Fisheries Population Analysis

Lecture 6 "Bayesian methods”

Toshihide Kitakado

Likelihood estimation

1. Mathematically describe the statistical nature of the data you have
2. Define a likelihood function of parameters of interest
3. Maximize the log-likelihood function with respect to the unknown parameters
4. Evaluate the standard errors of the parameter estimates

“A” framework of statistical inference

1. **Modeling**
2. **Inference**
   - point estimation
   - standard errors
   - confidence regions
   - hypothesis testing
3. **Model selection and diagnosis**

“A” framework of statistical inference

1. **Modeling** (Defining a likelihood function)
2. **Inference**
   - point estimation (Maximizing the likelihood)
   - standard errors (Fisher information matrix)
   - confidence regions (Likelihood profile etc.)
   - hypothesis testing (LRT)
3. **Model selection and diagnosis**
   (AIC and deviance)

“Maximum likelihood (ML) methods” suits well with this framework

Likelihood inference

Many good properites

★ Consistency, Asymptotic normality

★ Profile, LRT, Fisher information, AIC….

Sometimes biased

★Separate inference, Integrated likelihood etc

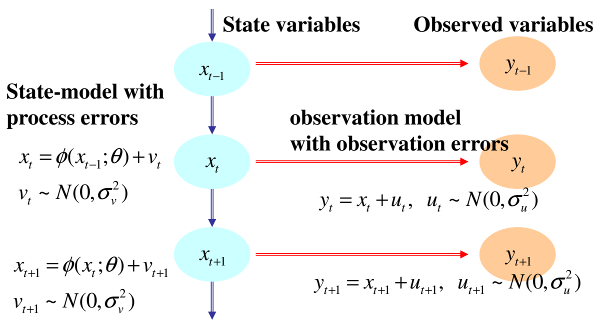
Complex structures (recently often)

★ random (mixed) effects

★ latent, hidden, missing variables

★ hierarchical structures (state-space, smoothing)

State-space model



Algorithm for state-space modelling for MLE

Algorithms, approximations

- EM algorithm
- Laplace approximation

Numerical approximation

- MCEM algorithm
- Relative likelihood
- Non-informative priors
  (but not actual MLE but integrated one)

etc

Bayesian Analysis

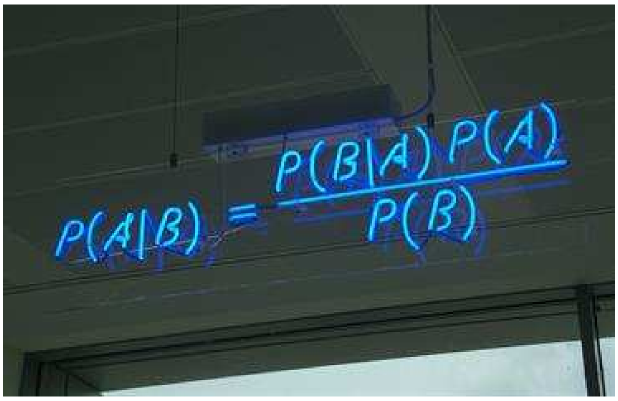
- Likelihood (given data) + prior information

$p(y|\theta)$

$p(\theta)$ $\Rightarrow$ posterior $$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta)\, d\theta}$$

- Of course, an inappropriate prior causes worse inference, but large sample can thin down it
- Often non-informative priors are employed, but the definition is ambiguous in cases of multivariate parameters
- Convenient framework for complex and hierarchical models especially in analyses where decision is necessary

Bayes theorem in a neon sign!



Binomial distribution

$$P(Y = y \mid p) = \binom{N}{y} p^{y} (1-p)^{N-y}$$

$$\pi(p) = \frac{1}{Be(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad where \quad Be(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\pi(p \mid y) = \frac{P(Y = y \mid p)\pi(p)}{P(Y = y)} = \frac{1}{Be(y+\alpha, N-y+\beta)} p^{y+\alpha-1}(1-p)^{N-y+\beta-1}$$

