

An introduction to Bayesian statistical inference

S. Robin



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

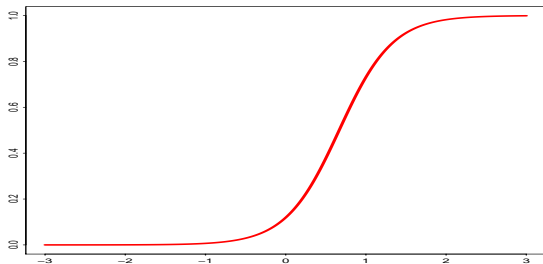
A statistical model

Logistic regression

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

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

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



Frequentist inference

θ = fixed parameter:

- ▶ Statistical model:

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

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

$$\hat{\theta} : \quad \log p_{\hat{\theta}}(\mathbf{Y}) = \max_{\theta} \log p_{\theta}(\mathbf{Y})$$

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

Frequentist inference

θ = fixed parameter:

- ▶ Statistical model:

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

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

$$\hat{\theta} : \quad \log p_{\hat{\theta}}(\mathbf{Y}) = \max_{\theta} \log p_{\theta}(\mathbf{Y})$$

- ▶ The estimate $\hat{\theta}$ 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

θ = random parameter:

- ▶ Statistical model:

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

- ▶ Inference: provide the conditional distribution of θ given the observed data \mathbf{Y} :

$$p(\theta \mid \mathbf{Y}) \quad (= \textit{posterior distribution})$$

→ credibility intervals

- ▶ Requires to define a marginal distribution:

$$\pi(\theta) := p(\theta) \quad (= \textit{prior distribution})$$

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:

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

- ▶ $P(B)$ = marginal probability of B
- ▶ $P(A, B)$ = joint probability of A and B
- ▶ $P(A|B)$ = conditional probability of A given B [# 57]

Why 'Bayes'

Bayes formula:

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

- ▶ $P(B)$ = marginal probability of B
- ▶ $P(A, B)$ = joint probability of A and B
- ▶ $P(A|B)$ = conditional probability of A given B [# 57]

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.

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

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

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

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

Bayes formula for Bayesian inference (2/2)

Some remarks.

Bayes formula for Bayesian inference (2/2)

Some remarks.

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

$$p(\theta | \mathbf{Y}) = \frac{\pi(\theta) \ell(\mathbf{Y} | \theta)}{p(\mathbf{Y})} \quad \Leftrightarrow \quad [\theta | \mathbf{Y}] = \frac{[\theta] [\mathbf{Y} | \theta]}{[\mathbf{Y}]}$$

Bayes formula for Bayesian inference (2/2)

Some remarks.

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

$$p(\theta | \mathbf{Y}) = \frac{\pi(\theta) \ell(\mathbf{Y} | \theta)}{p(\mathbf{Y})} \quad \Leftrightarrow \quad [\theta | \mathbf{Y}] = \frac{[\theta] [\mathbf{Y} | \theta]}{[\mathbf{Y}]}$$

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 denotes $[\cdot]$:

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

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

3. Obviously

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

→ $p(\boldsymbol{\theta} | \mathbf{Y})$ 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 denotes $[\cdot]$:

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

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

3. Obviously

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

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

4. Obviously, the posterior $p(\boldsymbol{\theta} | \mathbf{Y})$ 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
- ▶ \cdot : number sick / n

The posterior depends on the prior

Data & Model:

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

Param:

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

The posterior depends on the prior

Data & Model:

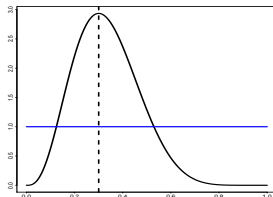
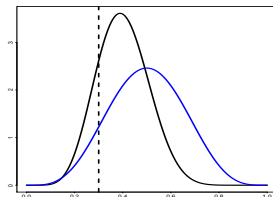
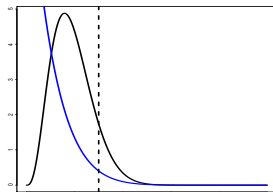
- ▶ $Y_i = 1$ if sick, 0 otherwise
- ▶ $n = 10$ patients
- ▶ \cdot : number sick / n

Param:

