MCMC I
Course Time Plan
July 13-15, 2015

Instructors: Vladimir Minin, Kari Auranen, M. Elizabeth Halloran

Course Description: This module is an introduction to Markov chain Monte Carlo methods with some simple applications in infectious disease studies. The course includes an introduction to Bayesian inference, Monte Carlo, MCMC, some background theory, and convergence diagnostics. Algorithms include Gibbs sampling and Metropolis-Hastings and combinations. Programming is in R. Familiarity with the R statistical package or other computing language is needed.

Course schedule: The course is composed of 10 90-minute sessions, for a total of 15 hours of instruction.

1 Introduction to Bayesian Inference

- Overview of the course.
- Bayesian inference: Likelihood, prior, posterior, normalizing constant
- Conjugate priors; Beta-binomial; Poisson-gamma; normal-normal
- Posterior summaries, mean, mode, posterior intervals
- Motivating examples: Chain binomial model (Reed-Frost), General Epidemic Model, SIS model.
- Lab:
  - Goals: Warm-up with R for simple Bayesian computation
  - Example: Posterior distribution of transmission probability with a binomial sampling distribution using a conjugate beta prior distribution
  - Summarizing posterior inference (mean, median, posterior quantiles and intervals)
  - Varying the amount of prior information
  - Writing an R function

2 Introduction to Gibbs Sampling

- Chain binomial model and data augmentation
- Brief introduction to Gibbs sampling
- Lab
  - Goals: Simple data augmentation using MCMC
  - Example: Gibbs sampler for the chain binomial model.

3 Classical Monte Carlo and Markov chain theory

- Random number generators
- Non-iterative Monte Carlo methods
  - Classical Monte Carlo and importance sampling
- Basic Markov Chain Theory
  - Definitions
  - Stationarity
  - The ergodic theorem
- Lab:
  - Goals: Understanding importance sampling and the ergodic theorem for Markov chains.

4 Metropolis-Hastings algorithm

- Construction
- Proof of detailed balance
- Lab:
  - Goals: elementary missing data imputation on S-I model

5 Gibbs sampling

- Relationship with Metropolis-Hastings
- Revisit simple Gibbs sampler for chain-binomial model

6 Metropolis-Hasting and Gibbs combined

- Example: Hierarchical model
- Lab:
  - Goals: Combining Metropolis and Gibbs in one algorithm
  - Example: Beta-binomial hierarchical model with rat data

7 Chain binomial model revisited

- Hierarchical chain binomial model with hyperparameters
  - Model checking
  - Allowing for heterogeneity
- Lab:
  - Goals: Combined M-H and Gibbs and learning model checking
  - Example: Hierarchical beta-binomial chain binomial model

8 General Epidemic Model

- The general epidemic model and incompletely observed data
- Algorithm
- Lab: General epidemic model
  - Goals: parameter estimation with data augmentation
  - Example: smallpox transmission

9 Diagnostics, etc

- Assessing convergence (more or less), Coda
- Variance reduction, Monte Carlo error
- Poisson process
- Lab: Diagnostics
  - Goals: learn how to do basic diagnostics on chain and output
  - Coda
  - Diagnostics on previous examples

10 SIS model

- Binary Markov process model for a recurrent infection
- Likelihood
- Algorithm
- Lab: Estimating rates in simple SIS model

- Goals: Data simulation and parameter estimation from complete data in a simple SIS model.
- Example: simulate one long chain in one person
- Estimating rates from complete data
- Diagnostics

MCMC Methods for Infectious Diseases
Introduction

Vladimir Minin, Kari Auranen, M. (Betz) Elizabeth Halloran

SISMID 2015
University of Washington
Seattle, WA, USA

June 28, 2015

Outline Introduction Transmission Probability Simple Gibbs sampler

Prior, likelihood, and posterior

- Let
  - $y = (y_1, \ldots, y_n)$: observed data
  - $f(y|\theta)$: model for the observed data, usually a probability distribution
  - θ: vector of unknown parameters, assumed a random quantity
  - $\pi(\theta)$: prior distribution of θ
- The posterior distribution for inference concerning θ is

$$f(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|u)\pi(u)du}.$$

Posterior and marginal density of y

- The integral $\int f(y|u)\pi(u)du$, the marginal density of the data y, does not depend on θ.
- When the data y are fixed, then the integral can be regarded as a normalizing constant C.
- In high dimensional problems, the integral can be very difficult to evaluate.
- Evaluation of the complex integral $\int f(y|u)\pi(u)du$ was a focus of much Bayesian computation.

Advent of MCMC Methods

- With the advent of the use of Markov chain Monte Carlo (MCMC) methods,
  $\longrightarrow$ one could avoid evaluating the integral, making use of the unnormalized posterior density.

$$f(\theta|y) \propto f(y|\theta)\pi(\theta).$$

- Equivalently, if we denote the likelihood function or sampling distribution by $L(\theta)$, then

$$\begin{array}{rcl} f(\theta|y) & \propto & L(\theta)\pi(\theta). \\ \text{posterior} & \propto & \text{likelihood} \times \text{prior} \end{array}$$

- We will show how this works.

Other Uses of MCMC Methods

- Can simplify otherwise difficult computations.
- Sometimes a likelihood would be easy to evaluate if some data had been observed that was not observed or is unobservable.
- Examples:
  - infection times,
  - time of clearing infection,
  - when someone is infectious,
  - chains of infection.
- MCMC methods can be used to augment the observed data to make estimation simpler.

Likehood and Data Transforms Prior to Posterior

- Likelihood and data take prior to posterior:

$$\text{Prior} \xrightarrow[\begin{array}{c}\text{–Likelihood}\\ \text{–Data}\end{array}]{\text{Transformation}} \text{Posterior}$$

- Bayesian data analysis is a study of the transformation.

