

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 methods

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

Extensions

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

Reminder: Joint, marginal, conditional (1/2)

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

Reminder: 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	$q_B = f_{AB} + f_{aB}$	$q_b = f_{Ab} + f_{ab}$	$f_{AB} + f_{Ab} + f_{aB} + f_{ab} = 1$

Reminder: 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

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

Reminder: 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

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

('Bayes formula')

Reminder: Joint, marginal, conditional (2/2)

Continuous case: 2 continuous random variables U and V

- Joint distribution:

	v	marginal
u	$p_{UV}(u, v)$	$p_U(u) = \int p_{UV}(u, v) dv$
marginal	$p_V(v) = \int p_{UV}(u, v) du$	$\int p_{UV}(u, v) du dv = 1$

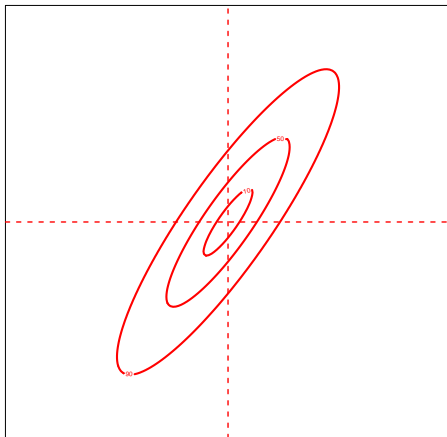
- Marginal distribution: 'integrate out' the other variable

$$p_U(u) = \int p_{UV}(u, v) dv$$

- Conditional distribution: fix the value of the other variable

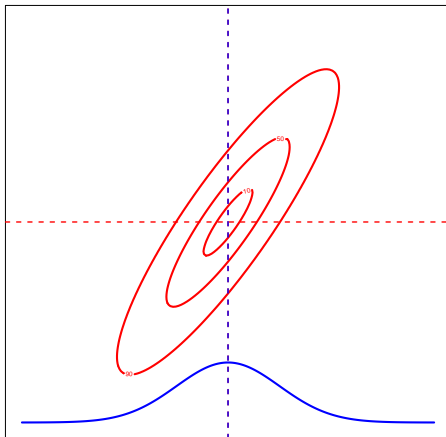
$$p_{V|U=u}(v) = \frac{p_{UV}(u, v)}{p_U(u)} = \frac{p_{UV}(u, v)}{\int p_{UV}(u, v) dv}$$

A picture



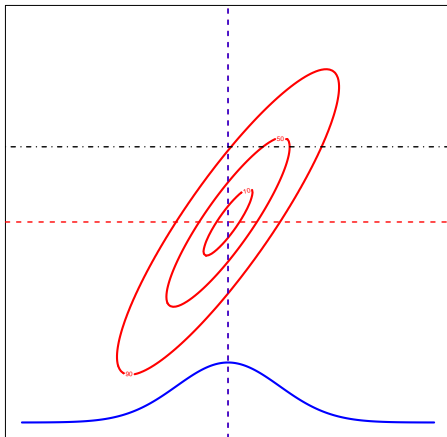
Joint,

A picture



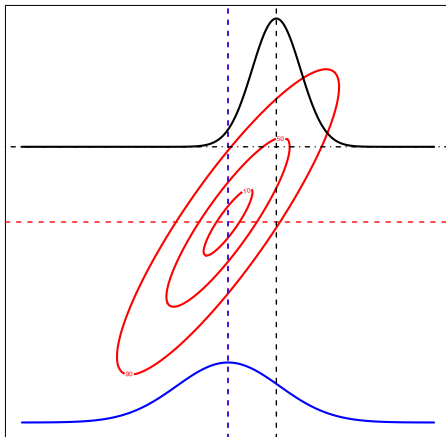
Joint, Marginal,

A picture



Joint, Marginal, Conditional

A picture



Joint, Marginal, Conditional

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 methods

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 methods

Conjugate priors

Monte Carlo integration

Monte Carlo Markov chains (MCMC)

Extensions

Sequential Monte-Carlo (SMC)



Approximate Bayesian computation (ABC)

An example

Example:

- ▶ n tissue samples: $i = 1 \dots n$
- ▶ $Y_i = \text{status}$ (0 = normal, 1 = tumor) of sample i
- ▶ $\mathbf{x}_i = (x_{i1}, \dots, x_{ip}) = \text{vector of gene expression for sample } i \text{ (gene } j = 1 \dots p)$

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
	 X				 Y

An example

Example:

- ▶ n tissue samples: $i = 1 \dots n$
- ▶ $Y_i = \text{status}$ ($0 = \text{normal}$, $1 = \text{tumor}$) of sample i
- ▶ $\mathbf{x}_i = (x_{i1}, \dots, x_{ip}) = \text{vector of gene expression for sample } i \text{ (gene } j = 1 \dots p)$

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
					
	X				Y

Similar question for genotyping data: $x_{ij} \in \{0, 1, 2\}$.

