{Y})$
- Compute

$$\Pr\{Y_0 = 1 \mid \mathbf{Y}, \widehat{M}\} = \int \Pr\{Y_0 = 1 \mid \boldsymbol{\theta}\} p(\boldsymbol{\theta} \mid \mathbf{Y}, \widehat{M}) \, \mathrm{d}\boldsymbol{\theta}$$

Model averaging (uncertainty on models)

Question: Probability for patient 0 to be sick?

Model selection.

- Select the 'best' model $\widehat{M}$, i.e. with largest posterior $p(M \mid \mathbf{Y})$
- Compute

$$\Pr\{Y_0 = 1 \mid \mathbf{Y}, \widehat{M}\} = \int \Pr\{Y_0 = 1 \mid \boldsymbol{\theta}\} p(\boldsymbol{\theta} \mid \mathbf{Y}, \widehat{M}) \, \mathrm{d}\boldsymbol{\theta}$$

Model averaging.

- Keep all models
- Compute

$$\Pr\{Y_0 = 1 \mid \mathbf{Y}\} = \sum_M \Pr\{Y_0 = 1 \mid \mathbf{Y}, M\} p(M \mid \mathbf{Y})$$

Model averaging: Illustration

Aim: Probability p_0 to be sick a patient with gene expression profile

$$\mathbf{x}_0 = (0.178, 0.116, \dots, 0.076, -0.231)$$

Results for models $M_1, \dots, M_d$:

Model	$p(M \mid \mathbf{Y})$	$\mathbb{E}(p_0 \mid \mathbf{Y}, M)$	$\sqrt{\mathbb{V}}(p_0 \mid \mathbf{Y}, M)$
M_1	1e-04	0.494	0.028
M_2	7e-04	0.611	0.097
M_3	5e-04	0.627	0.106
...			
M_{14}	0.1436	0.242	0.18
M_{15}	0.2859	0.203	0.168
M_{16}	0.2726	0.195	0.168

Averaging		$\mathbb{E}(p_0 \mid \mathbf{Y})$	$\sqrt{\mathbb{V}}(p_0 \mid \mathbf{Y})$
		0.249	0.198

Transfer of uncertainty from one experience to another

Combining samples. Consider two independent but similar datasets $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

Simple algebra gives:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}_1, \mathbf{Y}_2) = \frac{p(\boldsymbol{\theta} \mid \mathbf{Y}_1) p(\mathbf{Y}_2 \mid \boldsymbol{\theta}, \mathbf{Y}_1)}{p(\mathbf{Y}_2 \mid \mathbf{Y}_1)}$$

Transfer of uncertainty from one experience to another

Combining samples. Consider two independent but similar datasets $\mathbf{Y}_1$ and $\mathbf{Y}_2$.

Simple algebra gives:

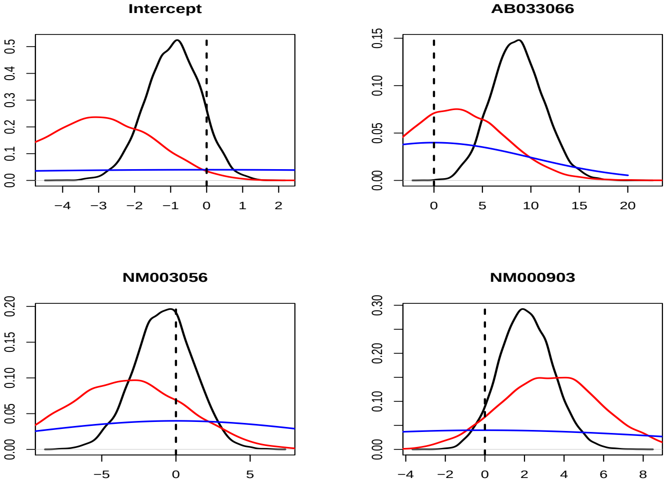
$$p(\boldsymbol{\theta} \mid \mathbf{Y}_1, \mathbf{Y}_2) = \frac{p(\boldsymbol{\theta} \mid \mathbf{Y}_1) p(\mathbf{Y}_2 \mid \boldsymbol{\theta}, \mathbf{Y}_1)}{p(\mathbf{Y}_2 \mid \mathbf{Y}_1)}$$

In practice:

1. Perform inference using $\mathbf{Y}_1$ to get $p(\boldsymbol{\theta} \mid \mathbf{Y}_1)$ from prior $\pi(\boldsymbol{\theta})$
2. Then perform inference using $\mathbf{Y}_2$ to get $p(\boldsymbol{\theta} \mid \mathbf{Y}_1, \mathbf{Y}_2)$ using $p(\boldsymbol{\theta} \mid \mathbf{Y}_1)$ as a prior

Combining experiments

Output: $n_1 = n_2 = 39$



Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

Posterior distribution

Aim: Evaluate

$$E[f(\boldsymbol{\theta})|\mathbf{Y}]$$

- Posterior mean: $f(\boldsymbol{\theta}) = \theta_j$
- Credibility interval: $f(\boldsymbol{\theta}) = \mathbb{I}\{\theta_j^\ell < \theta_j < \theta_j^u\}$
- Posterior variance: $f(\boldsymbol{\theta}) = \theta_j^2 \qquad (+$ posterior mean)

Posterior distribution

Aim: Evaluate

$$E[f(\boldsymbol{\theta})|\mathbf{Y}]$$

- Posterior mean: $f(\boldsymbol{\theta}) = \theta_j$
- Credibility interval: $f(\boldsymbol{\theta}) = \mathbb{I}\{\theta_j^\ell < \theta_j < \theta_j^u\}$
- Posterior variance: $f(\boldsymbol{\theta}) = \theta_j^2$ $\qquad$ $(+$ posterior mean)

Main problem: evaluate

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{\pi(\boldsymbol{\theta})\ell(\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y})}$$

which requires to evaluate

$$p(\mathbf{Y}) = \int \underbrace{\pi(\boldsymbol{\theta})}_{\textit{prior}} \underbrace{\ell(\mathbf{Y} \mid \boldsymbol{\theta})}_{\textit{likelihood}} \, \mathrm{d}\boldsymbol{\theta}$$

Outline

Statistical inference: Bayesian point-of-view

Statistical inference: frequentist / Bayesian

Basics of Bayes inference

Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

Conjugate priors

Monte Carlo integration

Monte Carlo Markov chains (MCMC)

Extensions

Sequential Monte-Carlo (SMC)

Approximate Bayesian computation (ABC)

Nice case: Conjugate priors

Example: Bernoulli[1]

Prior: $\theta =$ probability to be sick.

$$\theta \sim \text{Beta}(a, b), \qquad \pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

[1] #13: from top to bottom, $(a, b) = (1, 10), (5, 5), (1, 1)$

Nice case: Conjugate priors

Example: Bernoulli[1]

Prior: $\theta =$ probability to be sick.

$$\theta \sim \text{Beta}(a, b), \qquad \pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Likelihood: $Y_i = 1$ if sick, 0 otherwise. $S =$ number of sick

$$Y_i \mid \theta \sim \mathcal{B}(\theta), \qquad \ell(\mathbf{Y} \mid \theta) = \prod_i \theta^{Y_i}(1-\theta)^{1-Y_i} = \theta^S(1-\theta)^{n-S}$$

[1] #13: from top to bottom, $(a, b) = (1, 10), (5, 5), (1, 1)$

Nice case: Conjugate priors

Example: Bernoulli[1]

Prior: $\theta =$ probability to be sick.

$$\theta \sim \text{Beta}(a, b), \qquad \pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Likelihood: $Y_i = 1$ if sick, 0 otherwise. $S =$ number of sick

$$Y_i \,|\, \theta \sim \mathcal{B}(\theta), \qquad \ell(\mathbf{Y} \,|\, \theta) = \prod_i \theta^{Y_i}(1-\theta)^{1-Y_i} = \theta^S(1-\theta)^{n-S}$$