Outline Introduction Transmission Probability Simple Gibbs sampler

Transmission probability

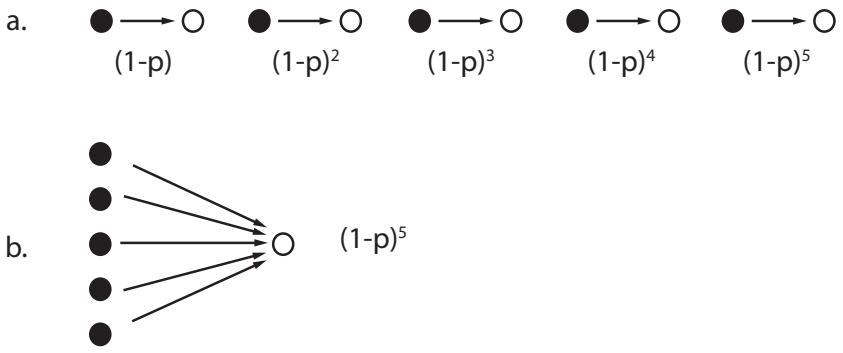
- p is the probability an infective infects a susceptible: transmission probability
- $q = 1 - p$ is the probability a susceptible escapes infection when exposed to an infective: escape probability
- Transmission versus escape ? which is the "success" and which the "failure"?
- Given there are n exposures, and y infections, what is the estimate of the transmission probability?
- Given there are n exposures, and $n - y$ escapes, what is the estimate of the escape probability?

Chain-binomial model

- Assume independent households
- One person in each household introduces the infection into the household (index case).
- Infections occur within households in generations of infection (discrete time).
- p is the probability an infective infects a susceptible in a household in a generation
- $q = 1 - p$ is the probability a susceptible escapes infection when exposed to an infective in a household

Reed-Frost Chain Binomial Model

Figure : Independent exposures = independent Bernoulli trials



Chain Binomial Model

Table : Chain binomial probabilities in the Reed-Frost model in N households of size 3 with 1 initial infective and 2 susceptibles, $S_0 = 2, I_0 = 1$

Chain	Chain probability	Frequency	at p=0.4	at p=0.7	Final number infected
$1 \longrightarrow 0$	q^2	n_1	0.360	0.090	1
$1 \longrightarrow 1 \longrightarrow 0$	$2pq^2$	n_{11}	0.288	0.126	2
$1 \longrightarrow 1 \longrightarrow 1$	$2p^2q$	n_{111}	0.192	0.294	3
$1 \longrightarrow 2$	p^2	n_{12}	0.160	0.490	3
Total	1	N	1.00	1.00	

Chain binomial model

- Data: The observations are based on outbreaks of measles in Rhode Island 1929–1934.
- The analysis is restricted to $N = 334$ families with three susceptible individuals at the outset of the epidemic.
- Assume there is a single index case that introduces infection into the family.
- The actual chains are not observed, just how many are infected at the end of the epidemic.
- So the frequency of chains $1 \longrightarrow 1 \longrightarrow 1$ and $1 \longrightarrow 2$ are not observed.
- MCMC can be used to augment the missing data, and estimate the transmission probability p.

Chain Binomial Model

Table : Rhodes Island measles data: chain binomial probabilities in the Reed-Frost model in $N = 334$ households of size 3 with 1 initial infective and 2 susceptibles, $N_3 = n_{111} + n_{12} = 275$ is observed

Chain	Chain probability	Frequency	Observed frequency	Final number infected
$1 \longrightarrow 0$	q^2	n_1	34	1
$1 \longrightarrow 1 \longrightarrow 0$	$2pq^2$	n_{11}	25	2
$1 \longrightarrow 1 \longrightarrow 1$	$2p^2q$	n_{111}	not observed	3
$1 \longrightarrow 2$	p^2	n_{12}	not observed	3
Total	1	N	334	

General epidemic (SIR) model

- The population of N individuals
- Denote the numbers of susceptible, infective, and removed individuals at time t by $S(t)$, $I(t)$, and $R(t)$.
- The process can be represented by the compartmental diagram

$$S(t) \longrightarrow I(t) \longrightarrow R(t)$$

- Thus, $S(t) + I(t) + R(t) = N$ for all t.
- Initially, $(S(0), I(0), R(0)) = (N - 1, 1, 0)$

General epidemic model

- Each infectious individual remains so for a length of time $T_I \sim \exp(\gamma)$.
- During this time, infectious contacts occur with each susceptible according to a Poisson process of rate β/N
- Thus, the overall hazard of infection at time t is $\beta I(t)/N$
- The two model parameters of interest are β and γ

General epidemic model

- In a well-known smallpox data set, the removal times are observed. That is, when the people are no longer infectious for others.
- However, the infection times are not observed.
- Thus, estimating the two model parameters is difficult.
- The missing infection times are treated as latent variables.
- MCMC methods are used to augment the missing infection times and estimate the parameters β and γ.

Susceptible-infected-susceptible (SIS) model

- Background: Many infections are recurrent, occurring as an alternating series of presence and absence of infection
- Nasopharyngeal carriage of *Streptococcus pneumoniae* (Auranen et al 2000; Cauchemez et al; Melegaro et al)
- Nasopharyngeal carriage of *Neisseria meningitidis* (Trotter and Gay 2003)
- Malaria (Nagelkerke et al,)
- multi-resistant *Staphylococcus aureus* (Cooper et al)

Susceptible-infected-susceptible (SIS) model

- The population of N individuals
- Denote the numbers of susceptible and infected individuals at time t by $S(t)$ and $I(t)$.
- The process can be represented by the compartmental diagram

$$S(t) \leftrightarrow I(t)$$

- Thus, $S(t) + I(t) = N$ for all t.
- Acquisition and clearance times often remain unobserved
- Active sampling of the population to determine the current status of being infected or susceptible in individuals.

Susceptible-infected-susceptible (SIS) model

- Could be formulated as an infectious disease transmission process, as the general epidemic model.
- Too complicated for this introductory course
- We consider here the simple transition process, with rate parameters λ for acquisition and μ for clearance.
- The acquisition and clearance times are treated as latent variables.
- MCMC methods are used to augment the missing infection and clearance times, and estimate the parameters λ and μ.

Conjugate prior distribtions

- Conjugacy: the property that the posterior distribution follows that same parametric form as the prior distribution.
- Beta prior distribution is conjugate family for binomial likelihood: posterior distribution is Beta
- Gamma prior distribution is conjugate family for Poisson likelihood: posterior distribution is Gamma

Conjugate prior distributions

- Simply put, conjugate prior distributions in tandem with the appropriate sampling distribution for the data have the same distribution as the posterior distribution.
- Conjugate prior distributions have computational convenience.
- They can also be interpreted as additional data.
- They have the disadvantage of constraining the form of the prior distribution.

Nonconjugate prior distributions

- Nonconjugate prior distributions can be used when the shape of the prior knowledge or belief about the distribution of the parameters of interest does not correspond to the conjugate prior distribution.
- Noninformative prior distributions carry little population information and are generally supposed to play a minimal role in the posterior distribution.
  $\longrightarrow$They are also called diffuse, vague, or flat priors.
- Computationally nonconjugate distributions can be more demanding.

