

Lecture 1

Bayesian inference

olivier.francois@imag.fr

April 2011

Outline of Lecture 1

- ▶ Principles of Bayesian inference
- ▶ Classical inference problems (frequency, mean, variance)
- ▶ Basic simulation algorithms

What is Bayesian data analysis?

- ▶ **Model building.** Build a joint distribution for both observable quantities (**data**) and non-observable quantities (**parameters**).
- ▶ **Parameter inference.** Compute the conditional distributions of the non-observable quantities given the data.
- ▶ **Model criticism and improvement.** Evaluate the fit of the model to the data and check their predictions.

Model definition.

- ▶ Parameter $\theta = (\theta_1, \dots, \theta_J)$, $J \geq 1$.
- ▶ Data $y = (y_1, \dots, y_n)$, $n \geq 1$.
- ▶ A model is a joint distribution

$$p(y, \theta) = p(y|\theta)p(\theta)$$

- ▶ $p(\theta)$ is the **prior** distribution.
- ▶ $p(y|\theta)$ is the **likelihood** or sampling distribution.

Inference.

- Use the Bayes formula to compute the **posterior distribution**

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

where $p(y) = \int p(y|\theta)p(\theta)d\theta$ is the **marginal** distribution.

- The marginal distribution is usually a highly dimensional impossible to compute integral, and we write

$$p(\theta|y) \propto p(y|\theta)p(\theta).$$

Prediction.

- ▶ The **posterior predictive** distribution is

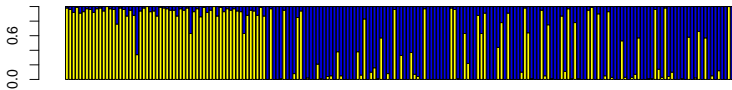
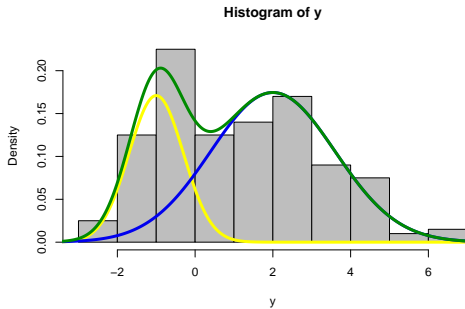
$$p(y_{\text{rep}}|y) = \int p(y_{\text{rep}}|\theta)p(\theta|y)d\theta.$$

- ▶ Models are **wrong**, and the posterior predictive distribution can be used to evaluate aspects of the model that do not fit to the data (**model checking**).

Examples of application (in this course)

- ▶ **Bayesian clustering**: How many groups in the data?
- ▶ What are the within-group means and variances?
- ▶ For a given individual, what is the assignment probability?
- ▶ **Population genetics**: For an individual genome, what fraction of DNA can be assigned to putative source (ancestral) populations?

Mixture models



Example 1: Inferring allele frequencies

- ▶ Natural populations are of finite size, N .
- ▶ New genetic variants can arise from mutation or migration
- ▶ Genes frequencies at a bi-allelic locus (ancestral/derived allele) can fluctuate

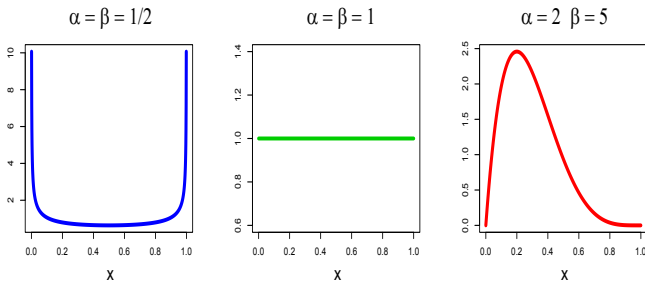
$$\frac{\#\{\text{carriers of the derived allele}\}}{N} \rightarrow \text{beta}(\alpha, \beta)$$

where the **beta distribution** is

$$\text{beta}(x, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad x \in (0, 1)$$

and $\alpha, \beta > 0$ depend on mutation or migration rates.

Beta distribution



- Expectation and mode of the beta distribution

$$E[X] = \frac{\alpha}{\alpha + \beta} \quad \text{Mode}(X) = \frac{\alpha - 1}{\alpha + \beta - 2}$$

Model

- ▶ **Prior distribution** on the allele frequency: $\theta \sim \text{beta}(1,1)$ (uniform).
- ▶ **Data**: We observe the derived allele $y = 9$ times in a sample of size $n = 20$ genes (frequency = .45)
- ▶ **Likelihood**

$$p(y|\theta) = \text{binom}(n, \theta)(y) \propto \theta^y (1 - \theta)^{n-y}$$

- ▶ **Posterior distribution** (Exercise)

$$p(\theta|y) = \text{beta}(y + 1, n + 1 - y)(\theta)$$

Remarks

- ▶ Point estimate (conditional mean) different from the maximum likelihood estimate