Posterior:

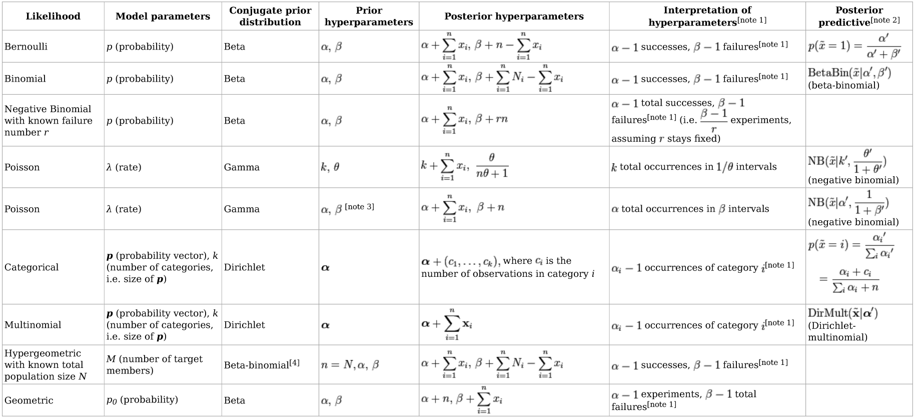
$$p(\theta \,|\, \mathbf{Y}) \propto \pi(\theta)\ell(\mathbf{Y} \,|\, \theta) = \theta^{a+S-1}(1-\theta)^{b+n-S-1}$$

which means that

$$\theta \,|\, \mathbf{Y} \sim \text{Beta}(a+S, b+n-S)$$

[1] #13: from top to bottom, $(a, b) = (1, 10), (5, 5), (1, 1)$

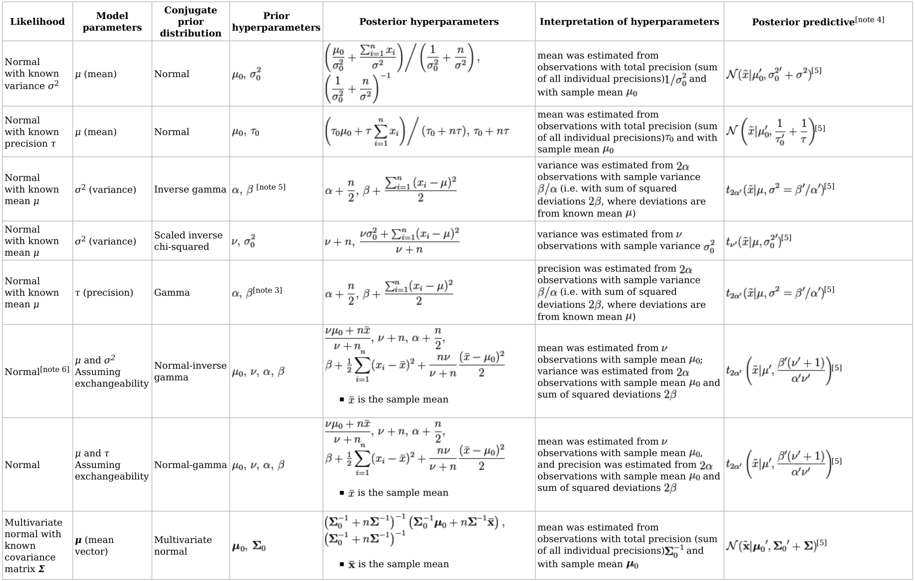
Conjugate priors: Discrete distributions



Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+n-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$p(\tilde{x}=1)=\frac{\alpha'}{\alpha'+\beta'}$
Binomial	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+\sum_{i=1}^n N_i-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x}\mid\alpha',\beta')$ (beta-binomial)
Negative Binomial with known failure number r	p (probability)	Beta	$\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+rn$	$\alpha-1$ total successes, $\beta-1$ failures[note 1] (i.e. $\frac{\beta-1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k+\sum_{i=1}^n x_i,\ \frac{\theta}{n\theta+1}$	k total occurrences in $1/\theta$ intervals	$\operatorname{NB}(\tilde{x}\mid k',\frac{\theta'}{1+\theta'})$ (negative binomial)
Poisson	λ (rate)	Gamma	$\alpha,\ \beta$ [note 3]	$\alpha+\sum_{i=1}^n x_i,\ \beta+n$	α total occurrences in β intervals	$\operatorname{NB}(\tilde{x}\mid\alpha',\frac{1}{1+\beta'})$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories, i.e. size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}+(c_1,\ldots,c_k)$, where c_i is the number of observations in category i	α_i-1 occurrences of category i[note 1]	$p(\tilde{x}=i)=\frac{{\alpha_i}'}{\sum_i {\alpha_i}'}$ $=\frac{\alpha_i+c_i}{\sum_i \alpha_i+n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories, i.e. size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}+\sum_{i=1}^n \mathbf{x}_i$	α_i-1 occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}}\mid\boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size N	M (number of target members)	Beta-binomial[4]	$n=N,\alpha,\ \beta$	$\alpha+\sum_{i=1}^n x_i,\ \beta+\sum_{i=1}^n N_i-\sum_{i=1}^n x_i$	$\alpha-1$ successes, $\beta-1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\ \beta$	$\alpha+n,\ \beta+\sum_{i=1}^n x_i$	$\alpha-1$ experiments, $\beta-1$ total failures[note 1]	

en.wikipedia.org/wiki/Conjugate_prior

Conjugate priors: Continuous distributions



Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\, \sigma_0^2$	$\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right)\Big/\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right),$ $\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions)$1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\|\mu_0', \sigma_0^{2'}+\sigma^2)$[5]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\, \tau_0$	$\left(\tau_0\mu_0+\tau\sum_{i=1}^n x_i\right)\Big/(\tau_0+n\tau),\, \tau_0+n\tau$	mean was estimated from observations with total precision (sum of all individual precisions)τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x}\|\mu_0', \frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[5]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\, \beta$ [note 5]	$\alpha+\frac{n}{2},\, \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\|\mu, \sigma^2=\beta'/\alpha')$[5]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\, \sigma_0^2$	$\nu+n,\, \frac{\nu\sigma_0^2+\sum_{i=1}^n(x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\|\mu, \sigma_0^{2'})$[5]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\, \beta$[note 3]	$\alpha+\frac{n}{2},\, \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\|\mu, \sigma^2=\beta'/\alpha')$[5]
Normal[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0,\, \nu,\, \alpha,\, \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\, \nu+n,\, \alpha+\frac{n}{2},$ $\beta+\tfrac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ ▪ $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\|\mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[5]
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0,\, \nu,\, \alpha,\, \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\, \nu+n,\, \alpha+\frac{n}{2},$ $\beta+\tfrac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ ▪ $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0, and precision was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\|\mu', \frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[5]
Multivariate normal with known covariance matrix $\boldsymbol{\Sigma}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0,\, \boldsymbol{\Sigma}_0$	$\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+n\boldsymbol{\Sigma}^{-1}\mathbf{\bar{x}}\right),$ $\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}$ ▪ $\mathbf{\bar{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions)$\boldsymbol{\Sigma}_0^{-1}$ and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}(\tilde{\mathbf{x}}\|\boldsymbol{\mu}_0{}', \boldsymbol{\Sigma}_0{}'+\boldsymbol{\Sigma})$[5]

en.wikipedia.org/wiki/Conjugate_prior

Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

Computing integrals

General case: $p(\boldsymbol{\theta} \mid \mathbf{Y})$ has no close form

Goal: compute

$$\mathbb{E}(f(\boldsymbol{\theta}) \mid \mathbf{Y}) = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \, \mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, \ell(\mathbf{Y} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \Big/ p(\mathbf{Y})$$

where

$$p(\mathbf{Y}) = \int \pi(\boldsymbol{\theta}) \, \ell(\mathbf{Y} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