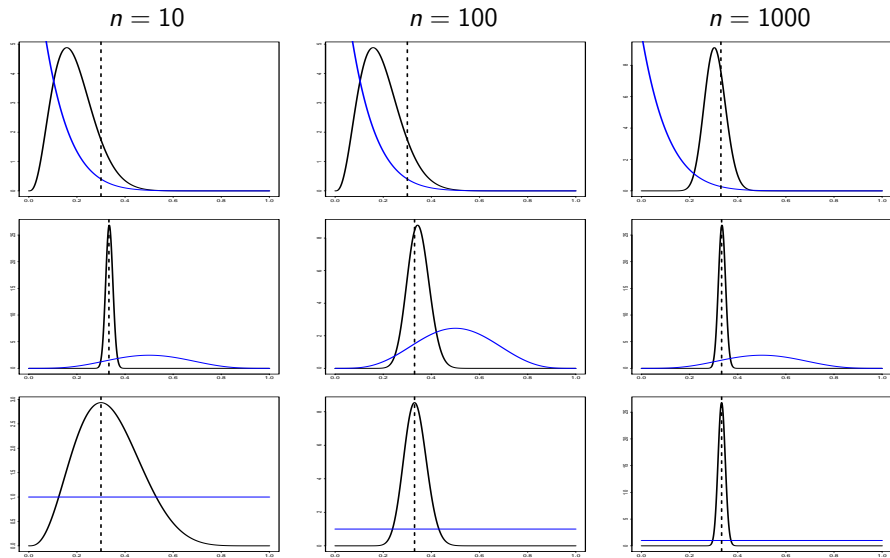
- ▶ θ = proba. sick
- ▶ — : prior $\pi(\theta)$

Output:

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



Dependency vanishes when n increases



Back to logistic regression

Model

- Prior: all coefficient θ_j independent:

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

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

$$\Pr\{Y_i = 1 \mid \boldsymbol{\theta}\} = e^{\mathbf{x}_i^T \boldsymbol{\theta}} / (1 + e^{\mathbf{x}_i^T \boldsymbol{\theta}})$$

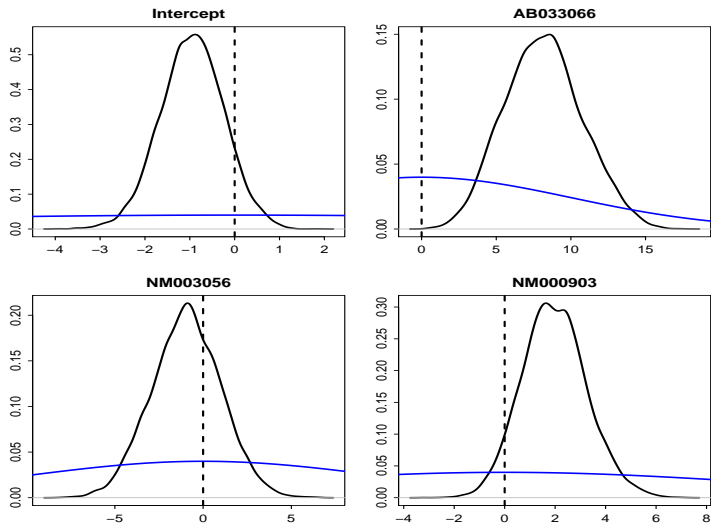
Inference:

$$\boldsymbol{\theta} \mid \mathbf{Y} \sim ?$$

(see later, but $p(\boldsymbol{\theta} \mid \mathbf{Y}) \neq \mathcal{N}(\cdot, \cdot)$, for sure.)

Bayesian inference

Output:



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'.

posterior mean: $\hat{\theta}_j = \mathbb{E}(\theta_j | \mathbf{Y})$

posterior mode: $\hat{\theta}_j = \arg \max_{\theta_j} p(\theta_j | \mathbf{Y})$

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

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

Posterior distribution and confidence intervals

Parameter 'estimate'.

posterior mean: $\hat{\theta}_j = \mathbb{E}(\theta_j | \mathbf{Y})$

posterior mode: $\hat{\theta}_j = \arg \max_{\theta_j} p(\theta_j | \mathbf{Y})$

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

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

Example. [# 4]

	post.mean	post.mode	lower.CI	upper.CI
Intercept	-0.9816181	-0.8652166	-2.41342	0.3704264
AB033066	8.395169	8.587083	3.271464	13.98373
NM003056	-1.042483	-0.9854149	-5.046639	2.819764
NM000903	1.911312	1.677234	-0.4240319	4.452512

Accounting for uncertainty

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

Model answer:

$$\Pr\{Y_0 = 1 \mid \boldsymbol{\theta}\} = e^{\mathbf{x}_0^\top \boldsymbol{\theta}} / (1 + e^{\mathbf{x}_0^\top \boldsymbol{\theta}})$$

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

Bayesian answer: *posterior predictive* probability

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

Model comparison (1/2)

Problem. Which model fits the data better:

M_0 : none of the genes has an effect, i.e. $\boldsymbol{\theta} = (\theta_0, 0, \dots, 0)$

M_1 : only the first gene has an effect, i.e. $\boldsymbol{\theta} = (\theta_0, \theta_1, 0, \dots, 0)$

...