Data and Sampling Distribution

- Goal: Inference on the posterior distribution of the transmission probability
- Suppose that n people are exposed once to infection
  - y become infected ("successes")
  - $n - y$ escape infection ("failures")
- Let
  - p = transmission probability
  - $1 - p = q$ = escape probability
- Binomial sampling distribution

$$L(y|p) = \text{Bin}(y|n,p) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} p^y q^{n-y}$$

Specify the Prior Distribution of p

- To perform Bayesian inference, we must specify a prior distribution for p.
- We specify a Beta prior distribution:

$$p \sim \text{Beta}(\alpha, \beta)$$

$$\text{Beta}(p|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \alpha > 0, \beta > 0.$$

- Mean: $E(p|\alpha,\beta) = \frac{\alpha}{\alpha+\beta}$
- Variance: $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{E(p|\alpha,\beta)[1-E(p|\alpha,\beta)]}{\alpha+\beta+1}$

Specify the Prior Distribution of p

- We specify a Beta prior distribution:

$$\begin{aligned} p &\sim \text{Beta}(\alpha, \beta) \\ \pi(p) &= \text{Beta}(p|\alpha, \beta) \\ \text{Beta} &\propto p^{\alpha-1}(1-p)^{\beta-1}. \end{aligned}$$

- Looks similar to binomial distribution
- $\alpha > 0$, $\beta > 0$, "prior sample sizes"

Posterior distribution of p

- The posterior distribution of the transmission probability p, $f(p|y)$:

$$\begin{aligned} f(p|y) &\propto p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1} \\ \text{posterior} &\quad \text{likelihood} \times \textit{prior} \\ &= p^{y+\alpha-1}(1-p)^{n-y+\beta-1} \\ &= \text{Beta}(p|\alpha+y, \beta+n-y) \end{aligned}$$

- The role of α and β as prior sample sizes is clear.

Posterior mean of θ

- Posterior mean of p
  $\longrightarrow$ posterior probability of success (transmission) for a future draw from the population:

$$E(p|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

- posterior mean always lies between the prior mean $\alpha/(\alpha+\beta)$ and the sample mean y/n.
- Posterior variance of p:

$$\text{var}(p|y) = \frac{E(p|y)[1 - E(p|y)]}{\alpha + \beta + n + 1}$$

Uniform prior distribution

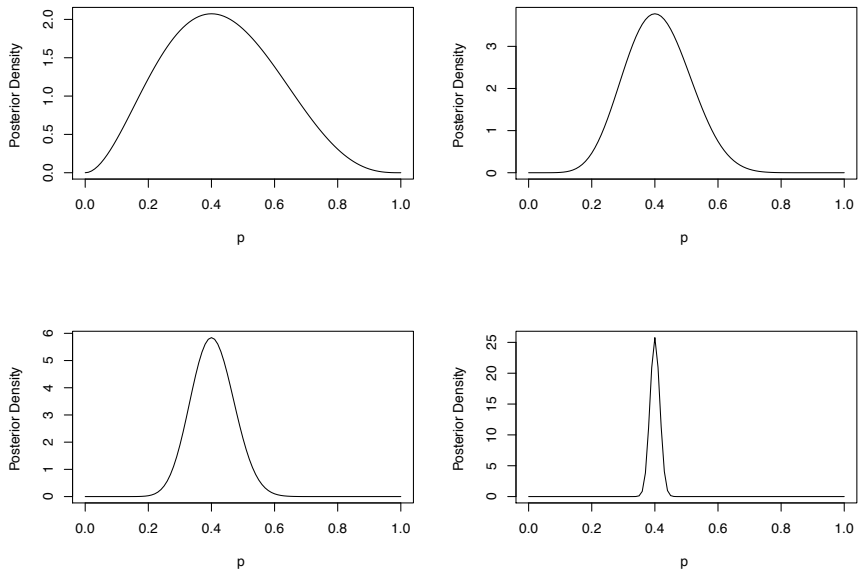
- The uniform prior distribution on [0,1] corresponds to $\alpha = 1$, $\beta = 1$. Essentially no prior information on p.

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

- Let's see how the posterior distribution of the transmission probability depends on the amount of data given a uniform prior distribution (Sample mean $y/n = 0.40$).

n, number exposed	y, number infected
5	2
20	8
50	20
1000	400

Figure : R program: Posterior distribution with differing amounts of data. Uniform Beta prior, Binomial sampling distribution.



Prediction

- After the data have been observed, we can predict a future unknown observable y_{n+1}.
- For example, we may observe n people who were exposed to infection, and whether they became infected.
- We may want to predict the probability that the next person to be observed would become infected.
- Posterior predictive distribution:
  $\longrightarrow$ posterior because conditional on the observed y
  $\longrightarrow$ predictive because it is a prediction for an observable y_{n+1}.

Prediction

- **Posterior predictive distribution of unknown observable** y_{n+1}:

$$\begin{aligned} f(y_{n+1}|y) &= \int f(y_{n+1}, p|y)dp \\ &= \int f(y_{n+1}|p, y)f(p|y)dp \\ &= \int f(y_{n+1}|p)f(p|y)dp \end{aligned}$$

- The last line follows because y and y_{n+1} are conditionally independent given p in this model.
- Useful in model checking.

References

- Gelman, A, Carlin, JB, Stern, HS, Dunson, DB, Vehtari, A, Rubin, DB. *Bayesian Data Analysis*, Chapman and Hall/CRC, third edition, 2014.
- Carlin, BP and Louis, TA. *Bayesian Methods for Data Analysis*, CRC Press, third edition, 2008.

Chain Binomial Model

Table : Rhodes Island measles data: chain binomial probabilities in the Reed-Frost model in $N = 334$ households of size 3 with 1 initial infective and 2 susceptibles, $N_3 = n_{111} + n_{12} = 275$ is observed

Chain	Chain probability	Frequency	Observed frequency	Final number infected
$1 \longrightarrow 0$	q^2	n_1	34	1
$1 \longrightarrow 1 \longrightarrow 0$	$2pq^2$	n_{11}	25	2
$1 \longrightarrow 1 \longrightarrow 1$	$2p^2q$	n_{111}	not observed	3
$1 \longrightarrow 2$	p^2	n_{12}	not observed	3
Total	1	N	334	

Complete data likelihood for q

- The multinomial complete data likelihood for q:

$$\begin{aligned} & f(n_1, n_{11}, N_3, n_{111}|q) \\ = & \binom{334}{n_1, n_{11}, n_{111}, N_3 - n_{111}} (q^2)^{n_1} (2q^2p)^{n_{11}} (2qp^2)^{n_{111}} (p^2)^{N_3 - n_{111}} \\ = & \text{ constant } \times q^{2n_1 + 2n_{11} + n_{111}} p^{n_{11} + 2N_3} \end{aligned}$$

- The observed data are (n_1, n_{11}, N_3), but we do not observe n_{111}.
- We could estimate q using a marginal model, but won't.