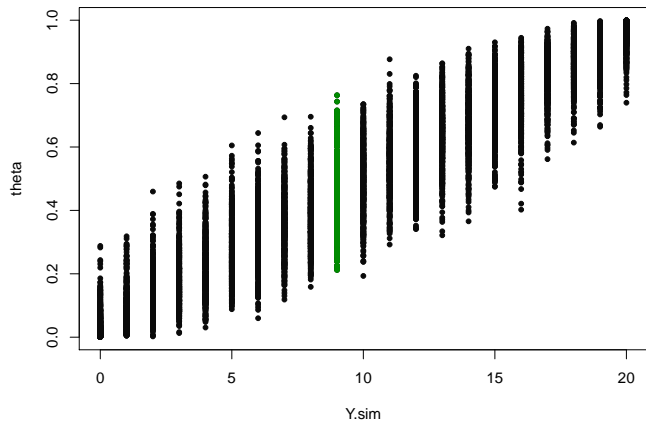
$$E[\theta|y] = \frac{y+1}{n+2} \approx \frac{y}{n}, \quad \text{as } n \rightarrow \infty$$

- ▶ Credible interval I so that $\Pr(\theta \in I|y) = .95$ (R command `quantile`)

$$I = (0.25, 0.65)$$

Not a confidence interval!

Joint distribution



Computing the posterior distribution from simulations

- ▶ Rejection algorithm

```
Repeat  
theta <- unif(0,1)  
y.s <- binom(n,theta)  
Until (y.s == y)  
return(theta)
```

- ▶ It generates samples from the posterior distribution $p(\theta|y)$ (Exercise).

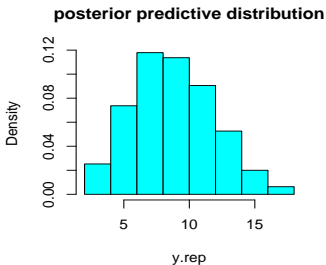
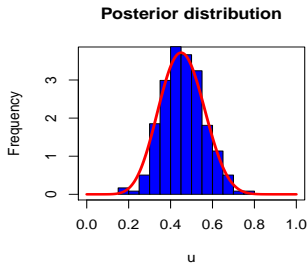
R scripts

- Rejection in R (sample of random size)

```
y = 9 ; n = 20  
theta <- runif(10000)  
y.s <- rbinom(10000, n, theta)  
theta.post <- theta[ y.s == y ]
```

- **Exercise:** Compute a 95% credible interval for θ and a histogram of the posterior predictive distribution given y .

Is the rejection algorithm efficient?



- The acceptance rate is only $\approx 4.5\%$. It leaves room for improvement (Lecture 2).

Gaussian model: $\theta = m$ (σ^2 known)

- ▶ Case 1: One-dimensional data: $y \in \mathbb{R}$
- ▶ Prior distribution

$$p(\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta - m_0)^2\right), \quad \beta_0 = \frac{1}{\sigma_0^2}$$

- ▶ Sampling distribution

$$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y - \theta)^2\right), \quad \beta = \frac{1}{\sigma^2}$$

- ▶ Posterior distribution (Exercise)

$$\theta|y \sim N(m_1, \sigma_1^2)$$

with $1/\sigma_1^2 = \beta_1 = \beta_0 + \beta$, and $m_1 = (\beta_0 m_0 + \beta y)/\beta_1$.

Gaussian model: $\theta = m$ (σ^2 known)

- ▶ Non-informative prior distribution

$$p(\theta) \propto 1, \quad \beta_0 \rightarrow 0 \quad (\sigma_0^2 = \infty).$$

- ▶ Posterior distribution

$$\theta|y \sim N(y, \sigma^2)$$

- ▶ Exercise: Posterior predictive distribution

$$\tilde{y}|y \sim N(m_1, \sigma^2 + \sigma_1^2) = N(m_1, 2\sigma^2)$$

Gaussian model: $\theta = m$ (σ^2 known)

- ▶ Case 2: n data, $y = (y_1, \dots, y_n)$
- ▶ Sampling distribution

$$p(y_1, \dots, y_n | \theta) \propto \prod_{i=1}^n \exp \left(-\frac{1}{2\sigma^2} (y_i - \theta)^2 \right),$$

- ▶ $\bar{y} = \sum_{i=1}^n y_i / n$ is sufficient

$$p(\theta | y) = p(\theta | \bar{y})$$

- ▶ Posterior distribution (Uninformative prior)

$$\theta | y \sim N(\bar{y}, \sigma^2 / n).$$

Gaussian model: $\theta = \sigma^2$ (m known)