Computing integrals

General case: $p(\boldsymbol{\theta} \,|\, \mathbf{Y})$ has no close form

Goal: compute

$$\mathbb{E}(f(\boldsymbol{\theta}) \,|\, \mathbf{Y}) = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathbf{Y}) \, \mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, \ell(\mathbf{Y} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \Big/ p(\mathbf{Y})$$

where

$$p(\mathbf{Y}) = \int \pi(\boldsymbol{\theta}) \, \ell(\mathbf{Y} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

We need to evaluate integrals of the form

$$\int [\cdots] \, \pi(\boldsymbol{\theta}) \, \ell(\mathbf{Y} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

Monte Carlo

Principle. To evaluate

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

Monte Carlo

Principle. To evaluate

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

1. sample

$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q$$

Monte Carlo

Principle. To evaluate

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

1. sample

$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q$$

2. compute

$$\widehat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{B} \sum_b f(\boldsymbol{\theta}^b)$$

Monte Carlo

Principle. To evaluate

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})q(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$$

1. sample

$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q$$

2. compute

$$\widehat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{B}\sum_b f(\boldsymbol{\theta}^b)$$

$\rightarrow$ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/B$. [# 64]

Monte Carlo

Principle. To evaluate

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

1. sample

$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q$$

2. compute

$$\widehat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{B} \sum_b f(\boldsymbol{\theta}^b)$$

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/B$. [# 64]

In practice:

- Works fine to evaluate $\mathbb{E}[f(\boldsymbol{\theta})]$, taking $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$

$$\widehat{\mathbb{E}}_{\mathcal{N}(0,10)}\left[e^{\theta}\right] = \texttt{mean(exp(rnorm(B, mean=0, sd=sqrt(10))))}$$

Monte Carlo

Principle. To evaluate

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

1. sample

$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q$$

2. compute

$$\widehat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{B} \sum_b f(\boldsymbol{\theta}^b)$$

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/B$. [# 64]

In practice:

- Works fine to evaluate $\mathbb{E}[f(\boldsymbol{\theta})]$, taking $q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$

$$\widehat{\mathbb{E}}_{\mathcal{N}(0,10)}\left[e^{\theta}\right] = \texttt{mean(exp(rnorm(B, mean=0, sd=sqrt(10))))}$$

- Useless to evaluate $\mathbb{E}[f(\boldsymbol{\theta})|\mathbf{Y}]$ as we do not know how to sample from $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \,|\, \mathbf{Y})$

Importance Sampling (IS)

Main trick = weighting particles.

Importance Sampling (IS)

Main trick = weighting particles. To evaluate $\mathbb{E}[f(\boldsymbol{\theta})]$, write it as

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})q(\boldsymbol{\theta}) \; \mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{\theta})\frac{q(\boldsymbol{\theta})}{q'(\boldsymbol{\theta})}q'(\boldsymbol{\theta}) \; \mathrm{d}\boldsymbol{\theta}$$

for some proposal $q' \gg q$, from which you *know how to sample*, then

Importance Sampling (IS)

Main trick = weighting particles. To evaluate $\mathbb{E}[f(\boldsymbol{\theta})]$, write it as

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \;\mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \frac{q(\boldsymbol{\theta})}{q'(\boldsymbol{\theta})} q'(\boldsymbol{\theta}) \;\mathrm{d}\boldsymbol{\theta}$$

for some proposal $q' \gg q$, from which you *know how to sample*, then

1. sample

$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q'(\boldsymbol{\theta}),$$

Importance Sampling (IS)

Main trick = weighting particles. To evaluate $\mathbb{E}[f(\boldsymbol{\theta})]$, write it as

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta})q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \frac{q(\boldsymbol{\theta})}{q'(\boldsymbol{\theta})} q'(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

for some proposal $q' \gg q$, from which you *know how to sample*, then

1. sample
$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q'(\boldsymbol{\theta}),$$
2. compute the weights
$$W(\boldsymbol{\theta}^b) = q(\boldsymbol{\theta}^b)/q'(\boldsymbol{\theta}^b),$$

Importance Sampling (IS)

Main trick = weighting particles. To evaluate $\mathbb{E}[f(\boldsymbol{\theta})]$, write it as

$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \; \mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \frac{q(\boldsymbol{\theta})}{q'(\boldsymbol{\theta})} q'(\boldsymbol{\theta}) \; \mathrm{d}\boldsymbol{\theta}$$

for some proposal $q' \gg q$, from which you *know how to sample*, then

1. sample
$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q'(\boldsymbol{\theta}),$$
2. compute the weights
$$W(\boldsymbol{\theta}^b) = q(\boldsymbol{\theta}^b) / q'(\boldsymbol{\theta}^b),$$
3. and compute
$$\widehat{\mathbb{E}}[f(\boldsymbol{\theta})] = \frac{1}{B} \sum_b W(\boldsymbol{\theta}^b) f(\boldsymbol{\theta}^b)$$

Importance Sampling (IS)

Main trick = weighting particles. To evaluate $\mathbb{E}[f(\boldsymbol{\theta})]$, write it as

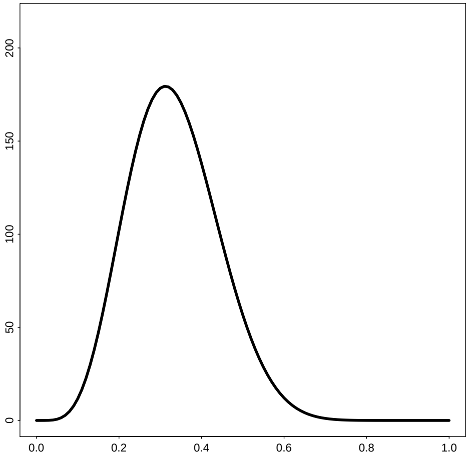
$$\mathbb{E}_q[f(\boldsymbol{\theta})] = \int f(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} = \int f(\boldsymbol{\theta}) \frac{q(\boldsymbol{\theta})}{q'(\boldsymbol{\theta})} q'(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

for some proposal $q' \gg q$, from which you *know how to sample*, then

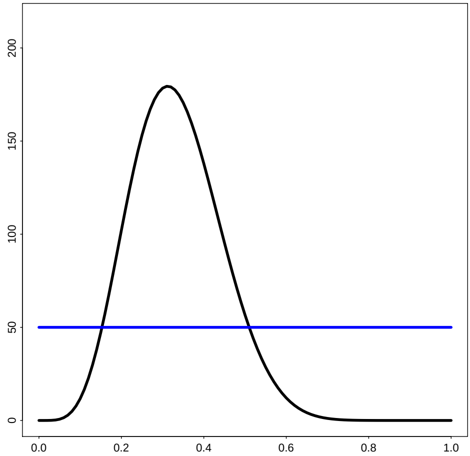
1. sample
$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q'(\boldsymbol{\theta}),$$
2. compute the weights
$$W(\boldsymbol{\theta}^b) = q(\boldsymbol{\theta}^b)/q'(\boldsymbol{\theta}^b),$$
3. and compute
$$\widehat{\mathbb{E}}[f(\boldsymbol{\theta})] = \frac{1}{B} \sum_b W(\boldsymbol{\theta}^b) f(\boldsymbol{\theta}^b)$$

→ unbiased estimate of $\mathbb{E}[f(\boldsymbol{\theta})]$ with variance $\propto \sum_b W(\theta^b)^2/B$.