M_p : all genes have an effect, i.e. $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_p)$

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

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

Model comparison (2/2)

Ingredients:

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

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

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

$$\theta_j | M_k \begin{cases} \sim \mathcal{N}(0, 100) & \text{if } j \leq k \\ = 0 & \text{otherwise} \end{cases}$$

Model comparison (2/2)

Ingredients:

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

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

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

$$\theta_j | M_k \begin{cases} \sim \mathcal{N}(0, 100) & \text{if } j \leq k \\ = 0 & \text{otherwise} \end{cases}$$

Recipe:

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

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

Model comparison (2/2)

Ingredients:

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

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

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

$$\theta_j | M_k \begin{cases} \sim \mathcal{N}(0, 100) & \text{if } j \leq k \\ = 0 & \text{otherwise} \end{cases}$$

Recipe:

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

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

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

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

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 \hat{M} , i.e. with largest posterior $p(M | \mathbf{Y})$
- ▶ Compute

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

Model averaging (uncertainty on models)

Question: Probability for patient 0 to be sick?

Model selection.

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

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

Model averaging.

- ▶ Keep all models
- ▶ Compute

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

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:

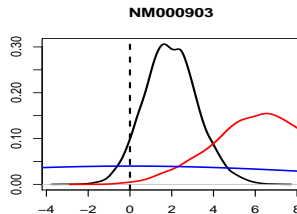
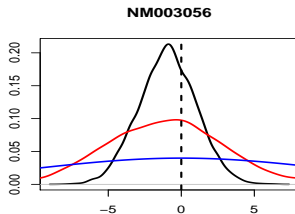
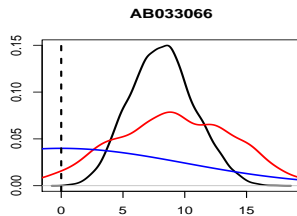
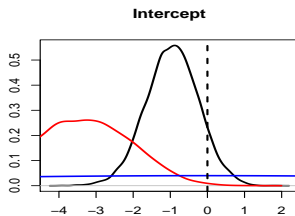
$$p(\theta | \mathbf{Y}_1, \mathbf{Y}_2) = \frac{p(\theta | \mathbf{Y}_1)p(\mathbf{Y}_2 | \theta, \mathbf{Y}_1)}{p(\mathbf{Y}_2 | \mathbf{Y}_1)}$$

In practice:

1. Perform inference using \mathbf{Y}_1 to get $p(\theta | \mathbf{Y}_1)$ from prior $\pi(\theta)$
2. Then perform inference using \mathbf{Y}_2 to get $p(\theta | \mathbf{Y}_1, \mathbf{Y}_2)$ using $p(\theta | \mathbf{Y}_1)$ 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: $f(\boldsymbol{\theta}) = \theta_j^2$ (+ 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: $f(\boldsymbol{\theta}) = \theta_j^2$ (+ posterior mean)

Main problem: evaluate

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

which requires to evaluate

$$p(\mathbf{Y}) = \int \underbrace{\pi(\boldsymbol{\theta})}_{\text{prior}} \underbrace{p(\mathbf{Y}|\boldsymbol{\theta})}_{\text{likelihood}} d\boldsymbol{\theta}$$

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: Bernoulli¹

Prior: θ = probability to be sick.

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

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

Nice case: Conjugate priors

Example: Bernoulli¹

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

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

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

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

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

Nice case: Conjugate priors

Example: Bernoulli¹

Prior: θ = probability to be sick.