Gibbs sampler for chain binomial model

- The general idea of the Gibbs sampler is to sample the model unknowns from a sequence of full conditional distributions and to loop iteratively through the sequence.
- To sample one draw from each full conditional distribution at each iteration, it is assumed that all of the other model quantities are known at that iteration.
- In the theoretical lectures, it will be shown that that the Gibbs sampler converges to the posterior distribution of the model unknowns.
- In the Rhode Island measles data, we are interested in augmenting the missing data n_{111} and estimating the posterior distribution of q, the escape probability.

Gibbs sampler for chain binomial model

- The joint distribution of the observations (n_1, n_{11}, N_3) and the model unknowns (n_{111}, q) is

$$f(n_1, n_{11}, N_3, n_{111}, q) = f(n_1, n_{11}, N_3, n_{111}|q) \times f(q)$$
$$\text{complete data likelihood} \times \text{prior}$$

- We want to make inference about the joint posterior distribution of the model unknowns

$$f(n_{111}, q|n_1, n_{11}, N_3)$$

- This is possible by sampling from the full conditionals (Gibbs sampling): $f(q|n_1, n_{11}, N_3, n_{111})$ and $f(n_{111}|n_1, n_{11}, N_3, q)$

Algorithm for Gibbs sampler for chain binomial model

1. Start with some initial values $(q^{(0)}, n_{111}^{(0)})$
2. For $t = 0$ to M do
3. Sample $q^{(t+1)} \sim f(q|n_1, n_{11}, N_3, n_{111}^{(t)})$
4. Sample $n_{111}^{(t+1)} \sim f(n_{111}|n_1, n_{11}, N_3, q^{(t+1)})$
5. end for
6. How to get the two full conditionals in this model?

Full conditional of chain $1 \longrightarrow 1 \longrightarrow 1$

- Assume q is known
- Compute the conditional probability of chain $1 \to 1 \to 1$ when outbreak size is $N = 3$:

$$\Pr(1 \to 1 \to 1 | N = 3, q) = \frac{\Pr(N = 3, 1 \to 1 \to 1 | q)}{\Pr(N = 3 | q)}$$

$$= \frac{\Pr(N = 3 | 1 \to 1 \to 1, q) \Pr(1 \to 1 \to 1 | q)}{\Pr(N = 3 | 1 \to 1 \to 1, q) \Pr(1 \to 1 \to 1 | q) + \Pr(N = 3 | 1 \to 2, q) \Pr(1 \to 2 | q)}$$

$$= \frac{2p^2 q}{2p^2 q + p^2} = \frac{2q}{2q + 1}, \quad (0 \leq q < 1)$$

The full conditional of n_{111}

- We have found that

$$\Pr(1 \to 1 \to 1 | N = 3, q) = \frac{2q}{2q + 1}$$

- So the full conditional distribution of n_{111} is

$$n_{111} | (n_1, n_{11}, N_3, q) \sim \text{Binomial}(275, 2q/(2q + 1))$$

The full conditional of q

- Assume that n_{111} is known, that is, assume we know the complete data $(n_1, n_{11}, N_3, n_{111})$
- Assume a prior distribution for q: $q \sim \text{Beta}(\alpha, \beta)$,

$$f(q) \equiv f(q|\alpha, \beta) \propto q^{\alpha-1}(1-q)^{\beta-1}$$

- The full conditional distribution of q :

$$f(q|n_1, n_{11}, N_3, n_{111}, \alpha, \beta) \propto f(n_1, n_{11}, N_3, n_{111}|q, \alpha, \beta) f(q|\alpha, \beta)$$

$$\propto \underbrace{q^{2n_1+2n_{11}+n_{111}} p^{n_{11}+2N_3}}_{\text{complete data likelihood}} \times \underbrace{q^{\alpha-1}(1-q)^{\beta-1}}_{\text{prior}}$$

The full conditional of q

- The full conditional distribution of q is thus a Beta distribution

$$q|\text{complete data}, \alpha, \beta \sim \text{Beta}(2n_1 + 2n_{11} + n_{111} + \alpha, n_{11} + 2N_3 + \beta)$$

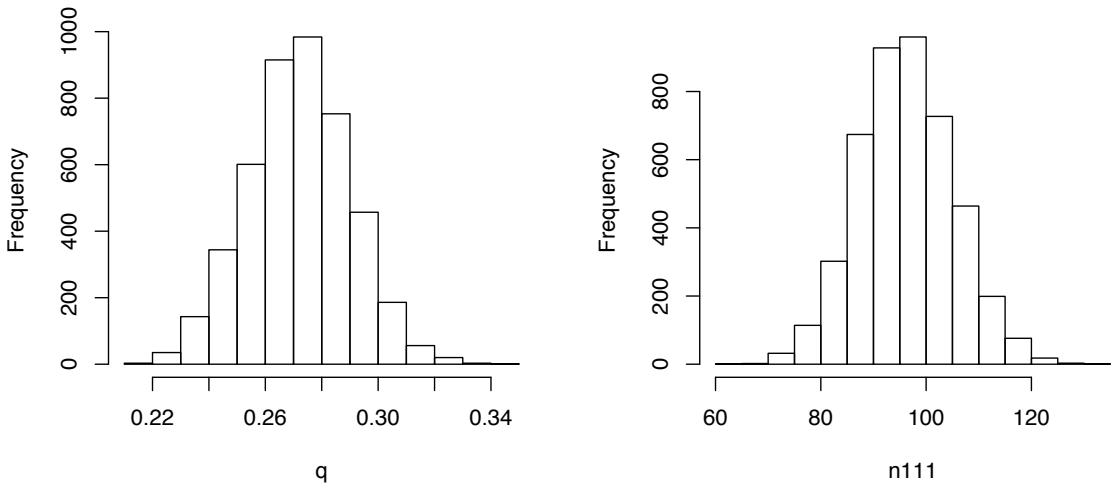
- A uniform prior on q corresponds to $\alpha = 1$, $\beta = 1$.
- With the complete data, a natural point estimate of the escape probability would be the mean of the Beta distribution, i.e., the proportion of "escapes" out of all exposures:

$$\frac{2n_1 + 2n_{11} + n_{111} + \alpha}{2n_1 + 3n_{11} + 3n_{111} + 2n_{12} + \alpha + \beta}$$

Algorithm for Gibbs sampler for chain binomial model

1. Start with some initial values $(q^{(0)}, n_{111}^{(0)})$
2. For $t = 0$ to M do
3. Sample $q^{(t+1)} \sim \text{Beta}(2n_1 + 2n_{11} + n_{111}^{(t)} + \alpha, n_{11} + 2N_3 + \beta)$
4. Sample $n_{111}^{(t+1)} \sim \text{Binomial}(275, 2q^{(t+1)}/(2q^{(t+1)} + 1))$
5. end for
6. Get summaries of the marginal posterior distributions.