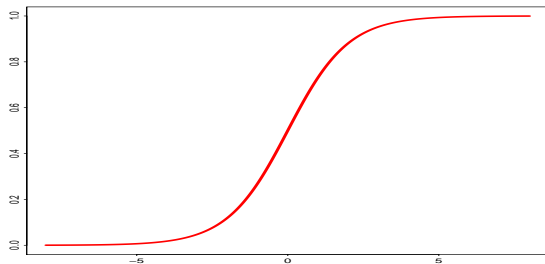
A statistical model

Model: Logistic regression

- ▶ The samples are independent.
- ▶ The probability for sample i to be tumor depends on \mathbf{x}_i :

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

- ▶ $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$: 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. maximum likelihood

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

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

Frequentist inference

θ = fixed parameter:

- ▶ Statistical model:

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

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

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

- ▶ The estimate $\hat{\theta}$ itself is random (depends on the data) \rightarrow conf. 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:

Bayesian inference

θ = random parameter:

- Statistical model:

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

Bayesian inference

θ = random parameter:

- Statistical model:

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

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

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

→ credibility intervals

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 methods

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

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

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.

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

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

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

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

Bayes formula for Bayesian inference (2/2)

Remarks.

Bayes formula for Bayesian inference (2/2)

Remarks.

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

Bayes formula for Bayesian inference (2/2)

Remarks.

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

Remarks.

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

2. 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 $p(\mathbf{Y})$**

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

Bayes formula for Bayesian inference (2/2)

Remarks.

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

2. 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 $p(\mathbf{Y})$**

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

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

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

The posterior depends on the prior

Data & Model:

- ▶ $Y_i = 1$ if disease, 0 otherwise
- ▶ $n = 10$ patients
- ▶ $\dot{\cdot}$: number disease carriers/ n

The posterior depends on the prior

Data & Model:

- ▶ $Y_i = 1$ if disease, 0 otherwise
- ▶ $n = 10$ patients
- ▶ $\dot{\cdot}$: number disease carriers/ n

Param:

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

The posterior depends on the prior

Data & Model:

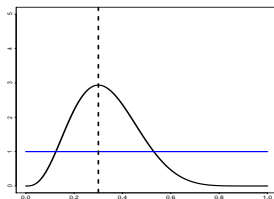
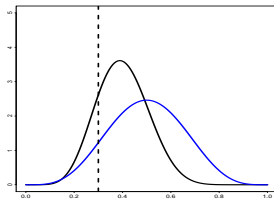
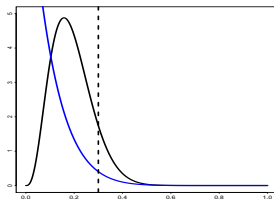
- ▶ $Y_i = 1$ if disease, 0 otherwise
- ▶ $n = 10$ patients
- ▶ $\dot{\cdot}$: number disease carriers/ n

Param:

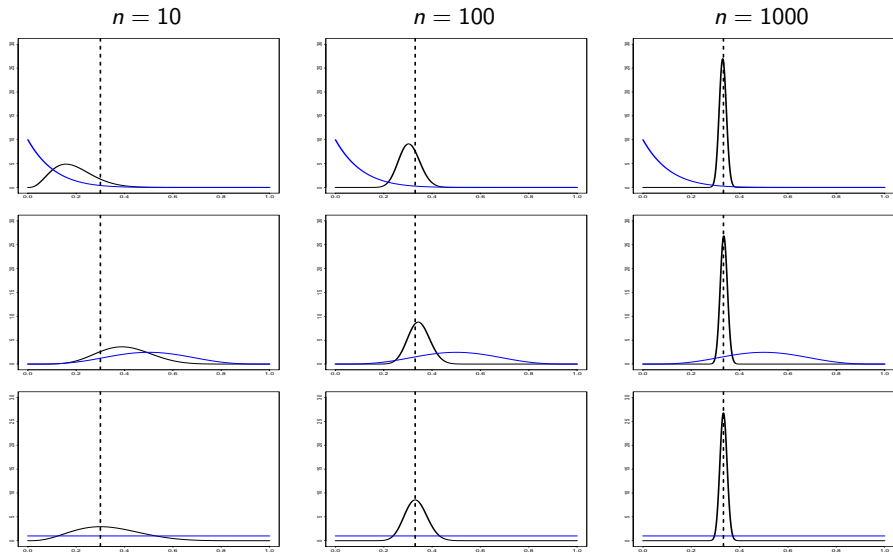
- ▶ θ = proba. disease
- ▶ — : prior $\pi(\theta)$

Output:

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



Dependency vanishes when n increases



Back to logistic regression

Model

- Prior $\pi(\boldsymbol{\theta})$: coefficients θ_j all independent:

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

- Likelihood $\ell(\mathbf{Y} | \boldsymbol{\theta})$: all samples independent, *conditionally* on $\boldsymbol{\theta}$:

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

Back to logistic regression

Model

- Prior $\pi(\boldsymbol{\theta})$: coefficients θ_j all independent:

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

- Likelihood $\ell(\mathbf{Y} | \boldsymbol{\theta})$: all samples independent, *conditionally* on $\boldsymbol{\theta}$:

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

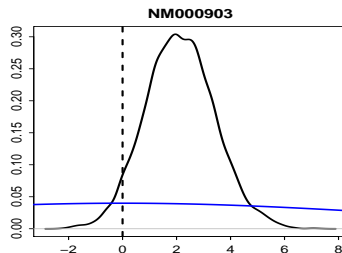
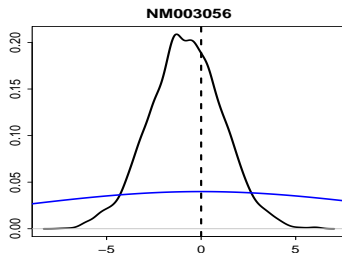
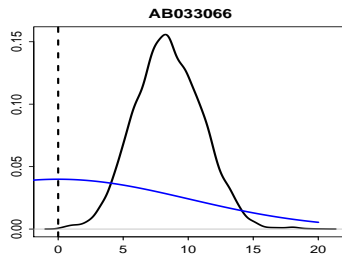
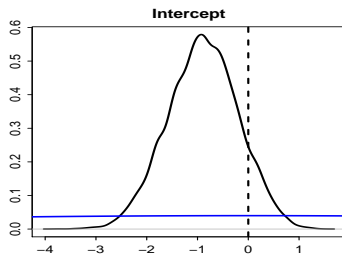
Inference:

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

(for sure: $\neq \mathcal{N}(\cdot, \cdot)$).

Bayesian inference

Output:



No test (and no estimator)

Frequentist hypothesis:

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

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

No test (and no estimator)

Frequentist hypothesis:

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

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

Bayesian assessment:

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

No test (and no estimator)

Frequentist hypothesis:

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

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

Bayesian assessment:

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

Parameter estimate. For the same reason:

$\hat{\theta}$ can no be an estimate of θ

(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 methods

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'.

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

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

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

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

Posterior distribution and confidence intervals

Parameter 'estimate'.

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

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

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

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

Example. [# 64]

	post.mean	post.mode	lower.CI	upper.CI
Intercept	-0.890718	-0.9281564	-2.266244	0.4477921
AB033066	8.483059	8.229936	3.595861	13.60357
NM003056	-0.8067056	-1.290723	-4.698653	3.000588
NM000903	2.13275	1.958102	-0.3106497	4.750834

Accounting for uncertainty

Question: What is the probability for a **new sample** 0 (with profile \mathbf{x}_0) to be tumor?

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: $\pi(\theta | M)$, 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 :

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

- 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 sample 0 to be tumor?

Model averaging (uncertainty on models)

Question: Probability for sample 0 to be tumor?

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 sample 0 to be tumor?

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

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

Model averaging: Illustration

Aim: Probability p_0 to be tumor for a sample with gene expression profile

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

Example. For models M_0, \dots, M_p :

Model	$\log pY M$	$pM Y$	Esp. $p_0 M.Y$	Sd. $p_0 M.Y$
M_0	-53.77	1e-04	0.435	0.056
M_1	-51.61	4e-04	0.609	0.095
...				
M_{11}	-47.57	0.0252	0.431	0.196
M_{12}	-45.48	0.2023	0.246	0.173
M_{13}	-45.9	0.1331	0.234	0.174
M_{14}	-45.1	0.2974	0.207	0.168
M_{15}	-45.19	0.27	0.197	0.161
<hr/>				
Averaging	$\log pY$		Esp. $p_0 Y$	Sd. $p_0 Y$
	-46.66		0.242	0.188

Transfer of uncertainty from one experience to another

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

Model:

- ▶ Prior: $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$
- ▶ Independent samples with same conditional likelihood:

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

Transfer of uncertainty from one experience to another

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

Model:

- ▶ Prior: $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$
- ▶ Independent samples with same conditional likelihood:

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

Simple algebra gives:

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

Transfer of uncertainty from one experience to another

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

Model:

- ▶ Prior: $\theta \sim \pi(\theta)$
- ▶ Independent samples with same conditional likelihood:

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

Simple algebra gives:

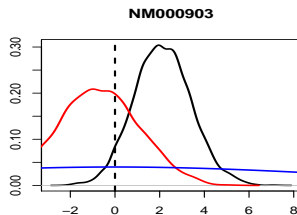
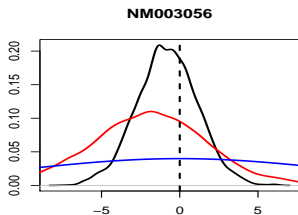
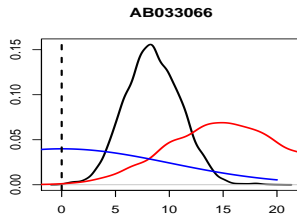
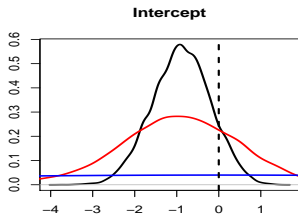
$$p(\theta | \mathbf{Y}_1, \mathbf{Y}_2) = \frac{p(\theta | \mathbf{Y}_1) p(\mathbf{Y}_2 | \theta, \mathbf{Y}_1)}{p(\mathbf{Y}_2 | \mathbf{Y}_1)} = \frac{p(\theta | \mathbf{Y}_1) \ell(\mathbf{Y}_2 | \theta)}{p(\mathbf{Y}_2 | \mathbf{Y}_1)}$$

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 methods

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 covariance: $f(\boldsymbol{\theta}) = \theta_j \theta_k$ (+ posterior means)

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 covariance: $f(\boldsymbol{\theta}) = \theta_j\theta_k$ (+ posterior means)

Main problem: evaluate

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

which requires to evaluate

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

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 methods

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 carry a disease.