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

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

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

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	α, β	$\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures ^[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures ^[note 1]	BetaBin($\tilde{x} \alpha', \beta'$) (beta-binomial)
Negative Binomial with known failure number r	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^n x_i, \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures ^[note 1] (i.e. $\frac{\beta-1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	k, θ	$k + \sum_{i=1}^n x_i, \frac{\theta}{n\theta + 1}$	k total occurrences in $1/\theta$ intervals	NB($\tilde{x} k', \frac{\theta'}{1+\theta'}$) (negative binomial)
Poisson	λ (rate)	Gamma	α, β ^[note 3]	$\alpha + \sum_{i=1}^n x_i, \beta + n$	α total occurrences in β intervals	NB($\tilde{x} \alpha', \frac{1}{1+\beta'}$) (negative binomial)
Categorical	\mathbf{p} (probability vector), k (number of categories, i.e. size of \mathbf{p})	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \dots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ 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	\mathbf{p} (probability vector), k (number of categories, i.e. size of \mathbf{p})	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^n \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i ^[note 1]	DirMult($\tilde{\mathbf{x}} \boldsymbol{\alpha}'$) (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	α, β	$\alpha + n, \beta + \sum_{i=1}^n x_i$	$\alpha - 1$ 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	μ_0, σ_0^2	$\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n x_i}{\sigma^2} \right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right),$ $\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\bar{x} \mu_0', \sigma_0'^2 + \sigma^2)^{[5]}$
Normal with known precision τ	μ (mean)	Normal	μ_0, τ_0	$\left(\tau_0 \mu_0 + \tau \sum_{i=1}^n x_i \right) / (\tau_0 + n\tau), \tau_0 + n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\bar{x} \mu_0', \frac{1}{\tau_0} + \frac{1}{\tau}\right)^{[5]}$
Normal with known mean μ	σ^2 (variance)	Inverse gamma	α, β ^[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)	$t_{2\alpha'}(\bar{x} \mu, \sigma^2 = \beta'/\alpha')^{[5]}$
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	ν, σ_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	$t_{\nu'}(\bar{x} \mu, \sigma_0'^2)^{[5]}$
Normal with known mean μ	τ (precision)	Gamma	α, β ^[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)	$t_{2\alpha'}(\bar{x} \mu, \sigma^2 = \beta'/\alpha')^{[5]}$
Normal ^[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2},$ $\beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{\nu + n} \frac{(\bar{x} - \mu_0)^2}{2}$ ▪ \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	$t_{2\alpha'}\left(\bar{x} \mu', \frac{\beta'(\nu' + 1)}{\alpha'\nu'}\right)^{[5]}$
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2},$ $\beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{\nu + n} \frac{(\bar{x} - \mu_0)^2}{2}$ ▪ \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	$t_{2\alpha'}\left(\bar{x} \mu', \frac{\beta'(\nu' + 1)}{\alpha'\nu'}\right)^{[5]}$
Multivariate normal with known covariance matrix Σ	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0, \Sigma_0$	$(\Sigma_0^{-1} + n\Sigma^{-1})^{-1} (\Sigma_0^{-1} \boldsymbol{\mu}_0 + n\Sigma^{-1} \bar{\mathbf{x}}),$ $(\Sigma_0^{-1} + n\Sigma^{-1})^{-1}$ ▪ $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) Σ_0^{-1} and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}(\bar{\mathbf{x}} \boldsymbol{\mu}_0', \Sigma_0' + \Sigma)^{[5]}$

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

Goal: compute

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

where

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

Computing integrals

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

Goal: compute

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

where

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

We need to evaluate integrals of the form

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

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

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

Monte Carlo

Principle. To evaluate

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

1. sample

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

2. compute

$$\hat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{M} \sum_m f(\boldsymbol{\theta}^m)$$

Monte Carlo

Principle. To evaluate

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

1. sample

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

2. compute

$$\hat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{M} \sum_m f(\boldsymbol{\theta}^m)$$

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/M$. [# 59]

Monte Carlo

Principle. To evaluate

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

1. sample

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

2. compute

$$\hat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{M} \sum_m f(\boldsymbol{\theta}^m)$$

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/M$. [# 59]

In practice:

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

$$\hat{\mathbb{E}}_{\mathcal{N}(0,10)} [e^\theta] = \text{mean}(\exp(\text{rnorm}(M, \text{mean}=0, \text{sd}=\text{sqrt}(10))))$$

Monte Carlo

Principle. To evaluate

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

1. sample

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

2. compute

$$\hat{\mathbb{E}}_q[f(\boldsymbol{\theta})] = \frac{1}{M} \sum_m f(\boldsymbol{\theta}^m)$$

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/M$. [# 59]

In practice:

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

$$\hat{\mathbb{E}}_{\mathcal{N}(0,10)}[e^\theta] = \text{mean}(\exp(\text{rnorm}(M, \text{mean}=0, \text{sd}=\text{sqrt}(10))))$$