Posterior mean of parameter

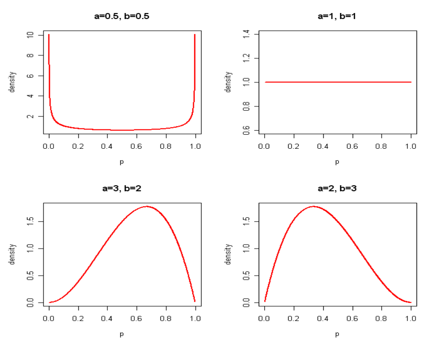
$$E[\,p \mid y\,] = \frac{y+\alpha}{N+\alpha+\beta} = \frac{N}{N+\alpha+\beta} \cdot \boxed{\frac{y}{N}} + \frac{\alpha+\beta}{N+\alpha+\beta} \cdot \boxed{\frac{\alpha}{\alpha+\beta}}$$

Smaller N is, larger the weight for the prior mean is.

When N is getting large, the effect of N is diminishing.

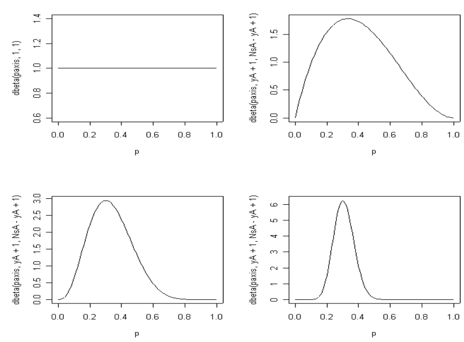
Prior distribution for binomial "p"

$$\pi(p) = \frac{1}{Be(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

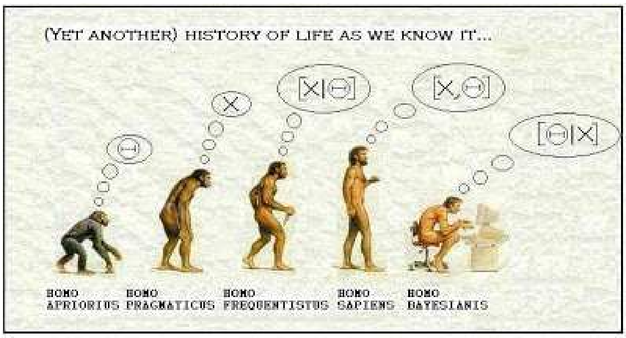


Posterior distribution for binomial "p"

$$\pi(p \mid y) = \frac{1}{Be(y+\alpha, N-y+\beta)} p^{y+\alpha-1}(1-p)^{N-y+\beta-1}$$



Bayesian method in a cartoon



I forgot where I downloaded this…

Bayesian inference: posterior is a key

Simple Bayes

$p(y|\theta)$

$p(\theta)$ ⇨ **posterior** $p(\theta \mid y) = \dfrac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta)\, d\theta}$

Empirical Bayes

$p(y|\theta)$

$p(\theta \mid \phi)$ ⇨ **posterior** $p(\theta \mid y, \hat{\phi}) = \dfrac{p(y \mid \theta)\, p(\theta \mid \hat{\phi})}{\int p(y \mid \theta)\, p(\theta \mid \hat{\phi})\, d\theta}$

$\hat{\phi} = \arg\max \int p(y \mid \theta)\, p(\theta \mid \phi)\, d\theta$

Hierarchical Bayes

$p(y|\theta)$

$p(\theta \mid \phi)$ ⇨ **posterior** $p(\theta, \phi \mid y) = \dfrac{p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi)}{\int p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi)\, d\theta d\phi}$

$p(\phi)$

Posterior and the mean needs integration

$p(y|\theta)$

$p(\theta)$ ⇨ $p(\theta \mid y) = \dfrac{p(y \mid \theta)\, p(\theta)}{\int p(y \mid \theta)\, p(\theta)\, d\theta}$ **Integration!**

⇨ $E[f(\theta) \mid y] = \int f(\theta)\, p(\theta \mid y)\, d\theta$

Analytical integration is quite often difficult for complex and hierarchical models -> numerical evaluation.