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

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

Nice case: Conjugate priors

Example: Bernoulli¹

Prior: θ = probability to carry a disease.

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

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

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

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

Nice case: Conjugate priors

Example: Bernoulli¹

Prior: θ = probability to carry a disease.

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

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

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

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#16: 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	$\mathcal{N}(\tilde{x} \mu'_0, \sigma_0'^2 + \sigma^2)^{[5]}$
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	$\mathcal{N}\left(\tilde{x} \mu'_0, \frac{1}{\tau_0} + \frac{1}{\tau}\right)^{[5]}$
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)	$t_{2\alpha'}(\tilde{x} \mu, \sigma^2 = \beta'/\alpha')^{[5]}$
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	$t_{\nu'}(\tilde{x} \mu, \sigma_0'^2)^{[5]}$
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)	$t_{2\alpha'}(\tilde{x} \mu, \sigma^2 = \beta'/\alpha')^{[5]}$
Normal ^[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2},$ $\beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{\nu + n} \frac{(\bar{x} - \mu_0)^2}{2}$ ▪ \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	$t_{2\alpha'}\left(\tilde{x} \mu', \frac{\beta'(\nu' + 1)}{\alpha'\nu'}\right)^{[5]}$
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0, \nu, \alpha, \beta$	$\frac{\nu\mu_0 + n\bar{x}}{\nu + n}, \nu + n, \alpha + \frac{n}{2},$ $\beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{\nu + n} \frac{(\bar{x} - \mu_0)^2}{2}$ ▪ \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	$t_{2\alpha'}\left(\tilde{x} \mu', \frac{\beta'(\nu' + 1)}{\alpha'\nu'}\right)^{[5]}$
Multivariate normal with known covariance matrix Σ	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0, \Sigma_0$	$\left(\Sigma_0^{-1} + n\Sigma^{-1} \right)^{-1} \left(\Sigma_0^{-1} \boldsymbol{\mu}_0 + n\Sigma^{-1} \bar{\mathbf{x}} \right),$ $\left(\Sigma_0^{-1} + n\Sigma^{-1} \right)^{-1}$ ▪ $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) Σ_0^{-1} and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}(\bar{\mathbf{x}} \boldsymbol{\mu}', \Sigma_0' + \Sigma)^{[5]}$

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 methods

- 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

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

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

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

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

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

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

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

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/B$. [# 65]

Monte Carlo

Principle. To evaluate

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

1. sample

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

2. compute

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

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/B$. [# 65]

In practice:

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

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

Monte Carlo

Principle. To evaluate

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

1. sample

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

2. compute

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

→ unbiased estimate of $\mathbb{E}_q[f(\boldsymbol{\theta})]$ with variance $\propto 1/B$. [# 65]

In practice:

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

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

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

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

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 $\mathbb{E}[f(\theta)]$, write it as

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

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

1. sample

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

2. compute the weights

$$W(\theta^b) = q(\theta^b)/q'(\theta^b),$$

Importance Sampling (IS)

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

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

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

1. sample

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

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

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

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

1. sample

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

2. compute the weights

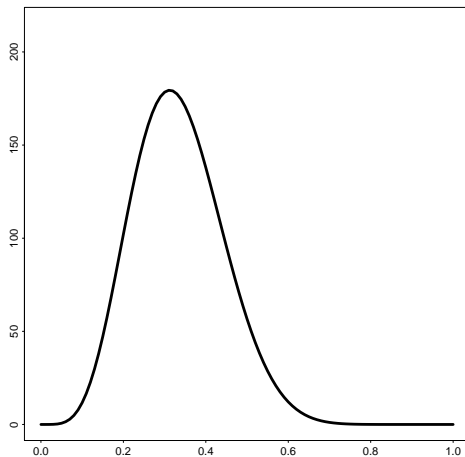
$$W(\boldsymbol{\theta}^b) = q(\boldsymbol{\theta}^b)/q'(\boldsymbol{\theta}^b),$$

3. and compute

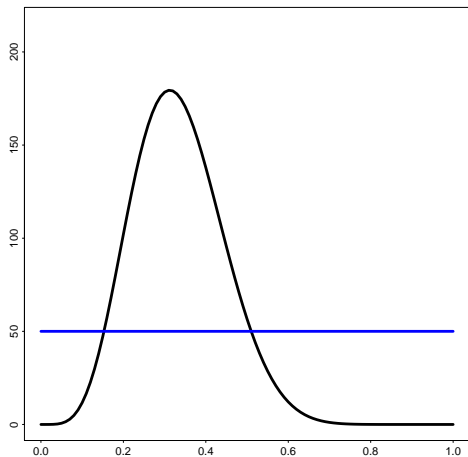
$$\widehat{\mathbb{E}}[f(\boldsymbol{\theta})] = \frac{1}{B} \sum_b W(\boldsymbol{\theta}^b)f(\boldsymbol{\theta}^b)$$