- 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

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

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

1. sample

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

Importance Sampling (IS)

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

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

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

1. sample

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

2. compute the weights

$$W(\boldsymbol{\theta}^m) = q(\boldsymbol{\theta}^m)/q'(\boldsymbol{\theta}^m),$$

Importance Sampling (IS)

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

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

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

1. sample

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

2. compute the weights

$$W(\boldsymbol{\theta}^m) = q(\boldsymbol{\theta}^m)/q'(\boldsymbol{\theta}^m),$$

3. and compute

$$\widehat{\mathbb{E}}[f(\boldsymbol{\theta})] = \frac{1}{M} \sum_m W(\boldsymbol{\theta}^m) f(\boldsymbol{\theta}^m)$$

Importance Sampling (IS)

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

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

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

1. sample

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

2. compute the weights

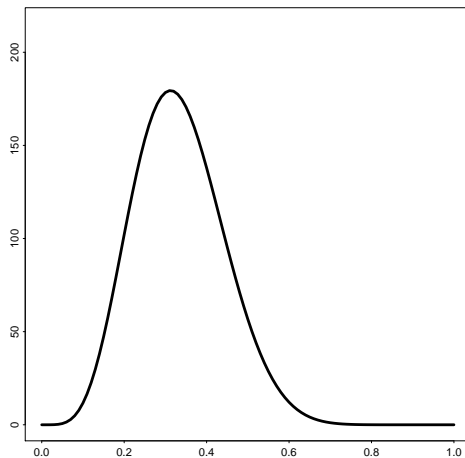
$$W(\boldsymbol{\theta}^m) = q(\boldsymbol{\theta}^m)/q'(\boldsymbol{\theta}^m),$$

3. and compute

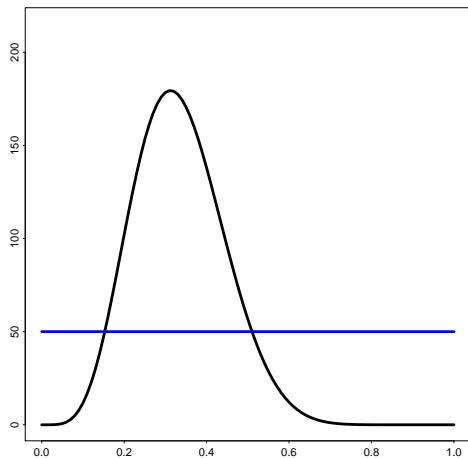
$$\widehat{\mathbb{E}}[f(\boldsymbol{\theta})] = \frac{1}{M} \sum_m W(\boldsymbol{\theta}^m) f(\boldsymbol{\theta}^m)$$

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

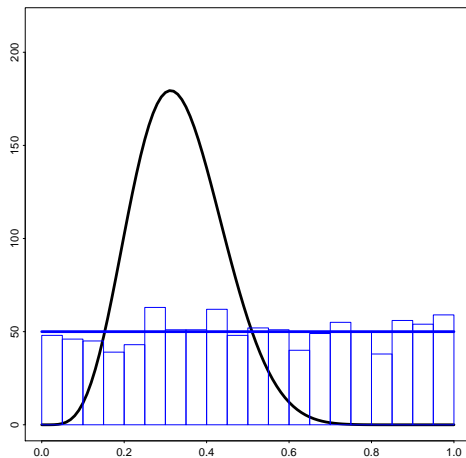
Importance Sampling (a picture)



Importance Sampling (a picture)



Importance Sampling (a picture)



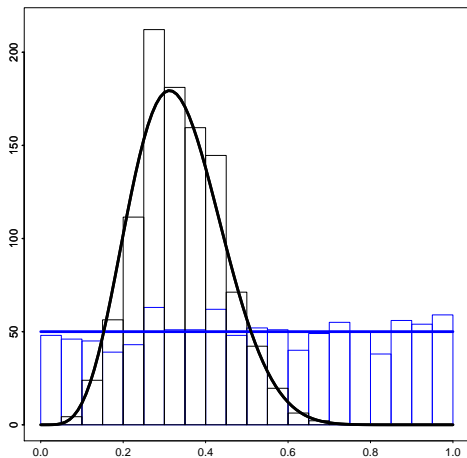
Importance Sampling (a picture)

Efficiency of sampling:

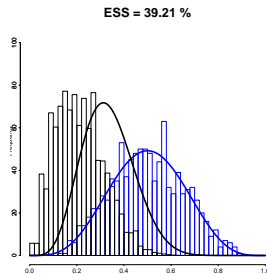
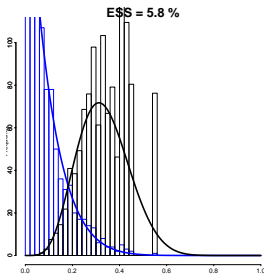
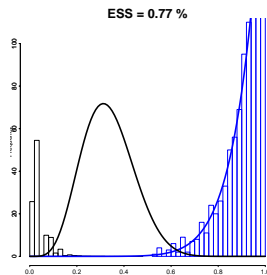
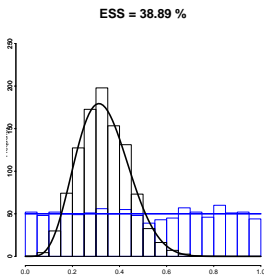
$$ESS = \overline{W}^2 / \overline{W^2}$$

$$q' = q$$

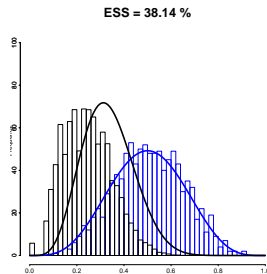
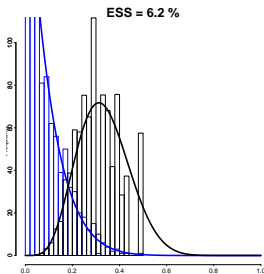
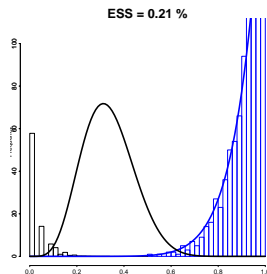
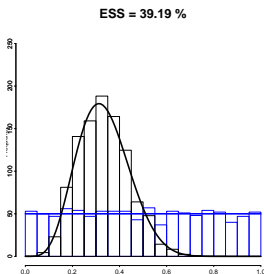
$$\Rightarrow ESS = 1$$



Importance Sampling: Importance of the proposal



Importance of the proposal: another draw



IS for posterior sampling

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

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{\theta})|\mathbf{Y}] &= \int f(\boldsymbol{\theta})p(\boldsymbol{\theta}, \mathbf{Y}) \, d\boldsymbol{\theta} / p(\mathbf{Y}) = \dots \\ &= \int f(\boldsymbol{\theta}) \frac{\pi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{Y})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \, d\boldsymbol{\theta} / \int \frac{\pi(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{Y})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \, d\boldsymbol{\theta}\end{aligned}$$

1. sample

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

2. compute the weights

$$W(\boldsymbol{\theta}^m) = \pi(\boldsymbol{\theta}^m)p(\boldsymbol{\theta}^m|\mathbf{Y}) / q(\boldsymbol{\theta}^m)$$

3. get

$$\widehat{\mathbb{E}}[f(\boldsymbol{\theta})|\mathbf{Y}] = \sum_m W(\boldsymbol{\theta}^m) f(\boldsymbol{\theta}^m) / \sum_m W(\boldsymbol{\theta}^m)$$

→ slightly biased.

Good proposals

Choosing q is critical

Good proposals

Choosing q is critical

Typical choices

- Prior

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

→ far from the target $p(\boldsymbol{\theta} | \mathbf{Y})$: small ESS

Good proposals

Choosing q is critical

Typical choices

- Prior

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

→ far from the target $p(\boldsymbol{\theta} | \mathbf{Y})$: small ESS

- MLE:

$$q(\boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{\theta}}_{MLE}, \mathbb{V}_{\infty}(\hat{\boldsymbol{\theta}}_{MLE}))$$

→ 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 $p(\boldsymbol{\theta} | \mathbf{Y})$: small ESS

- MLE:

$$q(\boldsymbol{\theta}) = \mathcal{N}(\hat{\boldsymbol{\theta}}_{MLE}, \mathbb{V}_{\infty}(\hat{\boldsymbol{\theta}}_{MLE}))$$