Posterior distributions of q and n_{111}



Summer Institute in Statistics and Modeling of Infectious Diseases
Module 7: MCMC Methods for Infectious Disease Studies

Instructors: Kari Auranen, Elizabeth Halloran and Vladimir Minin

July 13 – July 15, 2015

1 Probability refresher (self-study material)

We assume that we can assign probabilities to *events* — outcomes of a random experiment. For example, tossing a coin results in one of two possible events: H ="heads" and T="tails." We also need a concept of a random variable. Informally, a random variable X is a function or variable, whose value is generated by a random experiment. For example, we can define a binary random variable associated with a toss of a coin:

$$X = \begin{cases} 1 & \text{if heads,} \\ 0 & \text{if tails.} \end{cases}$$

Example: Discrete uniform random variable
Let $X \in \{1, 2, \ldots, n\}$, with $\Pr(X = i) = 1/n$ for all $i = 1, \ldots, n$.

Example: Bernoulli r.v.
$X \in \{0, 1\}$ with $\Pr(X = 1) = p$, $\Pr(X = 0) = 1 - p$ for $0 \leq p \leq 1$.

Example: Binomial r.v.
Let $X_i \sim \text{Bernoulli}(p)$. Then the number of successes $S_n = \sum_{i=1}^{n} X_i$ is called a *binomial r.v.* with

$$\Pr(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Example: Geometric r.v.
$X_1, X_2, \ldots$ ordered Bernoulli(p). Let $N = \min\{n : X_n = 1\}$ be the number of trials until the first success occurs, including the the successful trial.

$$\Pr(N = n) = (1-p)^{n-1} p \text{ for } n = 1, 2, \ldots$$

Note. There is an alternative definition of the geometric distribution does not count the successful trial so that $\Pr(N = n) = (1-p)^n p$.

We defined all discrete random variables above using probabilities of X taking a particular value. A function that assigns probabilities to random variable values is called a *probability mass function*. However, a more general way to define random variables is by specifying a cumulative distribution function.

Definition. $F(x) = \Pr(X \leq x)$ is called the *cumulative distribution function* (cdf) of X.

Properties of cdf:

1. $0 \leq F(x) \leq 1$.
2. $F(x) \leq F(y)$ for $x \leq y$.
3. $\lim_{x \to y^+} = F(y)$ ($F(x)$ is right-continuous).
4. $\lim_{x \to -\infty} = \Pr(X = -\infty)$ (usually $= 0$)
5. $\lim_{x \to \infty} = 1 - \Pr(X = \infty)$ (usually $= 1$)
6. $\Pr(X = x) = F(x) - F(x^-)$

Example: Discrete uniform random variable
For random variable U uniformly distributed over $\{1, 2, \ldots, n\}$, its cdf is given by

$$F(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{1}{n} & \text{if } 1 \leq x < 2, \\ \frac{2}{n} & \text{if } 2 \leq x < 3, \\ \vdots & \\ \frac{n-1}{n} & \text{if } n-1 \leq x < n, \\ 1 & \text{if } x \geq n. \end{cases}$$

The probability mass function and cdf of U, with $n = 10$, are shown in Figure 1, which also contains the probability mass function and cdf of a geometric random variable.

For continuous random variables, the analog of the probability mass function is a probability density function, defined as follows.

Definition. If $F(x) = \int_{-\infty}^{x} f(x)dx$ for some $f(x) \geq 0$, then $f(x)$ is called probability density function of X. If X has a probability density function, we say that X is absolutely continuous.

Note. $\int_a^b f(x)\mathrm{d}x = F(b) - F(a) = \Pr(a \leq X \leq b)$ for $a \leq b$. Moreover, $\frac{d}{dx}F(x) = f(x)$.

Example: Uniform random variable on $[0, 1]$
Random variable U with density

$$f(x) = \begin{cases} 1 & \text{if } x \in [0, 1], \\ 0 & \text{otherwise.} \end{cases}$$