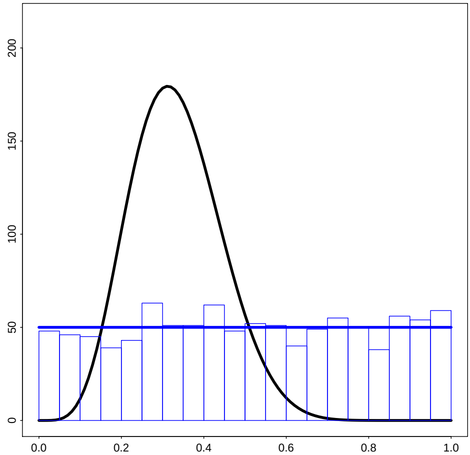
Importance Sampling (a picture)



Importance Sampling (a picture)



Importance Sampling (a picture)



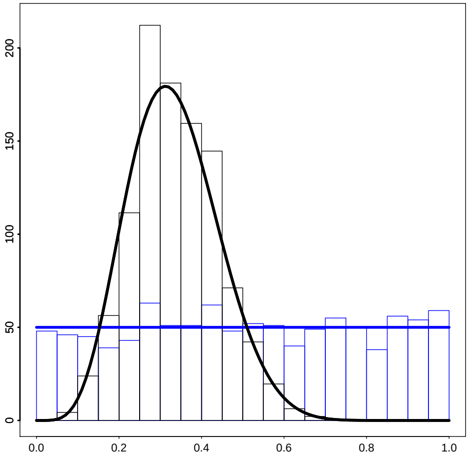
Importance Sampling (a picture)

Efficiency of sampling:

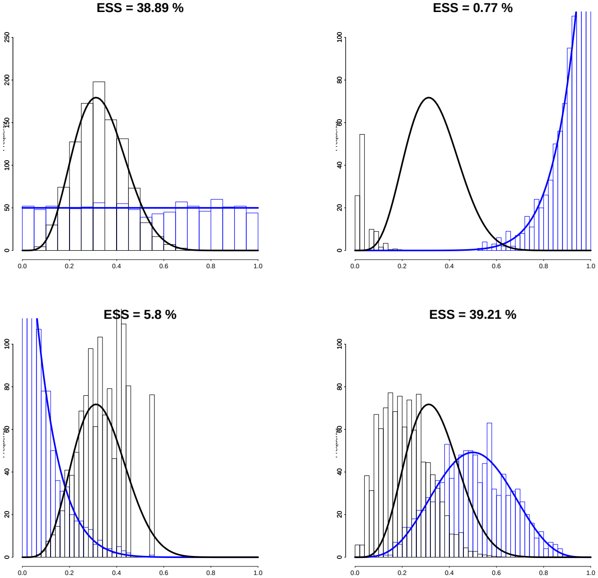
$$ESS = \overline{W}^2 / \overline{W^2}$$

$q' = q$

$$\Rightarrow \quad ESS = 1$$



Importance Sampling: Importance of the proposal

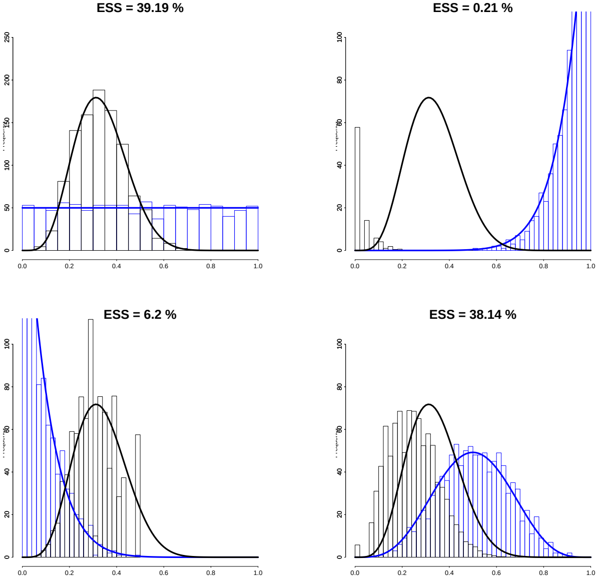

ESS = 38.89 %

ESS = 0.77 %

ESS = 5.8 %

ESS = 39.21 %

Importance of the proposal: another draw



IS for posterior sampling

To evaluate $\mathbb{E}[f(\boldsymbol{\theta})|\mathbf{Y}]$, write it as

$$\mathbb{E}[f(\boldsymbol{\theta}) \mid \mathbf{Y}] = \int f(\boldsymbol{\theta}) p(\boldsymbol{\theta}, \mathbf{Y}) \, \mathrm{d}\boldsymbol{\theta} \Big/ p(\mathbf{Y}) \; = \; \dots$$

$$= \int f(\boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \Big/ \int \frac{\pi(\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{Y})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$$

1. sample

$$(\boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^B) \text{ iid } \sim q$$

2. compute the weights

$$W(\boldsymbol{\theta}^b) = \pi(\boldsymbol{\theta}^b) p(\boldsymbol{\theta}^b \mid \mathbf{Y}) \Big/ q(\boldsymbol{\theta}^b)$$

3. get

$$\widehat{\mathbb{E}}[f(\boldsymbol{\theta}) \mid \mathbf{Y}] = \sum_b W(\boldsymbol{\theta}^b) f(\boldsymbol{\theta}^b) \Big/ \sum_b W(\boldsymbol{\theta}^b)$$

(slightly biased).

Good proposals

Choosing q is critical

Good proposals

Choosing q is critical

Typical choices

- Prior

$$q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$$

$\rightarrow$ far from the target $p(\boldsymbol{\theta} \mid \mathbf{Y})$: small *ESS*

Good proposals

Choosing q is critical

Typical choices

- Prior
$$q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$$
$\rightarrow$ far from the target $p(\boldsymbol{\theta} \mid \mathbf{Y})$: small *ESS*

- MLE:
$$q(\boldsymbol{\theta}) = \mathcal{N}(\widehat{\boldsymbol{\theta}}_{MLE}, \mathbb{V}_{\infty}(\widehat{\boldsymbol{\theta}}_{MLE}))$$
$\rightarrow$ fine, as long as MLE is available

Good proposals

Choosing q is critical

Typical choices

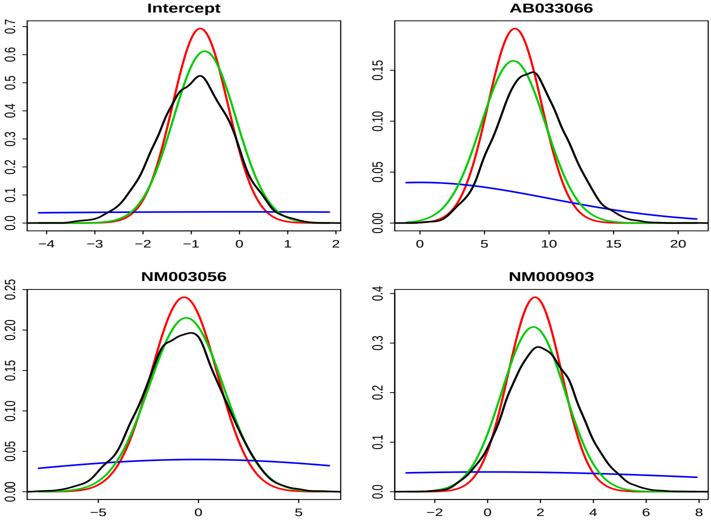
- Prior
$$q(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$$
→ far from the target $p(\boldsymbol{\theta} \mid \mathbf{Y})$: small *ESS*

- MLE:
$$q(\boldsymbol{\theta}) = \mathcal{N}(\widehat{\boldsymbol{\theta}}_{MLE}, \mathbb{V}_\infty(\widehat{\boldsymbol{\theta}}_{MLE}))$$
→ fine, as long as MLE is available

- Variational Bayes, expectation propagation, ...:
$$q(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} KL\left[q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta} \mid \mathbf{Y})\right]$$
→ fast and reasonably accurate