Importance sampling (IS), Sampling importance resampling (SIR)
- simulate values from a tractable distribution (hopefully similar to posterior, and adjust them by weighting (or weighted sampling)

Markov chain Monte Carlo (MCMC)
- simulate values from a Markov chain with the posterior as the equilibrium distribution

Posterior and the mean needs integration

Importance sampling (IS), Sampling importance sampling (SIR)

$$E[f(\theta) \mid y] = \int f(\theta)\, p(\theta \mid y)\, d\theta$$
$$= \int \frac{f(\theta)\, p(\theta \mid y)}{g(\theta)}\, g(\theta) d\theta \approx \frac{1}{n}\sum_{i=1}^{n} \frac{f(\theta_i)\, p(\theta_i \mid y)}{g(\theta_i)}$$

θ_i *from* $g(\theta_i)$

Markov chain Monte Carlo (MCMC)

$\theta_{i+1} | \theta_i$ *from* $g(\theta_{i+1}|\theta_i)$ ⇨ $\underbrace{\theta_1, \ldots, \theta_m}_{\text{Burn-in period (discarded)}}, \underbrace{\theta_{m+1}, \ldots, \theta_{m+n}}_{\text{Used for evaluation of posterior after thinning}}$

$$E[f(\theta) \mid y] = \int f(\theta)\, p(\theta \mid y)\, d\theta \approx \frac{1}{n}\sum_{i=1}^{n} f(\theta_i)$$

Metropolis-Hasting (MH) method

Target: generate random variables from p(θ|y)

1. $\theta^{(i)}$ **: i-th outcome**
2. $\theta^{*} \sim g(\bullet \mid \theta^{(i)})$ **(g is called a proposal distribution)**
3. **Acceptance ratio for** θ^*

$$\alpha = \min\left(\frac{p(\theta^* \mid y)\, g(\theta^{(i)} \mid \theta^*)}{p(\theta^{(i)} \mid y)\, g(\theta^* \mid \theta^{(i)})}, 1\right)$$

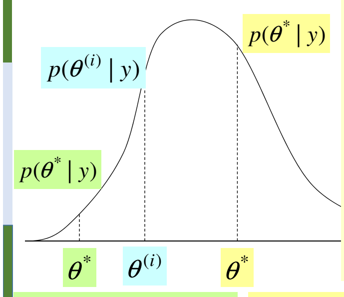
4. θ^* **is accepted with probability α**

$$\theta^{(i+1)} = \begin{cases} \theta^* & with\ \ prob\ \ \alpha \\ \theta^{(i)} & with\ \ prob\ \ 1-\alpha \end{cases}$$

If $g(\theta_1 \mid \theta_2) = g(\theta_2 \mid \theta_1)$**, the algorithm called Metropolis method**

$$\alpha = \min\left(\frac{p(\theta^* \mid y)}{p(\theta^{(i)} \mid y)}, 1\right)$$

Intuitive understanding for Metropolis



Optimization: planned journey
Search maximizer of target function, and stop if it is found

MCMC: wandering journey
- **Possible to move low probability values**
- **But, stay in high probability areas**
- **By traversing, capture the target distribution**

$\alpha = \min\left(\dfrac{p(\theta^* \mid y)}{p(\theta^{(i)} \mid y)}, 1\right) << 1$

$\alpha = \min\left(\dfrac{p(\theta^* \mid y)}{p(\theta^{(i)} \mid y)}, 1\right) = 1$

Accept! $\theta^{(i+1)} = \theta^*$

We don't need to evaluate any integral!

$$\alpha = \min\left(\frac{p(\theta^* \mid y)\, g(\theta^{(i)} \mid \theta^*)}{p(\theta^{(i)} \mid y)\, g(\theta^* \mid \theta^{(i)})}, 1\right)$$

$$\frac{p(\theta^* \mid y)}{p(\theta^{(i)} \mid y)} = \frac{p(y \mid \theta^*)\, p(\theta^*) / p(y)}{p(y \mid \theta^{(i)})\, p(\theta^{(i)}) / p(y)}$$

Just drawing sample from g(θ|θ$^{(i)}$), and evaluate the acceptance ratio based on likelihood, prior and sampler