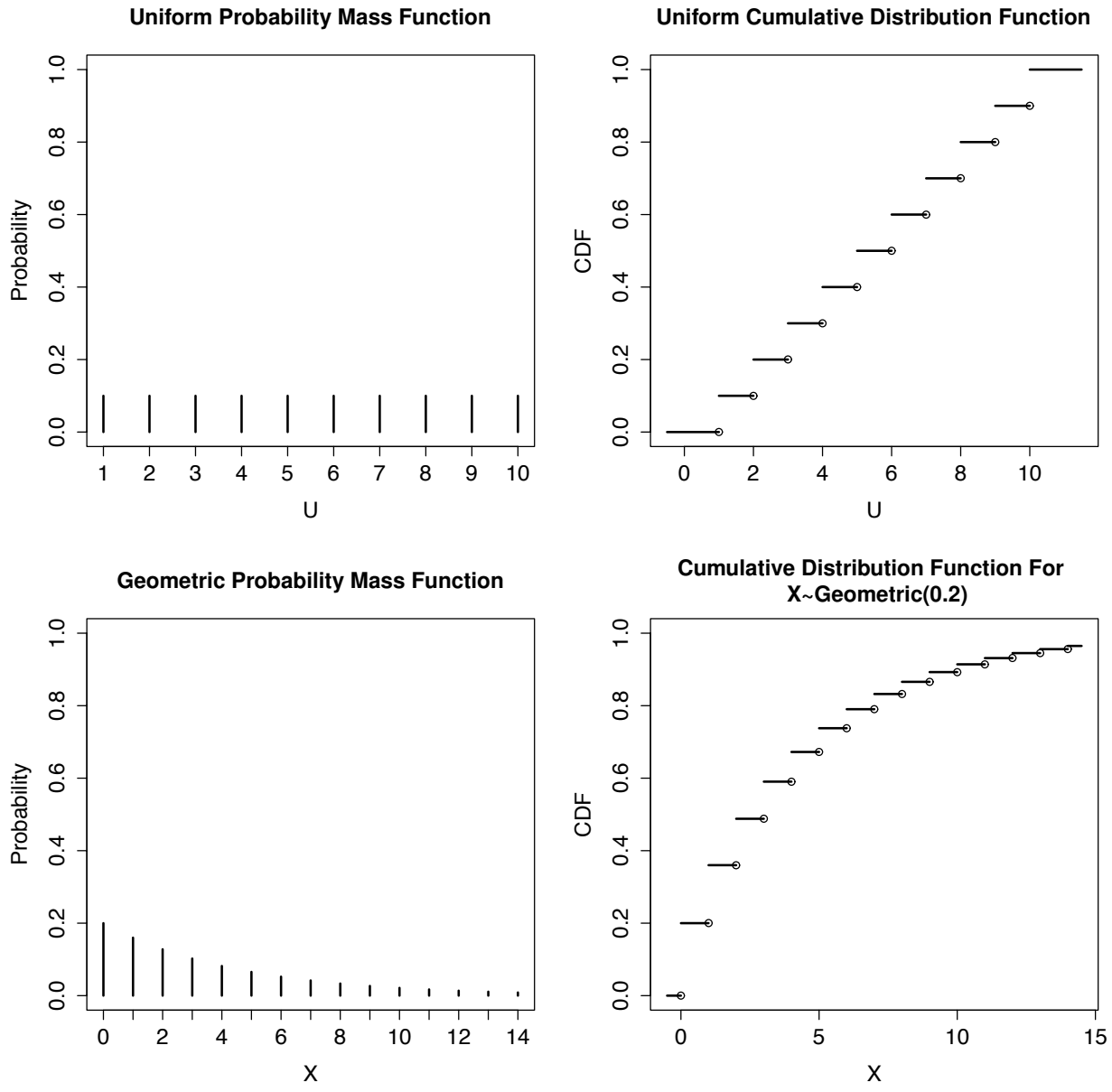


Figure 1: Probability mass functions (left column) and cumulative distribution functions (right column) of the discrete uniform random variable over $\{1, 2, \ldots, 10\}$ (top row) and geometric random variable with success probability $p = 0.2$ (bottom row).

The cdf of U is

$$F(x) = \begin{cases} 0 & \text{if } x < 0. \\ x & \text{if } 0 \leq x \leq 1, \\ 1 & \text{if } x > 1. \end{cases}$$

The top row of Figure 2 shows the probability mass function and cdf of U.

Definition. *Expectation* is defined as $E[g(X)] = \int_{-\infty}^{\infty} g(x)dF(x)$, where the integral is taken with respect to the measure induced by the cdf, aka probability measure. More concretely,

1. For discrete random variable X, $E[g(X)] = \sum_{k=1}^{\infty} g(x_k)\Pr(X = x_k)$.
2. For absolutely continuous random variable X, $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$.

Example: Exponential r.v.

Exponential random variable has density $f(x) = \lambda e^{-\lambda x}1_{\{x \geq 0\}}$, where $\lambda > 0$ is the rate parameter. Let $X \sim \text{Exp}(\lambda)$. The probability mass function and cdf of an exponential random variable are shown in the bottom row of Figure 2. Then

$$E(X) = \int_0^{\infty} x\lambda e^{-\lambda x}dx = \begin{bmatrix} u = x & e^{-\lambda x}dx = dv \\ du = dx & -\frac{e^{-\lambda x}}{\lambda} = v \end{bmatrix} = \lambda \left[-x\frac{e^{-\lambda x}}{\lambda}\Big|_0^{\infty} + \int_0^{\infty} \frac{e^{-\lambda x}}{\lambda}dx \right] = \lambda \left[0 + \frac{1}{\lambda^2} \right] = \frac{1}{\lambda}.$$

Expectations are linear operators, meaning that for any collection of random variables $X_1, \ldots, X_n$,

$$\text{E}\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i \text{E}(X_i).$$

Linearity does not hold for the variance in general. However, if random variables $X_1, \ldots, X_n$ are independent, then

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \text{Var}(X_i).$$

Definition. For events A and B in Ω we define *conditional probability*

$$\Pr(B \mid A) = \frac{\Pr(A \bigcap B)}{\Pr(A)}.$$

If we have a r.v. X defined on Ω, then we can define *conditional expectation*

$$\text{E}(X \mid A) = \frac{E(X1_{\{A\}})}{\Pr(A)}.$$

Conditioning on random variables is a little tricky, so we'll limit our discussion of this concept to