→ fine, as long as MLE is available

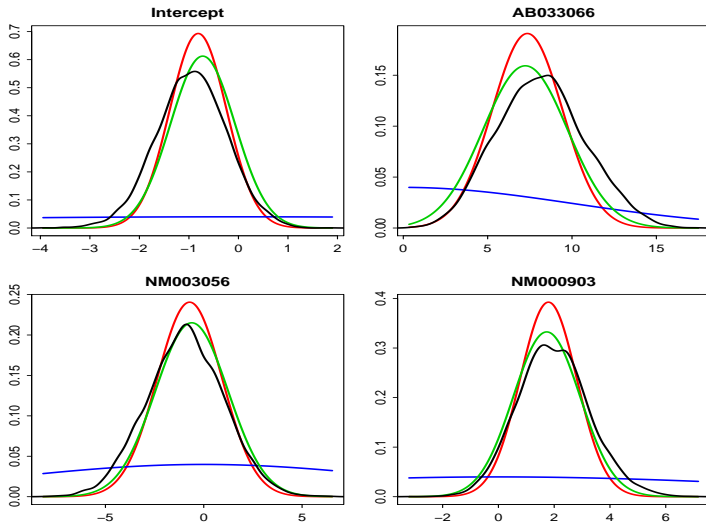
- Variational Bayes, expectation propagation, ...:

$$q(\boldsymbol{\theta}) = \arg \min_{q \in \mathcal{Q}} KL[q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} | \mathbf{Y})]$$

→ fast and reasonably accurate

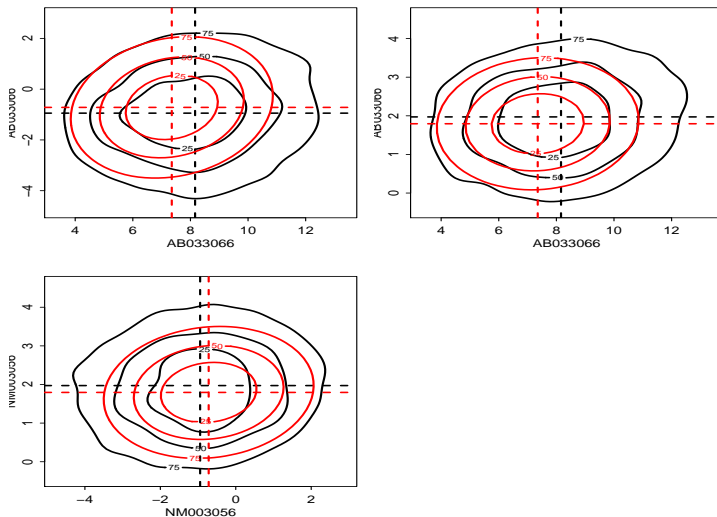
Variational Bayes & ML as a prior

— : 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 $\{\phi^t\}_{t \geq 0}$ is an ergodic Markov chain (irreducible, aperiodic, ...) with

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

$$p(\{\phi^t\}) = \nu(\phi^0) \times \kappa(\phi^1 | \phi^0) \times \kappa(\phi^2 | \phi^1) \times \kappa(\phi^3 | \phi^2) \times \dots$$

then

- ▶ it admits a **unique stationary distribution** μ :

$$\phi^{t-1} \sim \mu \quad \Rightarrow \quad \phi^t \sim \mu$$

- ▶ ϕ^t converges towards μ in distribution

$$\phi^t \xrightarrow[t \rightarrow \infty]{\Delta} \mu$$

for **any initial distribution** ν

Use for Bayesian inference

Aim. Sample from

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

Idea.

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

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

- ▶ Choose 'any' initial ν and simulate $\{\boldsymbol{\theta}^t\}_{t \geq 0}$
- ▶ Until it 'reaches' its stationary distribution

Metropolis-Hastings

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

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot | \theta)$

- ▶ Start with θ^0

Metropolis-Hastings

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

- ▶ Start with θ^0
- ▶ At step t ,

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot | \theta)$

- ▶ Start with θ^0
- ▶ At step t ,
 1. sample $\theta' \sim \lambda(\cdot | \theta^{t-1})$;

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot | \theta)$

- ▶ Start with θ^0
- ▶ At step t ,
 1. sample $\theta' \sim \lambda(\cdot | \theta^{t-1})$;
 2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\theta', \theta^{t-1}) = \frac{\lambda(\theta^{t-1} | \theta')}{\lambda(\theta' | \theta^{t-1})} \frac{p(\theta' | \mathbf{Y})}{p(\theta^{t-1} | \mathbf{Y})}$$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot | \theta)$