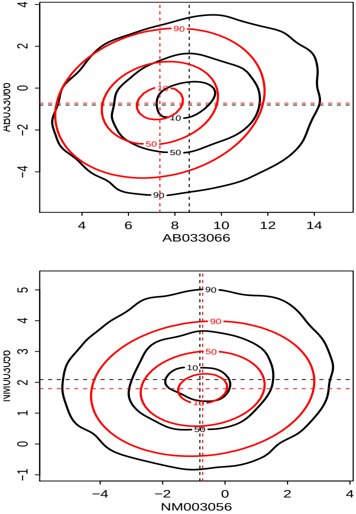
Variational Bayes & ML as a prior

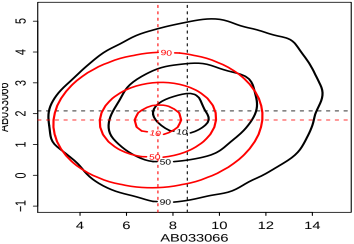
— : prior, — : VB, — : MLE, — : posterior



Variational Bayes as a prior: joint distribution

— : VB, — : posterior





Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

Limit distribution of Markov chain

Property. If $\{\phi^b\}_{b\geq 0}$ is an ergodic Markov chain (irreducible, aperiodic, ...) with

- initial distribution $\phi^0 \sim \nu$,
- transition kernel $\phi^b \mid \phi^{b-1} \sim \kappa(\cdot \mid \phi^{b-1})$:

$$p\left(\{\phi^b\}\right) = \nu(\phi^0) \times \kappa(\phi^1 \mid \phi^0) \times \kappa(\phi^2 \mid \phi^1) \times \kappa(\phi^3 \mid \phi^2) \times \dots$$

then

- it admits a unique stationary distribution μ:

$$\phi^{b-1} \sim \mu \qquad \Rightarrow \qquad \phi^b \sim \mu$$

- ϕ^b converges towards μ in distribution

$$\phi^b \xrightarrow[b\to\infty]{\Delta} \mu$$

for any initial distribution ν

Use for Bayesian inference

Aim. Sample from

$$p(\boldsymbol{\theta} \mid \mathbf{Y})$$

Idea.

- Construct an ergodic Markov chain $\{\boldsymbol{\theta}^b\}_{b \geq 0}$ with stationary distribution

$$\mu(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{Y})$$

- Choose 'any' initial ν and simulate $\{\boldsymbol{\theta}^b\}_{b \geq 0}$

- Until it 'reaches' its stationary distribution

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \mid \boldsymbol{\theta})$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \,|\, \boldsymbol{\theta})$

- Start with $\boldsymbol{\theta}^0$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \mid \boldsymbol{\theta})$

- Start with $\boldsymbol{\theta}^0$
- At step b,

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \,|\, \boldsymbol{\theta})$

- Start with $\boldsymbol{\theta}^0$
- At step b,
  1. sample $\boldsymbol{\theta}' \sim \lambda(\cdot \,|\, \boldsymbol{\theta}^{b-1})$;

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \,|\, \boldsymbol{\theta})$

- Start with $\boldsymbol{\theta}^0$
- At step b,
  1. sample $\boldsymbol{\theta}' \sim \lambda(\cdot \,|\, \boldsymbol{\theta}^{b-1})$;
  2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{b-1}) = \frac{\lambda(\boldsymbol{\theta}^{b-1} \,|\, \boldsymbol{\theta}')}{\lambda(\boldsymbol{\theta}' \,|\, \boldsymbol{\theta}^{b-1})} \frac{p(\boldsymbol{\theta}' \,|\, \mathbf{Y})}{p(\boldsymbol{\theta}^{b-1} \,|\, \mathbf{Y})}$$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \mid \boldsymbol{\theta})$

- Start with $\boldsymbol{\theta}^0$
- At step b,
  1. sample $\boldsymbol{\theta}' \sim \lambda(\cdot \mid \boldsymbol{\theta}^{b-1})$;
  2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{b-1}) = \frac{\lambda(\boldsymbol{\theta}^{b-1} \mid \boldsymbol{\theta}')}{\lambda(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{b-1})} \frac{p(\boldsymbol{\theta}' \mid \mathbf{Y})}{p(\boldsymbol{\theta}^{b-1} \mid \mathbf{Y})} = \frac{\lambda(\boldsymbol{\theta}^{b-1} \mid \boldsymbol{\theta}')}{\lambda(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{b-1})} \frac{\pi(\boldsymbol{\theta}')\ell(\mathbf{Y} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{b-1})\ell(\mathbf{Y} \mid \boldsymbol{\theta}^{b-1})};$$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \mid \boldsymbol{\theta})$

- Start with $\boldsymbol{\theta}^0$
- At step b,
  1. sample $\boldsymbol{\theta}' \sim \lambda(\cdot \mid \boldsymbol{\theta}^{b-1})$;
  2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{b-1}) = \frac{\lambda(\boldsymbol{\theta}^{b-1} \mid \boldsymbol{\theta}')}{\lambda(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{b-1})} \frac{p(\boldsymbol{\theta}' \mid \mathbf{Y})}{p(\boldsymbol{\theta}^{b-1} \mid \mathbf{Y})} = \frac{\lambda(\boldsymbol{\theta}^{b-1} \mid \boldsymbol{\theta}')}{\lambda(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{b-1})} \frac{\pi(\boldsymbol{\theta}')\ell(\mathbf{Y} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{b-1})\ell(\mathbf{Y} \mid \boldsymbol{\theta}^{b-1})};$$

  3. set $\boldsymbol{\theta}^b = \begin{cases} \boldsymbol{\theta}' & \text{with probability } \max(1, \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{b-1})), \\ \boldsymbol{\theta}^{b-1} & \text{otherwise.} \end{cases}$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot \mid \boldsymbol{\theta})$

- Start with $\boldsymbol{\theta}^0$
- At step b,
  1. sample $\boldsymbol{\theta}' \sim \lambda(\cdot \mid \boldsymbol{\theta}^{b-1})$;
  2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{b-1}) = \frac{\lambda(\boldsymbol{\theta}^{b-1} \mid \boldsymbol{\theta}')}{\lambda(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{b-1})} \frac{p(\boldsymbol{\theta}' \mid \mathbf{Y})}{p(\boldsymbol{\theta}^{b-1} \mid \mathbf{Y})} = \frac{\lambda(\boldsymbol{\theta}^{b-1} \mid \boldsymbol{\theta}')}{\lambda(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^{b-1})} \frac{\pi(\boldsymbol{\theta}')\ell(\mathbf{Y} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{b-1})\ell(\mathbf{Y} \mid \boldsymbol{\theta}^{b-1})};$$

  3. set $\boldsymbol{\theta}^b = \begin{cases} \boldsymbol{\theta}' & \text{with probability } \max(1, \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{b-1})), \\ \boldsymbol{\theta}^{b-1} & \text{otherwise.} \end{cases}$

Properties.

1. λ and α define a Markov chain with stationary distribution $\mu(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{Y})$.
2. If $\lambda(\cdot \mid \boldsymbol{\theta})$ is symmetric, α reduce to $\pi(\boldsymbol{\theta}')\ell(\mathbf{Y} \mid \boldsymbol{\theta}')/[\pi(\boldsymbol{\theta}^{b-1})\ell(\mathbf{Y} \mid \boldsymbol{\theta}^{b-1})]$

Metropolis-Hastings for logistic regression

Model.

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}_p, 100\,\mathbf{I}_p)$$

$$\mathbf{Y} \mid \boldsymbol{\theta} \sim \ell(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_i \left(\frac{e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}\right)^{y_i} \left(\frac{e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}\right)^{1-y_i}$$

Metropolis-Hastings for logistic regression

Model.