Generation of random number for N(0,1) by Metropolis method

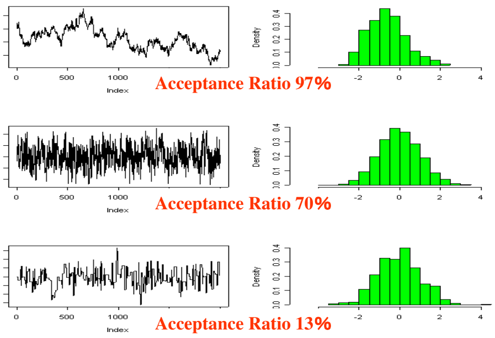
Proposal : $\theta^* \mid \theta^{(i)} \sim N(\theta^{(i)}, \sigma^2)$

Histogram based on samples in 1001~2000 iteratons

$\sigma = 0.1$ Acceptance Ratio 97%

$\sigma = 1$ Acceptance Ratio 70%

$\sigma = 10$ Acceptance Ratio 13%



Single component MH for cases of two parameters

Aim: Generate random variables from $p(\theta_1, \theta_2 \mid y)$

1. $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)})$: *i*-th outcome
2. From a proposal $Q(\bullet \mid \theta_1^{(i)}, \theta_2^{(i)})$, generate θ_1^*
3. θ_1^* is accepted with a probability α_1

$$\alpha_1 = \min\left(\frac{p(\theta_1^* \mid \theta_2^{(i)}, y)\, Q(\theta_1^{(i)} \mid \theta_1^*, \theta_2^{(i)})}{p(\theta_1^{(i)} \mid \theta_2^{(i)}, y)\, Q(\theta_1^* \mid \theta_1^{(i)}, \theta_2^{(i)})}, 1\right)$$

4. From a proposal $Q(\bullet \mid \theta_1^{(i+1)}, \theta_2^{(i)})$, generat θ_2^*
5. θ_2^* is accepted with a probability θ_2^*

$$\alpha_2 = \min\left(\frac{p(\theta_2^* \mid \theta_1^{(i+1)}, y)\, Q(\theta_1^{(i)} \mid \theta_1^{(i+1)}, \theta_2^*)}{p(\theta_2^{(i)} \mid \theta_1^{(i+1)}, y)\, Q(\theta_2^* \mid \theta_1^{(i+1)}, \theta_2^{(i)})}, 1\right)$$

Gibbs sampling (A special case of MH)

In single component MH, conditional distributions are used as proposal distributions

$$Q(\theta_1^* \mid \theta_1^{(i)}, \theta_2^{(i)}) = p(\theta_1^{(i)} \mid \theta_2^{(i)}, y)$$
$$Q(\theta_2^* \mid \theta_1^{(i+1)}, \theta_2^{(i)}) = p(\theta_2^{(i)} \mid \theta_1^{(i+1)}, y)$$

$$\alpha_1 = \min\left(\frac{p(\theta_1^* \mid \theta_2^{(i)}, y)\, Q(\theta_1^{(i)} \mid \theta_1^*, \theta_2^{(i)})}{p(\theta_1^{(i)} \mid \theta_2^{(i)}, y)\, Q(\theta_1^* \mid \theta_1^{(i)}, \theta_2^{(i)})}, 1\right)$$
$$= \min\left(\frac{p(\theta_1^* \mid \theta_2^{(i)}, y)\, p(\theta_1^{(i)} \mid \theta_2^{(i)}, y)}{p(\theta_1^{(i)} \mid \theta_2^{(i)}, y)\, p(\theta_1^* \mid \theta_2^{(i)}, y)}, 1\right) = 1$$
$$\alpha_2 = 1$$

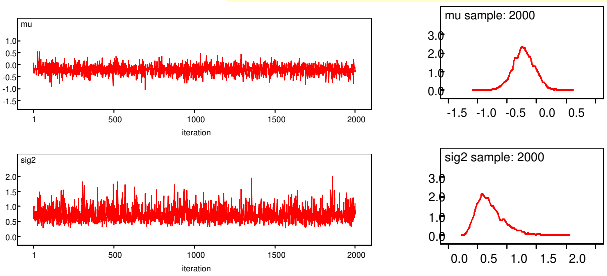
single-component MH with acceptance ratio 1

Sometimes, hybridization of MH and Gibbs sampling are used for getting samples from conditional distributions (MH withing Gibbs)