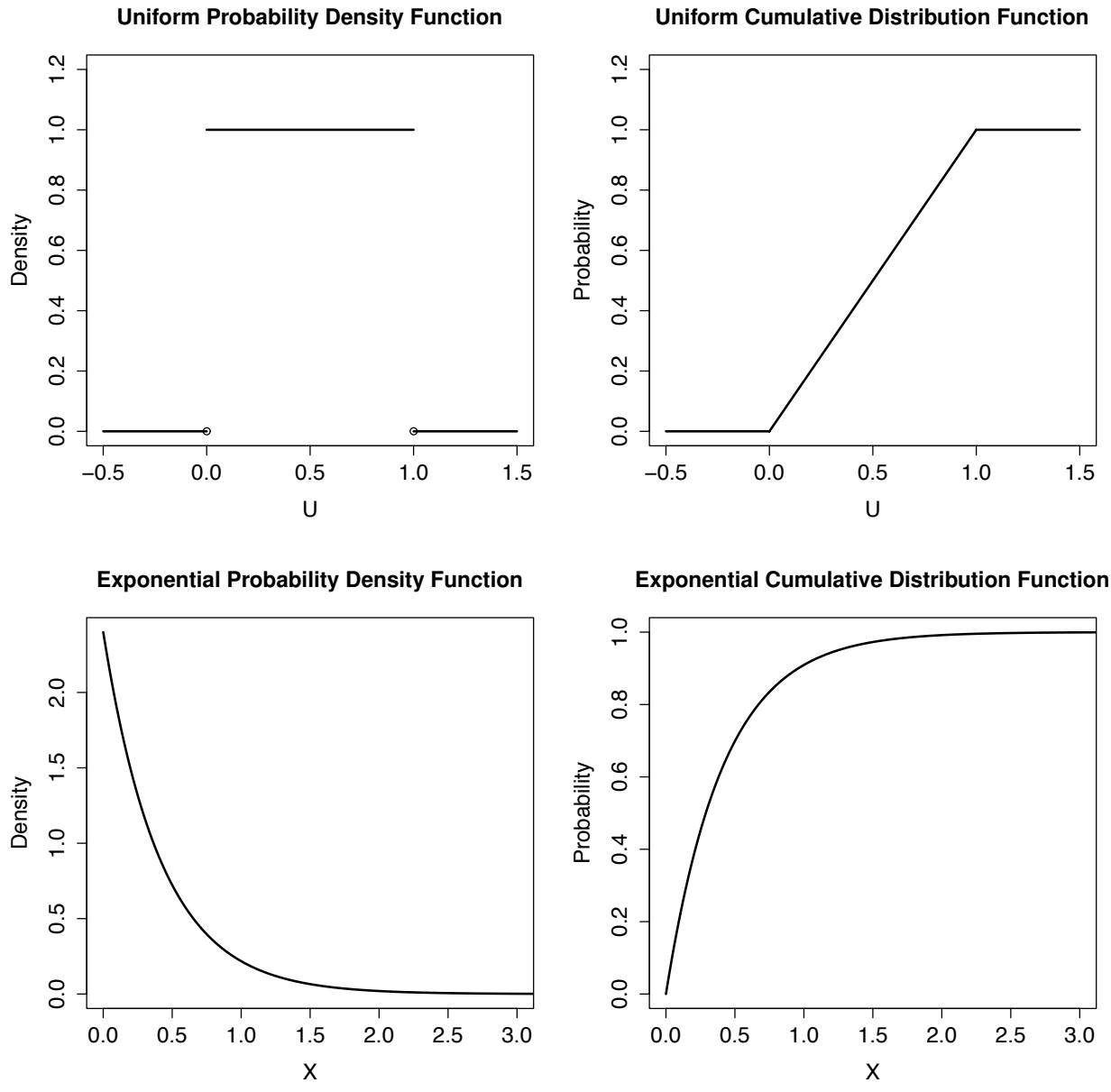


Figure 2: Probability density functions (left column) and cumulative distribution functions (right column) of the continuous uniform random variable on $[0, 1]$ (top row) and exponential random variable with rate parameter $\lambda = 2.4$ (bottom row).

1. discrete case:
$$\Pr(X = x \mid Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)},$$
and

2. absolutely continuous case:
$$F_{X\mid Y}(x|y) = \frac{\int_{-\infty}^{x} f_{XY}(z,y)dz}{f_Y(y)} \qquad \text{and} \qquad f_{X\mid Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)},$$
where $f_{XY}(x,y)$ is the joint density of X and Y and $f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx$ is the marginal density of Y.

Definition. Events A and B are *independent* if $\Pr(A \bigcap B) = \Pr(A)\Pr(B)$. Random variables X and Y are called independent if events $\{X \le a\}$ and $\{Y \le b\}$ are independent for all $a, b \in \mathbb{R}$, i.e. $\Pr(X \le a, Y \le b) = \Pr(X \le a)\Pr(Y \le b)$.

Note. If r.v.s X and Y are independent, then $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ and $\mathrm{E}(X \mid Y) = \mathrm{E}(X)$. The last equality says that Y carries no information about X.

Example: Hypergeometric distribution
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, $S_n = \sum_{i=1}^n X_i$, and $S_m = \sum_{i=1}^m X_i$ for $m < n$. We want to find the distribution of S_m conditional on S_n. We start with probability mass function

$$\Pr(S_m = j \mid S_n = k) = \frac{\Pr(S_m = j, S_n = k)}{\Pr(S_n = k)} = \frac{\Pr(\sum_{i=1}^m X_i = j, \sum_{i=1}^n X_i = k)}{\Pr(S_n = k)}$$
$$= \frac{\Pr\left(\sum_{i=1}^m X_i = j, \sum_{i=m+1}^n X_i = k - j\right)}{\Pr(S_n = k)} = [\text{independence}] = \frac{\Pr(\sum_{i=1}^m X_i = j)\Pr\left(\sum_{i=m+1}^n X_i = k - j\right)}{\Pr(S_n = k)}$$
$$= \frac{\binom{m}{j}p^j(1-p)^{m-j}\binom{n-m}{k-j}p^{k-j}(1-p)^{n-m-k+j}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{\binom{m}{j}\binom{n-m}{k-j}}{\binom{n}{k}}.$$

This is the probability mass function of the hypergeometric distribution, which usually is defined as the number of red balls among the m balls drawn from an urn with k red and $n-k$ blue balls.

$$\mathrm{E}(S_m \mid S_n = k) = \sum_{i=1}^m \mathrm{E}(X_i \mid S_n = k) = [\text{symmetry}] = m\mathrm{E}(X_1 \mid S_n = k) = \frac{m}{n}\sum_{i=1}^n \mathrm{E}(X_i \mid S_n = k)$$
$$= \frac{m}{n}\mathrm{E}(S_n \mid S_n = k) = \frac{mk}{n}.$$

Notice that $X_1, \ldots, X_n$ don't have to be Bernoulli for $\mathrm{E}(S_m \mid S_n) = mS_n/n$ to hold.

Law of total probability If $B_1, \ldots, B_n$ are mutually exclusive events and $\bigcup_{i=1}^n B_i = \Omega$, then

$$\Pr(A) = \sum_{i=1}^n \Pr\left(A \bigcap B_i\right) = \sum_{i=1}^n \Pr(A \mid B_i)\Pr(B_i).$$

Law of total expectation Recall that $\mathrm{E}(X)$ is a scalar, but $\mathrm{E}(X \mid Y)$ is a random variable. Let X and Y be discrete r.v.s.

$$\mathrm{E}(X \mid Y = y) = \sum_{k=1}^{\infty} x_k \mathrm{Pr}(X = x_k \mid Y = y).$$

Proof.

$$\mathrm{E}[\mathrm{E}(X \mid Y)] = \sum_{k=1}^{\infty} \mathrm{E}(X \mid Y = y_k) \mathrm{Pr}(Y = y_k) = \sum_{k=1}^{\infty} \frac{\mathrm{E}\left(X 1_{\{Y=y_k\}}\right)}{\mathrm{Pr}(Y = y_k)} \mathrm{Pr}(Y = y_k)$$
$$= \sum_{k=1}^{\infty} \mathrm{E}\left(X 1_{\{Y=y_k\}}\right) = \mathrm{E}\left(X 1_{\{\bigcup_{k=1}^{\infty}\{Y=y_k\}\}}\right) = \mathrm{E}(X).$$

□

In general, $\mathrm{E}[\mathrm{E}(X \mid Y)] = \mathrm{E}(X)$. In fact, this equality is often used as a definition of the conditional expectation, when conditioning on a random variable [Durret, 2004].