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}_p, 100\,\mathbf{I}_p)$$
$$\mathbf{Y} \mid \boldsymbol{\theta} \sim \ell(\mathbf{Y} \mid \boldsymbol{\theta}) = \prod_i \left(\frac{e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}\right)^{y_i} \left(\frac{e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i^\intercal \boldsymbol{\theta}}}\right)^{1-y_i}$$

Algorithm settings.

$$\boldsymbol{\theta}^0 = \mathbf{0}_p$$
$$\lambda(\cdot \mid \boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}_p, .5\,\mathbf{I}_p)$$

M-H for logistic regression: R code

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
    logprior.tmp = dmvnorm(theta.tmp, mean=mu.prior, sigma=Sigma.prior, log=T)
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
    logprior.tmp = dmvnorm(theta.tmp, mean=mu.prior, sigma=Sigma.prior, log=T)
    prob.tmp = plogis(X%*%theta.tmp)
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
    logprior.tmp = dmvnorm(theta.tmp, mean=mu.prior, sigma=Sigma.prior, log=T)
    prob.tmp = plogis(X%*%theta.tmp)
    loglik.tmp = sum(dbinom(Y, 1, prob.tmp, log=T))
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
    logprior.tmp = dmvnorm(theta.tmp, mean=mu.prior, sigma=Sigma.prior, log=T)
    prob.tmp = plogis(X%*%theta.tmp)
    loglik.tmp = sum(dbinom(Y, 1, prob.tmp, log=T))
    alpha = exp(logprior.tmp + loglik.tmp - logprior.cur - loglik.cur)
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
    logprior.tmp = dmvnorm(theta.tmp, mean=mu.prior, sigma=Sigma.prior, log=T)
    prob.tmp = plogis(X%*%theta.tmp)
    loglik.tmp = sum(dbinom(Y, 1, prob.tmp, log=T))
    alpha = exp(logprior.tmp + loglik.tmp - logprior.cur - loglik.cur)
    if(runif(1) < alpha){
        theta.sample[b, ] = theta.cur = theta.tmp
        logprior.cur = logprior.tmp
        loglik.cur = loglik.tmp
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
    logprior.tmp = dmvnorm(theta.tmp, mean=mu.prior, sigma=Sigma.prior, log=T)
    prob.tmp = plogis(X%*%theta.tmp)
    loglik.tmp = sum(dbinom(Y, 1, prob.tmp, log=T))
    alpha = exp(logprior.tmp + loglik.tmp - logprior.cur - loglik.cur)
    if(runif(1) < alpha){
        theta.sample[b, ] = theta.cur = theta.tmp
        logprior.cur = logprior.tmp
        loglik.cur = loglik.tmp
    }else{
        theta.sample[b, ] = theta.sample[b-1, ]
    }
```

M-H for logistic regression: R code