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

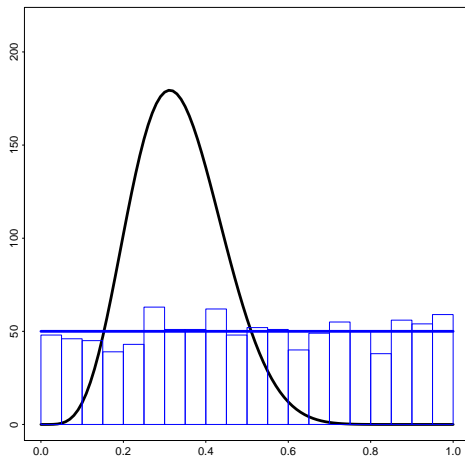
Importance Sampling (a picture)



Importance Sampling (a picture)



Importance Sampling (a picture)



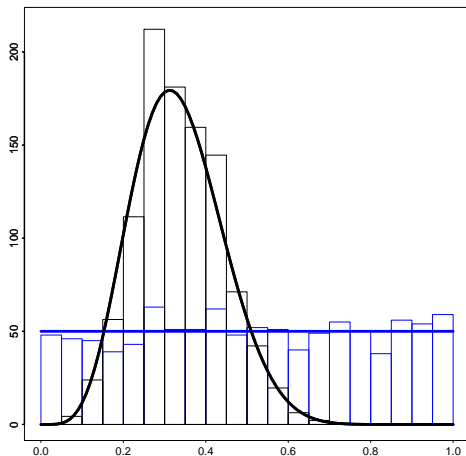
Importance Sampling (a picture)

Efficiency of sampling:

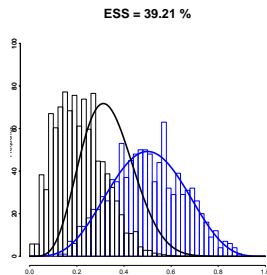
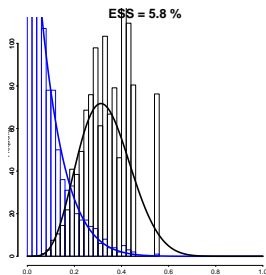
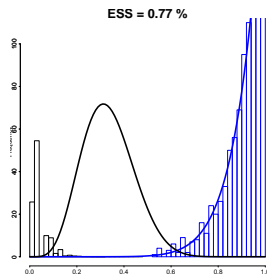
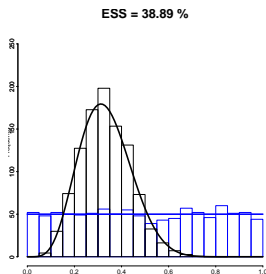
$$ESS = \overline{W}^2 / \overline{W^2}$$

$$q' = q$$

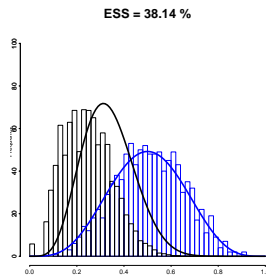
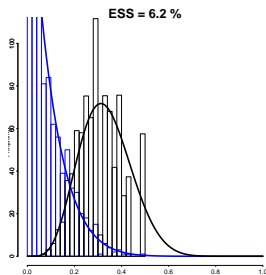
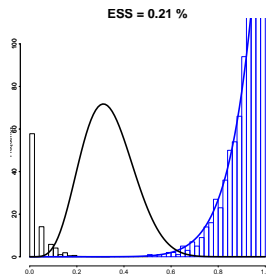
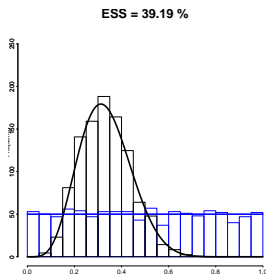
$$\Rightarrow ESS = 1$$



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

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

1. sample

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

2. compute the weights

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

3. get

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

(slightly **biased**).

Good proposals

Choosing q is critical

Good proposals

Choosing q is critical

Typical choices

- Prior

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

→ 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

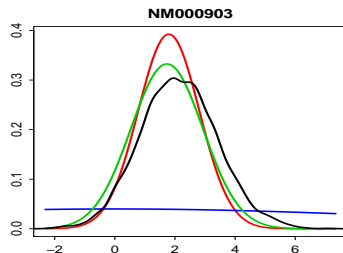
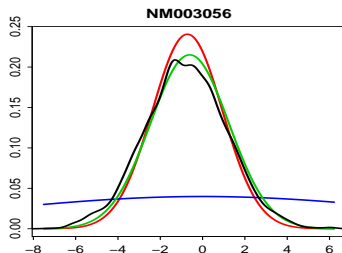
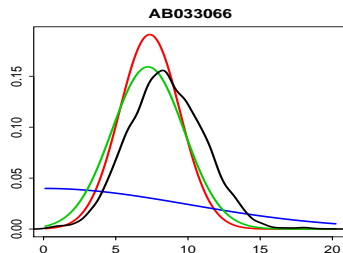
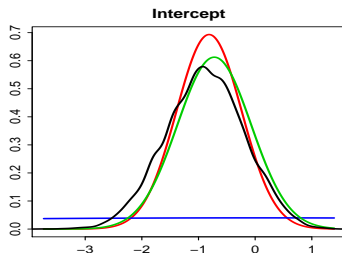
- Variational Bayes, expectation propagation, ...:

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

→ fast and reasonably accurate

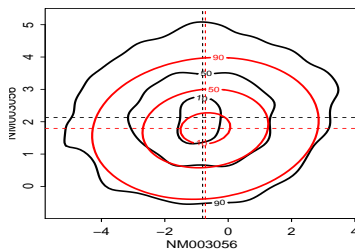
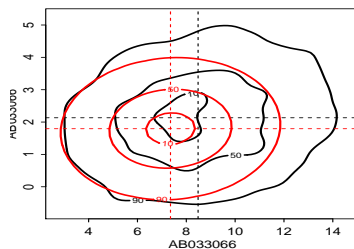
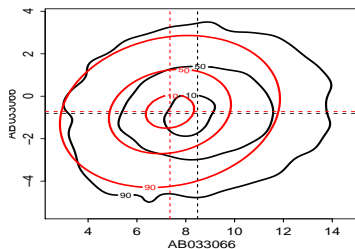
Variational Bayes & ML as a prior

— : 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 methods

Conjugate priors

Monte Carlo integration

Monte Carlo Markov chains (MCMC)

Extensions

Sequential Monte-Carlo (SMC)

Approximate Bayesian computation (ABC)

Limit distribution of Markov chain

Definition. Let $\{\phi^b\}_{b \geq 0}$ be a Markov chain with

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

$\{\phi^b\}_{b \geq 0}$ is said to be *ergodic* if

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

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

- ▶ for **any initial distribution** ν , ϕ^b converges towards μ in distribution

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

Ergodicity conditions.

- ▶ Finite state space: irreducibility, aperiodicity
- ▶ Infinite state space: the same + positive recurrence

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 $\lambda(\cdot | \theta)$

Metropolis-Hastings

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

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

Metropolis-Hastings

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

- ▶ Start with θ^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 $\lambda(\cdot | \theta)$

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

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

Metropolis-Hastings

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

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

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

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

Metropolis-Hastings

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

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

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

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

Properties.

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

Metropolis-Hastings for logistic regression

Model.

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

Metropolis-Hastings for logistic regression

Model.

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

Algorithm settings.

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

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);
```

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.prop = 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.prop = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
  logprior.prop = dmvnorm(theta.prop, 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.prop = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
  logprior.prop = dmvnorm(theta.prop, mean=mu.prior, sigma=Sigma.prior, log=T)
  prob.prop = plogis(X%*%theta.prop)
```

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.prop = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
  logprior.prop = dmvnorm(theta.prop, mean=mu.prior, sigma=Sigma.prior, log=T)
  prob.prop = plogis(X%*%theta.prop)
  loglik.prop = sum(dbinom(Y, 1, prob.prop, 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.prop = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
  logprior.prop = dmvnorm(theta.prop, mean=mu.prior, sigma=Sigma.prior, log=T)
  prob.prop = plogis(X%*%theta.prop)
  loglik.prop = sum(dbinom(Y, 1, prob.prop, log=T))

  alpha = exp(logprior.prop + loglik.prop - 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.prop = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
  logprior.prop = dmvnorm(theta.prop, mean=mu.prior, sigma=Sigma.prior, log=T)
  prob.prop = plogis(X%*%theta.prop)
  loglik.prop = sum(dbinom(Y, 1, prob.prop, log=T))

  alpha = exp(logprior.prop + loglik.prop - logprior.cur - loglik.cur)
  if(runif(1) < alpha){
    theta.sample[b, ] = theta.cur = theta.prop
    logprior.cur = logprior.prop; loglik.cur = loglik.prop
  }
}
```

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.prop = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
  logprior.prop = dmvnorm(theta.prop, mean=mu.prior, sigma=Sigma.prior, log=T)
  prob.prop = plogis(X%*%theta.prop)
  loglik.prop = sum(dbinom(Y, 1, prob.prop, log=T))

  alpha = exp(logprior.prop + loglik.prop - logprior.cur - loglik.cur)
  if(runif(1) < alpha){
    theta.sample[b, ] = theta.cur = theta.prop
    logprior.cur = logprior.prop; loglik.cur = loglik.prop
  }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.prop = rmvnorm(1, mean=theta.sample[b-1, ], sigma=Sigma.shift)[1, ]
  logprior.prop = dmvnorm(theta.prop, mean=mu.prior, sigma=Sigma.prior, log=T)
  prob.prop = plogis(X%*%theta.prop)
  loglik.prop = sum(dbinom(Y, 1, prob.prop, log=T))

  alpha = exp(logprior.prop + loglik.prop - logprior.cur - loglik.cur)
  if(runif(1) < alpha){
    theta.sample[b, ] = theta.cur = theta.prop
    logprior.cur = logprior.prop; loglik.cur = loglik.prop
  }else{
    theta.sample[b, ] = theta.sample[b-1, ]
  }
}
```