Ex: Estimation of parameter in N(μ,σ^2) using Gibbs sampler

$$y_1, \ldots, y_n \sim (iid)\ N(\mu, \sigma^2)$$
$$\mu \sim N(0, \gamma^2)$$
$$\sigma^2 \sim InvGa(\alpha, \beta)$$

$$\mu \mid \sigma^2, y \sim N\left(\frac{n\gamma^2}{n\gamma^2 + \sigma^2}\bar{y}, \frac{\sigma^2\gamma^2}{n\gamma^2 + \sigma^2}\right)$$
$$\sigma^2 \mid \mu, y \sim InvGa\left(\alpha + 0.5, \beta + \sum (y_i - \mu)^2 / 2\right)$$



A hierarchical model

$p(y \mid \theta)$

$p(\theta \mid \phi)$

$p(\phi \mid \lambda)$

$p(\lambda)$

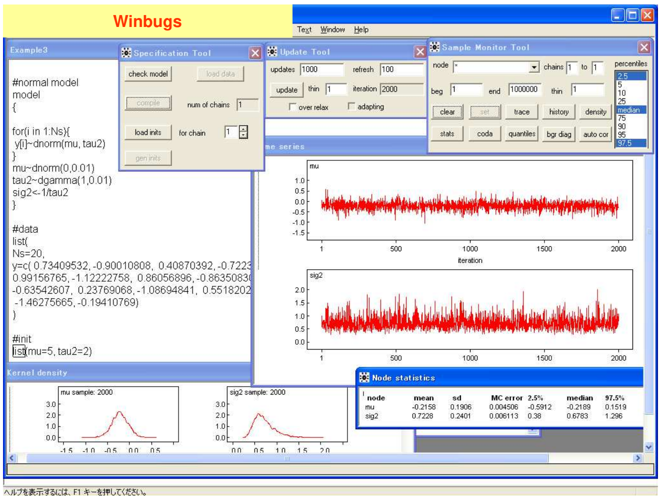
posterior

$$p(\theta, \phi, \lambda \mid y) = \frac{p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)}{\iiint p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)\, d\theta\, d\phi\, d\lambda}$$

$$p(\theta \mid y, \phi, \lambda) = \frac{p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)}{\int p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)\, d\theta} = \frac{p(y \mid \theta)\, p(\theta \mid \phi)}{\int p(y \mid \theta)\, p(\theta \mid \phi)\, d\theta}$$

$$p(\phi \mid y, \theta, \lambda) = \frac{p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)}{\int p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)\, d\phi} = \frac{p(\theta \mid \phi)\, p(\phi \mid \lambda)}{\int p(\theta \mid \phi)\, p(\phi \mid \lambda)\, d\phi}$$

$$p(\lambda \mid y, \theta, \phi) = \frac{p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)}{\int p(y \mid \theta)\, p(\theta \mid \phi)\, p(\phi \mid \lambda)\, p(\lambda)\, d\lambda} = \frac{p(\phi \mid \lambda)\, p(\lambda)}{\int p(\phi \mid \lambda)\, p(\lambda)\, d\lambda}$$



Model selection (Determination of the number of stocks)

1. Posterior probabilities

$\Pr(y \mid \theta, M=k)$

$\Pr(\theta \mid M=k)$

$\Pr(M=k)$

$$\Pr(M=k \mid y) = \frac{\Pr(y \mid M=k)\Pr(M=k)}{\sum_{k=1}^{K}\Pr(y \mid M=k)\Pr(M=k)}$$

Marginal likelihood $\Pr(y \mid M=k) = \int \Pr(y \mid \theta, M=k)\Pr(\theta \mid M=k)d\theta$

- Newton&Raftery(1994), Chib(1995) etc
- SIR is convenient in terms of techniques

2. Reversible jump (Green, 1995)

MCMC including model index

3. DIC (Deviance Information Criterion) Inplemented in Winbugs

posterior mean of deviance＋complexity of model

Bayesian method with MCMC

- Posterior distribution can be drawn by MCMC
  - possible to numerical integration
  - easily derive credible intervals
- Only likelihood functions and priors are essentially required (but needs better proposal)
- Handle complicated models with hierarchical structures
- Convergence diagnosis is tedious
- No magic bullet!
- Assess impact of prior distributions
- Winbugs / ADMB / Stan / JAGS