```
mu.prior = rep(0, p); Sigma.prior = 100*diag(p); Sigma.shift = .5*diag(p)
theta.sample = matrix(0, B, p); theta.cur = theta.sample[1, ]
logprior.cur = dmvnorm(theta.cur, mean=mu.prior, sigma=Sigma.prior, log=T)
prob.cur = plogis(X%*%theta.cur)
loglik.cur = sum(dbinom(Y, 1, prob.cur, log=T))
for (b in 2:B){
    theta.tmp = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
    logprior.tmp = dmvnorm(theta.tmp, mean=mu.prior, sigma=Sigma.prior, log=T)
    prob.tmp = plogis(X%*%theta.tmp)
    loglik.tmp = sum(dbinom(Y, 1, prob.tmp, log=T))
    alpha = exp(logprior.tmp + loglik.tmp - logprior.cur - loglik.cur)
    if(runif(1) < alpha){
        theta.sample[b, ] = theta.cur = theta.tmp
        logprior.cur = logprior.tmp
        loglik.cur = loglik.tmp
    }else{
        theta.sample[b, ] = theta.sample[b-1, ]
    }
}
```

Sanity checks

Setting. Sample 1.2 10^6 $\boldsymbol{\theta}$, remove first 2 10^5, extract every 10 $\rightarrow$ $B = 10^5$.

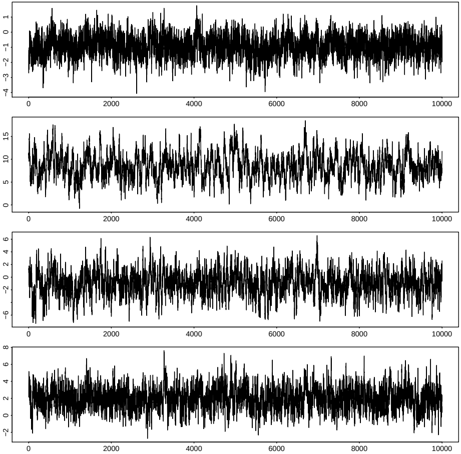
- Acceptance rate

variance shift	0.1	0.5	1
acceptance rate	0.035	0.011	0.004

Sanity checks

Setting. Sample 1.2 10^6 $\boldsymbol{\theta}$, remove first 2 10^5, extract every 10 $\rightarrow$ $B = 10^5$.

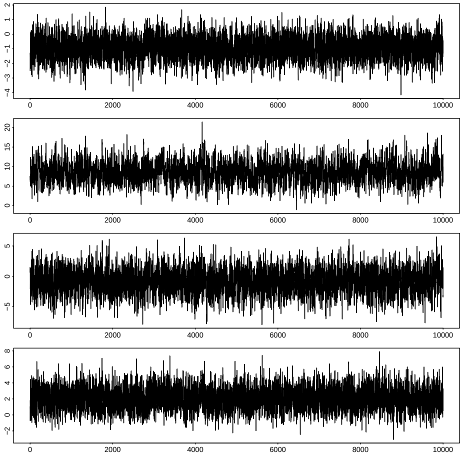
- Acceptance rate
- Stationarity:
  var. shift = .1



Sanity checks

Setting. Sample 1.2 10^6 $\boldsymbol{\theta}$, remove first 2 10^5, extract every 10 $\rightarrow$ $B = 10^5$.

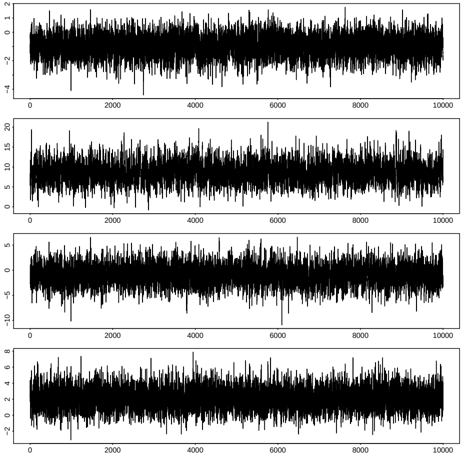
- Acceptance rate
- Stationarity:
  var. shift = .1 , .5



Sanity checks

Setting. Sample 1.2 10^6 $\boldsymbol{\theta}$, remove first 2 10^5, extract every 10 $\rightarrow$ $B = 10^5$.

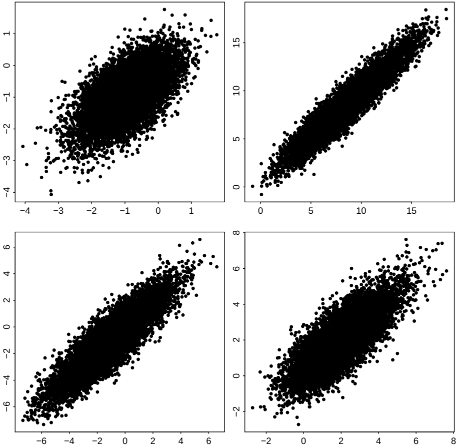
- Acceptance rate
- Stationarity:
  var. shift = .1 , .5 , 1



Sanity checks

Setting. Sample 1.2 10^6 $\boldsymbol{\theta}$, remove first 2 10^5, extract every 10 $\rightarrow$ $B = 10^5$.

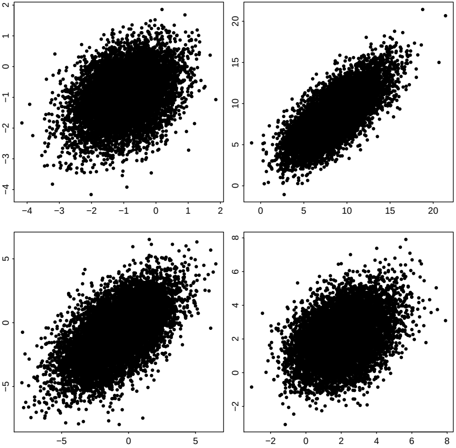
- Acceptance rate
- Stationarity:
  var. shift = .1 , .5 , 1
- Autocorrelation
  $\mathbb{C}\text{or}(\theta_j^{b-1}, \theta_j^b)$:
  var. shift = .1



Sanity checks

Setting. Sample 1.2 10^6 $\boldsymbol{\theta}$, remove first 2 10^5, extract every 10 $\rightarrow$ $B = 10^5$.

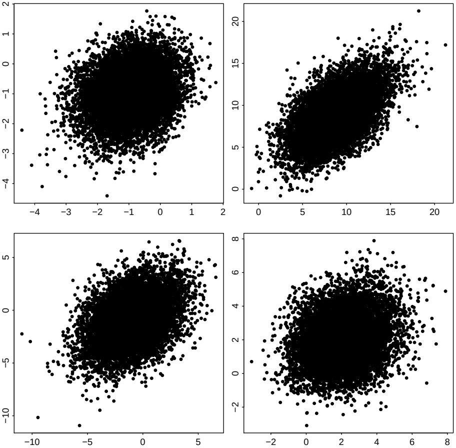
- Acceptance rate
- Stationarity:
  var. shift = .1 , .5 , 1
- Autocorrelation
  $\mathbb{C}\text{or}(\theta_j^{b-1}, \theta_j^b)$:
  var. shift = .1 , .5



Sanity checks

Setting. Sample 1.2 10^6 $\boldsymbol{\theta}$, remove first 2 10^5, extract every 10 $\rightarrow$ $B = 10^5$.

- Acceptance rate
- Stationarity:
  var. shift = .1 , .5 , 1
- Autocorrelation
  $\mathbb{C}\text{or}(\theta_j^{b-1}, \theta_j^b)$:
  var. shift = .1 , .5 , 1



Gibbs

Framework. We do not know how to sample the whole vector $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \,|\, \mathbf{Y})$$

but we may know how to sample each coordinate (conditional on the others):

$$p(\theta_j \,|\, \mathbf{Y}, \boldsymbol{\theta}_{-j})$$

$\boldsymbol{\theta}_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots \theta_p)$.

Gibbs

Framework. We do not know how to sample the whole vector $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \,|\, \mathbf{Y})$$

but we may know how to sample each coordinate (conditional on the others):

$$p(\theta_j \,|\, \mathbf{Y}, \boldsymbol{\theta}_{-j})$$

$\boldsymbol{\theta}_{-j} = (\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots \theta_p)$.

Sampling a genotype.

- Hard to sample a whole genotype (accounting for linkage disequilibrium)
- Easy to sample the genotype at one locus, conditional on the rest of the genotype

Gibbs sampling for Bayesian inference

Algorithm. Sample $\{\boldsymbol{\theta}^b\}_{b=0,\dots B}$ as follows.

Gibbs sampling for Bayesian inference

Algorithm. Sample $\{\boldsymbol{\theta}^b\}_{b=0,\dots B}$ as follows.

- Start with $\boldsymbol{\theta}^0$

Gibbs sampling for Bayesian inference

Algorithm. Sample $\{\boldsymbol{\theta}^b\}_{b=0,\dots B}$ as follows.

- Start with $\boldsymbol{\theta}^0$
- At step b, for $j = 1, \dots p$, sample θ_j^b:

$$\theta_j^b \mid \mathbf{Y}, \theta_1^b, \dots, \theta_{j-1}^b, \theta_{j+1}^{b-1}, \dots, \theta_{j+1}^{b-1}$$

Gibbs sampling for Bayesian inference

Algorithm. Sample $\{\boldsymbol{\theta}^b\}_{b=0,\dots B}$ as follows.

- Start with $\boldsymbol{\theta}^0$
- At step b, for $j = 1, \dots p$, sample θ_j^b:

$$\theta_j^b \mid \mathbf{Y}, \theta_1^b, \dots, \theta_{j-1}^b, \theta_{j+1}^{b-1}, \dots, \theta_{j+1}^{b-1}$$

Property.

- Obviously, $p(\boldsymbol{\theta} \mid \mathbf{Y})$ is a stationary distribution.
- Does not suffices to prove ergodicity.

Outline

Statistical inference: Bayesian point-of-view

Statistical inference: frequentist / Bayesian

Basics of Bayes inference

Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

Conjugate priors

Monte Carlo integration

Monte Carlo Markov chains (MCMC)

Extensions

Sequential Monte-Carlo (SMC)

Approximate Bayesian computation (ABC)

Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

Sequential Monte-Carlo

Example: Hidden Markov models

- $\mathbf{Z} = (Z_t)_{t \leq t}$ hidden Markov chain
- $\mathbf{Y} =$ observed sequence
- $\boldsymbol{\theta} = (\Pi, \gamma)$: transition matrix and emission probabilities

Sequential Monte-Carlo

Example: Hidden Markov models

- $\mathbf{Z} = (Z_t)_{t \leq t}$ hidden Markov chain
- $\mathbf{Y} =$ observed sequence
- $\boldsymbol{\theta} = (\Pi, \gamma)$: transition matrix and emission probabilities

Inference. Need to sample from

- $p(\boldsymbol{\theta} \,|\, \mathbf{Y})$ (parameter inference)
- $p(\mathbf{Z} \,|\, \mathbf{Y})$ (classification)

Sequential Monte-Carlo

Example: Hidden Markov models

- $\mathbf{Z} = (Z_t)_{t \leq t}$ hidden Markov chain
- $\mathbf{Y} =$ observed sequence
- $\boldsymbol{\theta} = (\Pi, \gamma)$: transition matrix and emission probabilities

Inference. Need to sample from

- $p(\boldsymbol{\theta} \mid \mathbf{Y})$ (parameter inference)
- $p(\mathbf{Z} \mid \mathbf{Y})$ (classification)

Sequential Monte Carlo.

- Monte Carlo (stochastic) counterpart of the forward-backward recurrence
- Sequentially sample from $p(Z_t \mid \mathbf{Y}_1^t, \mathbf{Z}_1^{t-1})$.

Outline

Statistical inference: Bayesian point-of-view

- Statistical inference: frequentist / Bayesian
- Basics of Bayes inference
- Some typical uses of Bayesian inference

Evaluating the posterior distribution: Monte-Carlo

- Conjugate priors
- Monte Carlo integration
- Monte Carlo Markov chains (MCMC)

Extensions

- Sequential Monte-Carlo (SMC)
- Approximate Bayesian computation (ABC)

When the likelihood is intractable

Ex.: Population genetics. Complex demographic model for which

- we do not know how to compute the likelihood:

$$\ell(\mathbf{Y} \mid \boldsymbol{\theta}) \text{ intractable}$$

- but we know how to sample from it

$$\mathbf{Y}^b \sim \ell(\mathbf{Y} \mid \boldsymbol{\theta}).$$

→ Importance sampling, Metropolis-Hastings, ... can not be implemented.

When the likelihood is intractable

Ex.: Population genetics. Complex demographic model for which

- we do not know how to compute the likelihood:

$$\ell(\mathbf{Y} \mid \boldsymbol{\theta}) \text{ intractable}$$

- but we know how to sample from it

$$\mathbf{Y}^b \sim \ell(\mathbf{Y} \mid \boldsymbol{\theta}).$$

→ Importance sampling, Metropolis-Hastings, ... can not be implemented.

Principle. Get a sample $\{\theta^b\}$ such that

$$\mathbf{Y}^b \sim p(\mathbf{Y} \mid \boldsymbol{\theta}^b) \text{ is 'similar' to } \mathbf{Y}_{\text{obs}}$$

Approximate Bayesian computation (ABC)

Ingredients.

- A set a *summary statstics* $\mathbf{s}(\mathbf{Y})$
- A 'distance' $d(\mathbf{s}, \mathbf{s}')$
- A threshold ε

Approximate Bayesian computation (ABC)

Ingredients.

- A set a *summary statstics* $\mathbf{s}(\mathbf{Y})$
- A 'distance' $d(\mathbf{s}, \mathbf{s}')$
- A threshold ε

Algorithm.

- Compute $\mathbf{s}_{\text{obs}} = \mathbf{s}(\mathbf{Y}_{\text{obs}})$
- Until we get B realizations
  1. sample $\boldsymbol{\theta}' \sim \pi(\boldsymbol{\theta})$ (from the prior)
  2. sample $\mathbf{Y}' \sim \ell(\mathbf{Y} \mid \boldsymbol{\theta}')$ (from the model)
  3. compute $\mathbf{s}' = \mathbf{s}(\mathbf{Y}')$
  4. if $d(\mathbf{s}' - \mathbf{s}_{\text{obs}}) < \varepsilon$, keep $\boldsymbol{\theta}'$ in the sample

Approximate Bayesian computation (ABC)

Ingredients.

- A set a *summary statstics* $\mathbf{s}(\mathbf{Y})$
- A 'distance' $d(\mathbf{s}, \mathbf{s}')$
- A threshold ε

Algorithm.

- Compute $\mathbf{s}_{\text{obs}} = \mathbf{s}(\mathbf{Y}_{\text{obs}})$
- Until we get B realizations
    1. sample $\boldsymbol{\theta}' \sim \pi(\boldsymbol{\theta})$ (from the prior)
    2. sample $\mathbf{Y}' \sim \ell(\mathbf{Y} \mid \boldsymbol{\theta}')$ (from the model)
    3. compute $\mathbf{s}' = \mathbf{s}(\mathbf{Y}')$
    4. if $d(\mathbf{s}' - \mathbf{s}_{\text{obs}}) < \varepsilon$, keep $\boldsymbol{\theta}'$ in the sample

Rational. Do not sample from $p(\boldsymbol{\theta} \mid \mathbf{Y})$ but from

$$p(\boldsymbol{\theta} \mid d(\mathbf{s}(\mathbf{Y}) - \mathbf{s}(\mathbf{Y}_{\text{obs}})) < \varepsilon).$$

References

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

J.-M. Marin and Ch. P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag: New-York, 2007.

Outline

Appendix

Joint, Marginal, Conditional (1/2)

Reminder: 2 loci with 2 alleles each: (A, a), (B, b)

- Joint distribution:

	B	b	marginal
A	f_{AB}	f_{Ab}	$p_A = f_{AB} + f_{Ab}$
a	f_{aB}	f_{ab}	$p_a = f_{aB} + f_{ab}$
marginal	$q_B = f_{AB} + f_{aB}$	$q_b = f_{Ab} + f_{ab}$	$f_{AB} + f_{Ab} + f_{aB} + f_{ab} = 1$

- Marginal distribution: 'integrate out' the allele of the other locus

$$\Pr\{B\} = q_B = f_{AB} + f_{aB}$$

- Conditional distribution: fix the allele of the other locus

$$\Pr\{A \mid b\} = \frac{\Pr\{A, b\}}{\Pr\{b\}} = \frac{f_{Ab}}{q_b} = \frac{f_{Ab}}{f_{Ab} + f_{ab}}$$

Joint, Marginal, Conditional (2/2)

Continuous case: 2 continuous random variables X and Y

- Joint distribution:

	y	marginal
x	$p_{XY}(x, y)$	$p_X(x) = \int p_{XY}(x, y) \, \mathrm{d}y$
marginal	$p_Y(y) = \int p_{XY}(x, y) \, \mathrm{d}x$	$\int p_{XY}(x, y) \, \mathrm{d}x \, \mathrm{d}y = 1$

- Marginal distribution: 'integrate out' the other variable

$$p_X(x) = \int p_{XY}(x, y) \, \mathrm{d}y$$

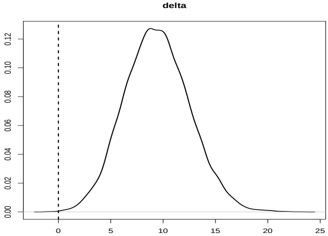
- Conditional distribution: fix the value of the other variable

$$p_{Y|X=x}(y) = \frac{p_{XY}(x, y)}{p_X(x)} = \frac{p_{XY}(x, y)}{\int p_{XY}(x, y) \, \mathrm{d}y}$$

Posterior distribution and CI

The same holds for combination of parameters, e.g.

$$\delta = \theta_2 - \theta_3$$



	post.mean	post.mode	lower.CI	upper.CI
delta	9.523487	9.45102	3.628611	15.78938

Monte Carlo: Illustration (1/3)

Example. $\pi(\theta) = \mathcal{N}(0, 10)$, $g(\theta) = e^{\theta}$:

- `theta.sample = rnorm(M, mean=0, sd=sqrt(10))`
- `mean(exp(theta.sample))`

Monte Carlo: Illustration (2/3)

Properties.

- Easy to implement