Sanity checks

Setting. Sample $1.2 \cdot 10^5$ θ , remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

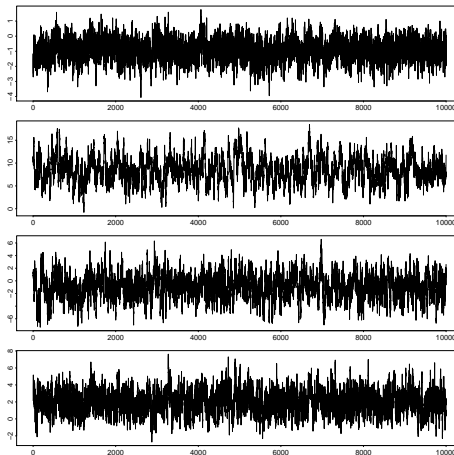
► Acceptance rate

variance shift	0.1	0.5	1
acceptance rate	0.421	0.125	0.053

Sanity checks

Setting. Sample $1.2 \cdot 10^5$ θ , remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

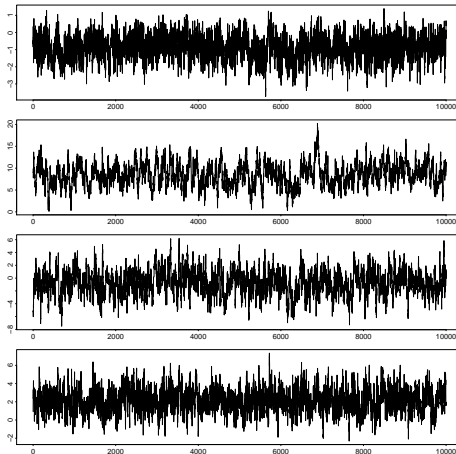
- Acceptance rate
- Stationarity:
var. shift = .1



Sanity checks

Setting. Sample $1.2 \cdot 10^5 \theta$, remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

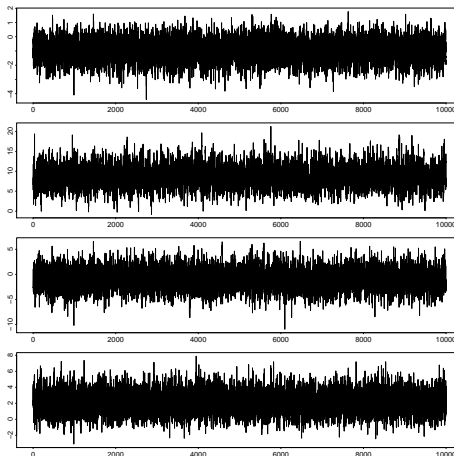
- Acceptance rate
- Stationarity:
var. shift = .1 , .5



Sanity checks

Setting. Sample $1.2 \cdot 10^5 \theta$, remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

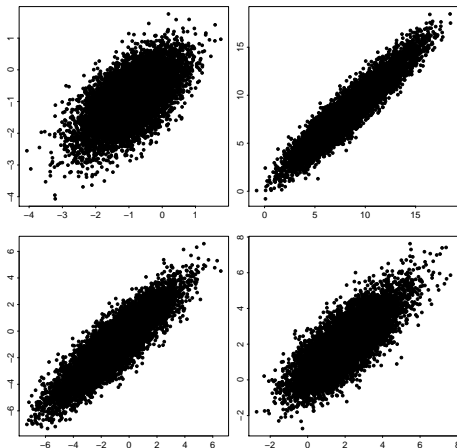
- Acceptance rate
- Stationarity:
var. shift = .1 , .5 , 1



Sanity checks

Setting. Sample $1.2 \cdot 10^5$ θ , remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

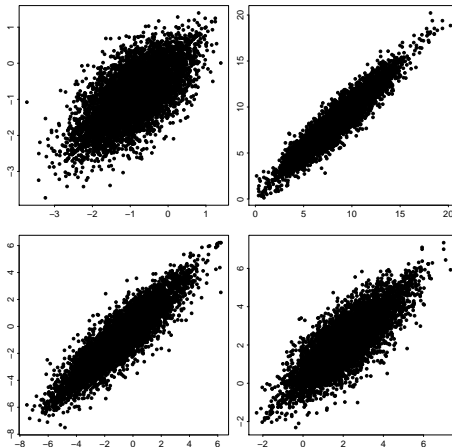
- ▶ Acceptance rate
- ▶ Stationarity:
var. shift = .1 , .5 , 1
- ▶ Autocorrelation
 $\text{Cor}(\theta_j^{b-1}, \theta_j^b)$:
var. shift = .1



Sanity checks

Setting. Sample $1.2 \cdot 10^5$ θ , remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