- ▶ χ_n^2 distribution

$$p(x) \propto x^{n/2-1} e^{-x/2}, \quad x > 0$$

- ▶ $\text{Inv}\chi^2(\nu, s^2)$ distribution: $X = \frac{\nu s^2}{\chi_\nu^2}$ (Exercise)

$$p(x) \propto \frac{1}{x^{\nu/2+1}} e^{-\frac{\nu s^2}{2x}}, \quad x > 0.$$

Gaussian model: $\theta = \sigma^2$ (m known)

- Prior distribution (not a density)

$$p(\theta) \propto \frac{1}{\theta}, \quad p(\log(\theta)) \propto 1.$$

- Sampling distribution

$$p(y_1, \dots, y_n | \theta) \propto \frac{1}{\theta^{n/2}} \exp\left(-\frac{n}{2\theta} s_n^2\right)$$

where

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (y_i - m)^2$$

- Posterior distribution (Exercise)

$$\sigma^2 | y \sim \text{Inv}\chi^2(n, s_n^2)$$

Joint inference $\theta = (m, \sigma^2)$

- Prior distribution (not a density)

$$p(m, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- Posterior distribution

$$p(m, \sigma^2 | y) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp \left(-\frac{1}{2\sigma^2} ((n-1)s_{n-1}^2 + n(\bar{y} - m)^2) \right)$$

where the unbiased empirical variance is

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

- The (marginal) posterior distribution of σ^2 is (exercise)

$$\sigma^2 | y \sim \text{Inv}\chi^2(n-1, s_{n-1}^2)$$

Multi-dimensional parameters: A basic algorithm

- ▶ $\theta = (\theta_1, \theta_2)$
- 1. Simulate θ_1 from the marginal distribution $p(\theta_1)$
- 2. Given θ_1 , simulate θ_2 from the conditional distribution $p(\theta_2|\theta_1)$.

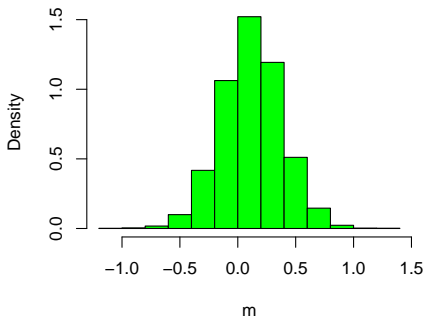
Simulation $\theta = (m, \sigma^2)$

1. $\sigma^2|y \sim (n-1)\text{var}(y)/\chi_{n-1}^2$
2. $m|\sigma^2, y \sim N(\text{mean}(y), \sigma^2/n)$

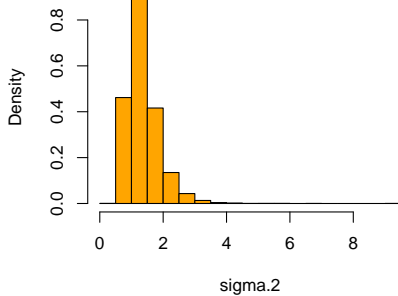
```
# simulated data
n = 20; y = rnorm(n)
# Posterior distribution sampling
sigma.2 = (n-1)*var(y)/rchisq(10000, n-1)
m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))
```


Posterior distribution (Gaussian model)

Histogram of m



Histogram of sigma.2



Model checking

- Our data are perhaps not from a Gaussian model

Example

```
y = rcauchy(n)
```

```
sigma.2 = (n-1)*var(y)/rchisq(10000, n-1)
```

```
m = rnorm(10000, mean(y), sd = sqrt(sigma.2/n))
```

- Use a test statistic (skewness)

```
post.pred = NULL
```

```
for (i in 1:1000) {
```

```
  ind = sample(10000, 1)
```

```
  post.pred[i] = skewness(rnorm(20, m[ind],
```

```
    sqrt(sigma.2[ind]))) }
```

```
hist(post.pred)
```

```
skewness(y)
```

Take-home messages

- ▶ Bayesian inference is about computing the conditional distribution of a parameter given the data.
- ▶ This can be achieved by using computational Monte Carlo methods
- ▶ More to come in lecture 2.

Exercises

- Ex1. Find the posterior distribution in the beta-binomial model (answer: $\text{beta}(y + \alpha, n + \beta - y)$).
- Ex2. Prove the rejection algorithm.
- Ex3. Compute the 95% credible interval for θ and the posterior predictive distribution given y from the rejection algorithm
- Ex4. Simulate from the posterior distribution in the Gaussian model (two parameters). Use your own statistic for model checking.
- Ex5. Run inference for the sepal length in `data(iris)`

Bibliography and resources

- ▶ Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis 2nd ed. Chapman & Hall, New-York.
- ▶ E. Paradis (2005) R pour les débutants. Univ. Montpellier II.
- ▶ R website: <http://cran.r-project.org/>