```
mean(exp(rnorm(M, mean=0, sd=sqrt(10))))
```

Monte Carlo: Illustration (2/3)

Properties.

- Easy to implement

```
mean(exp(rnorm(M, mean=0, sd=sqrt(10))))
```

- Unbiased: $\mathbb{E}\left[\widehat{\mathbb{E}}(g(\boldsymbol{\theta}))\right] = \mathbb{E}(g(\boldsymbol{\theta}))$

Monte Carlo: Illustration (2/3)

Properties.

- Easy to implement

```
mean(exp(rnorm(M, mean=0, sd=sqrt(10))))
```

- Unbiased: $\mathbb{E}\left[\widehat{\mathbb{E}}(g(\boldsymbol{\theta}))\right] = \mathbb{E}(g(\boldsymbol{\theta}))$
- Precision proportional to $1/\sqrt{M}$

Monte Carlo: Illustration (2/3)

Properties.

- Easy to implement

```
mean(exp(rnorm(M, mean=0, sd=sqrt(10))))
```

- Unbiased: $\mathbb{E}\left[\widehat{\mathbb{E}}(g(\boldsymbol{\theta}))\right] = \mathbb{E}(g(\boldsymbol{\theta}))$

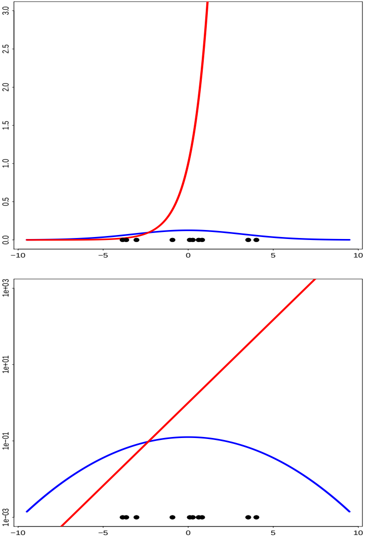
- Precision proportional to $1/\sqrt{M}$

- Still, very variant in practice (see next)

Monte Carlo: Illustration (3/3)

$$\theta \sim \mathcal{N}(0, 10), \quad g(\theta) = e^{\theta}$$

	mean	sd
1000	194.67	338.96
10000	139.63	47.24
1e+05	155.65	86.93
1e+06	147.76	15.68
truth	148.41	–


3.0
2.5
2.0
1.5
1.0
0.5
0.0
−10
−5
0
5
10
1e+03
1e+01
1e−01
1e−03
−10
−5
0
5
10