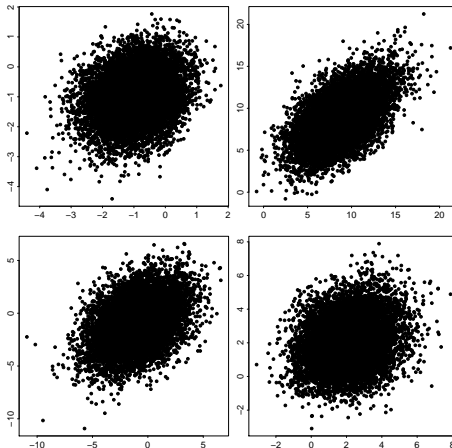
- ▶ Acceptance rate
- ▶ Stationarity:
var. shift = .1 , .5 , 1
- ▶ Autocorrelation
 $\text{Cor}(\theta_j^{b-1}, \theta_j^b)$:
var. shift = .1 , .5



Sanity checks

Setting. Sample $1.2 \cdot 10^5$ θ , remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

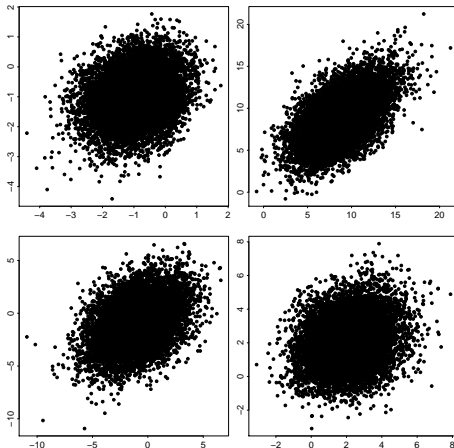
- Acceptance rate
- Stationarity:
var. shift = .1 , .5 , 1
- Autocorrelation
 $\text{Cor}(\theta_j^{b-1}, \theta_j^b)$:
var. shift = .1 , .5 , 1



Sanity checks

Setting. Sample $1.2 \cdot 10^5$ θ , remove first $2 \cdot 10^4$, extract every 10 $\rightarrow B = 10^4$.

- ▶ Acceptance rate
- ▶ Stationarity:
var. shift = .1 , .5 , 1
- ▶ Autocorrelation
 $\text{Cor}(\theta_j^{b-1}, \theta_j^b)$:
var. shift = .1 , .5 , 1
- ▶ And many others
(e.g. Gelman-Rubin)



Gibbs

Framework. We do not know how to sample the whole vector θ :

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

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

$$p(\theta_j \mid \mathbf{Y}, \theta_{-j})$$

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

Gibbs

Framework. We do not know how to sample the whole vector θ :

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

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

$$p(\theta_j \mid \mathbf{Y}, \theta_{-j})$$

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

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

- ▶ Start with θ^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 :

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

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 :

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

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 methods

- 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 methods

- 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
- ▶ $\theta = (\Pi, \gamma)$: transition matrix and emission probabilities

Inference. Need to sample from

- ▶ $p(\theta | \mathbf{Y})$ (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
- ▶ $\theta = (\Pi, \gamma)$: transition matrix and emission probabilities

Inference. Need to sample from

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

Sequential Monte Carlo.

- ▶ Monte Carlo (stochastic) counterpart of the forward-backward recurrence
- ▶ Sequentially sample from $p(Z_t | \mathbf{Y}_1^t, \mathbf{Z}_1^{t-1})$, then 'backtrack'.

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 methods

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:

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

- ▶ but we know how to sample from it

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

→ 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:

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

- ▶ but we know how to sample from it

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

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

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

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

Approximate Bayesian computation (ABC)

Ingredients.

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

Approximate Bayesian computation (ABC)

Ingredients.

- ▶ A set a *summary statistics* $\mathbf{s}(\mathbf{Y})$
- ▶ 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 $\theta' \sim \pi(\theta)$ (from the prior)
 2. sample $\mathbf{Y}' \sim \ell(\mathbf{Y} | \theta')$ (from the model)
 3. compute $\mathbf{s}' = \mathbf{s}(\mathbf{Y}')$
 4. if $d(\mathbf{s}' - \mathbf{s}_{\text{obs}}) < \varepsilon$, keep θ' in the sample

Approximate Bayesian computation (ABC)

Ingredients.

- ▶ A set a *summary statistics* $\mathbf{s}(\mathbf{Y})$
- ▶ 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 $\theta' \sim \pi(\theta)$ (from the prior)
 2. sample $\mathbf{Y}' \sim \ell(\mathbf{Y} | \theta')$ (from the model)
 3. compute $\mathbf{s}' = \mathbf{s}(\mathbf{Y}')$
 4. if $d(\mathbf{s}' - \mathbf{s}_{\text{obs}}) < \varepsilon$, keep θ' in the sample

Rational. Does not sample from $p(\theta | \mathbf{Y})$ but from

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

→ Related to kernel density estimate.

References

-  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.

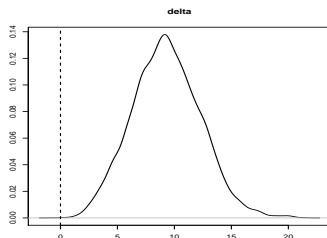
Outline

Appendix

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	9.289765	9.171537	3.614869	15.38288

Monte Carlo: Illustration (1/2)

Properties.

- Easy to implement

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


Monte Carlo: Illustration (1/2)

Properties.

- Easy to implement

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

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

Monte Carlo: Illustration (1/2)

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 (1/2)

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 (2/2)

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

	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	—