- ▶ Start with θ^0
- ▶ At step t ,
 1. sample $\theta' \sim \lambda(\cdot | \theta^{t-1})$;
 2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\theta', \theta^{t-1}) = \frac{\lambda(\theta^{t-1} | \theta')}{\lambda(\theta' | \theta^{t-1})} \frac{p(\theta' | \mathbf{Y})}{p(\theta^{t-1} | \mathbf{Y})} = \frac{\lambda(\theta^{t-1} | \theta')}{\lambda(\theta' | \theta^{t-1})} \frac{\pi(\theta') \ell(\mathbf{Y} | \theta')}{\pi(\theta^{t-1}) \ell(\mathbf{Y} | \theta^{t-1})};$$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot | \theta)$

- ▶ Start with θ^0
- ▶ At step t ,
 1. sample $\theta' \sim \lambda(\cdot | \theta^{t-1})$;
 2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\theta', \theta^{t-1}) = \frac{\lambda(\theta^{t-1} | \theta')}{\lambda(\theta' | \theta^{t-1})} \frac{p(\theta' | \mathbf{Y})}{p(\theta^{t-1} | \mathbf{Y})} = \frac{\lambda(\theta^{t-1} | \theta')}{\lambda(\theta' | \theta^{t-1})} \frac{\pi(\theta') \ell(\mathbf{Y} | \theta')}{\pi(\theta^{t-1}) \ell(\mathbf{Y} | \theta^{t-1})};$$

3. set $\theta^t = \begin{cases} \theta' & \text{with probability } \min(1, \alpha(\theta', \theta^{t-1})), \\ \theta^{t-1} & \text{otherwise.} \end{cases}$

Metropolis-Hastings

Algorithm. Define a shift kernel $\lambda(\cdot | \theta)$

- ▶ Start with θ^0
- ▶ At step t ,
 1. sample $\theta' \sim \lambda(\cdot | \theta^{t-1})$;
 2. compute the Metropolis-Hastings ratio (acceptance probability)

$$\alpha(\theta', \theta^{t-1}) = \frac{\lambda(\theta^{t-1} | \theta')}{\lambda(\theta' | \theta^{t-1})} \frac{p(\theta' | \mathbf{Y})}{p(\theta^{t-1} | \mathbf{Y})} = \frac{\lambda(\theta^{t-1} | \theta')}{\lambda(\theta' | \theta^{t-1})} \frac{\pi(\theta') \ell(\mathbf{Y} | \theta')}{\pi(\theta^{t-1}) \ell(\mathbf{Y} | \theta^{t-1})};$$

3. set $\theta^t = \begin{cases} \theta' & \text{with probability } \min(1, \alpha(\theta', \theta^{t-1})), \\ \theta^{t-1} & \text{otherwise.} \end{cases}$

Properties.

1. λ and α defined a Markov chain with stationary distribution $\mu(\theta) = p(\theta | \mathbf{Y})$.
2. If $\lambda(\cdot | \theta)$ is symmetric, α reduce to $\pi(\theta') \ell(\mathbf{Y} | \theta') / [\pi(\theta^{t-1}) \ell(\mathbf{Y} | \theta^{t-1})]$

Metropolis-Hastings for logistic regression

Model.

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

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

Metropolis-Hastings for logistic regression

Model.

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

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

Algorithm settings.

$$\boldsymbol{\theta}^0 = \mathbf{0}_p$$

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

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, M, 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, M, 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, M, 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, M, 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, M, 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, M, 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 (m in 2:M){
```


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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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){
```

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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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[m, ] = 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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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[m, ] = theta.tmp;
    logprior.cur = logprior.tmp;
    loglik.cur = loglik.tmp
  }else{
    theta.sample[m, ] = theta.sample[m-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, M, 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 (m in 2:M){
  theta.tmp = rmvnorm(1, mean=theta.sample[m-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[m, ] = theta.tmp;
    logprior.cur = logprior.tmp;
    loglik.cur = loglik.tmp
  }else{
    theta.sample[m, ] = theta.sample[m-1, ]
  }
}
}

```

Gibbs

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)

When the likelihood is intractable

References

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

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

Joint, Marginal, Conditional (2/2)

Continuous case: 2 continuous random variables X and Y

- Joint distribution:

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

- Marginal distribution: 'integrate out' the other variable

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

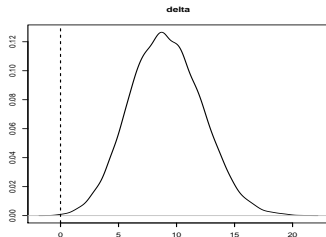
- Conditional distribution: fix the value of the other variable

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

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	9.389825	8.824822	3.539045	15.84367

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	—