Law of total variance Decomposing variance using conditioning is only slightly more complicated:

$$\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - \mathrm{E}(X)^2 = [\text{law of total expectation}] = \mathrm{E}\left[\mathrm{E}\left(X^2 \mid Y\right)\right] - \mathrm{E}[\mathrm{E}(X \mid Y)]^2$$
$$= [\text{def of variance}] = \mathrm{E}\left[\mathrm{Var}(X \mid Y) + \mathrm{E}(X \mid Y)^2\right] - \mathrm{E}[\mathrm{E}(X \mid Y)]^2 = \mathrm{E}[\mathrm{Var}(X \mid Y)]$$
$$+ \left\{\mathrm{E}\left[\mathrm{E}(X \mid Y)^2\right] - \mathrm{E}[\mathrm{E}(X \mid Y)]^2\right\} = [\text{def of variance}] = \mathrm{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[\mathrm{E}(X \mid Y)].$$

Later in the course, we will be using the following two limit theorems that describe asymptotic behavior of empirical averages of random variables.

Theorem. *Strong Law of Large Numbers (SLLN). Let $X_1, X_2, \ldots$ be independent and identically distributed (iid) random variables with $\mu = E(X_1) < \infty$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu.$$

SLLN says that the empirical average of iid random variables converges to the theoretical average/expectation.

Theorem. *Central Limit Theorem (CLT). Let $X_1, X_2, \ldots$ be independent and identically distributed (iid) random variables with $\mu = E(X_1) < \infty$ and $0 < \sigma^2 = Var(X_1) < \infty$ and let $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim \mathcal{N}(0,1) \text{ approximately for large } n.$$

Informally, CLT says that for large n, the empirical average behaves as $\mathcal{N}(\mu, \sigma^2/n)$. Scaling of the variance by $1/n$ implies that averaging reduces variability, which makes intuitive sense.

2 Monte Carlo methods

The rest of the notes are largely based on [Robert and Casella, 2004]. Although our driving applications of Monte Carlo integration will mostly revolve around Bayesian inference, we would like to point out that all Monte Carlo methods can (should?) be viewed as a numerical integration problem. Such problems usually start with either discrete ($\mathbf{x}$) or continuous ($\boldsymbol{\theta}$) vector of random variables. Despite the fact that distributions of these vectors are known only up to a proportionality constant, we are interested in taking expectations with respect to these distributions. Compare the following integration problems faced by physicists and Bayesian statisticians.

Statistical mechanics	*Bayesian statistics*
$\Pr(\mathbf{x}) = \frac{1}{Z} e^{-\mathcal{E}(\mathbf{x})}$	$\Pr(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{1}{C} \Pr(\mathbf{y} \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta})$
Objective: $\mathrm{E}[f(\mathbf{x})] = \sum_{\mathbf{x}} f(\mathbf{x}) \Pr(\mathbf{x})$	Objective: $\mathrm{E}[f(\boldsymbol{\theta}) \mid \mathbf{y}] = \int f(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta} \mid \mathbf{y}) \, \mathrm{d}\boldsymbol{\theta}$

Note. Many applications involve both, intractable summation and integration:

$$\mathrm{E}[f(\mathbf{x}, \boldsymbol{\theta})] = \sum_{\mathbf{x}} \int f(\mathbf{x}, \boldsymbol{\theta}) \Pr(\mathbf{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

The above integration problems are difficult to solve even numerically, especially in high dimensions, e.g. when the length of $\mathbf{x}$ and/or $\boldsymbol{\theta}$ is on the order of $10^3 - 10^6$. All Monte Carlo techniques attempt to solve such high dimensional integration problems by stochastic simulation.

2.1 Classical Monte Carlo

In general, Monte Carlo integration aims at approximating expectations of the form

$$\mathrm{E}[h(X)] = \int h(x) f(x) \mathrm{d}x. \tag{1}$$

If $X_1, ..., X_n \overset{\text{iid}}{\sim} f(x)$ and $\mathrm{E}[h(X_1)] < \infty$, then we know from the strong law of large number (SLLN) that

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \overset{\text{a.s.}}{\to} \mathrm{E}_f[h(X_1)].$$

Therefore, we can approximate the desired expectation with

$$\bar{h}_n = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \approx \mathrm{E}_f[h(X_1)]$$

for some large, yet finite n. Conveniently, the variance of this Monte Carlo estimator can be approximated as

$$\mathrm{Var}(\bar{h}_n) = \frac{1}{n^2} \times n \times \mathrm{Var}[h(X_1)] = \frac{1}{n} \int \{h(x) - \mathrm{E}_f[h(x)]\}^2 f(x) \mathrm{d}x \approx \frac{1}{n^2} \sum_{i=1}^{n} \left[h(X_i) - \bar{h}_n\right]^2 = v_n$$

Moreover, the central limit theorem says that

$$\frac{\bar{h}_n - \mathrm{E}_f[h(X_1)]}{\sqrt{v_n}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1),$$

allowing us to estimate the Monte Carlo error, e.g. $\bar{h}_n \pm 1.96\sqrt{v_n}$.

Importance Sampling

In many situations classical Monte Carlo is impossible, because we can not sample from the target distribution $f(x)$. Therefore, we would like to be able to compute the integral (1) by sampling from some other, perhaps simpler, distribution $g(x)$. Importance sampling allows us to accomplish this task. The main idea is to rewrite the expectation of interest as

$$\mathrm{E}_f[h(X)] = \int h(x)\frac{f(x)}{g(x)}g(x)\mathrm{d}x = \mathrm{E}_g\left[h(X)\frac{f(X)}{g(X)}\right].$$

This representation suggests that we can generate $X_1, ..., X_n \overset{\text{iid}}{\sim} g(x)$ and use the SLLN again to arrive at the approximation

$$\mathrm{E}_f[h(X)] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)}h(X_i).$$

Notice that the above approximation still requires knowledge of the normalizing constant of $f(x)$, which is unrealistic in most applications of importance sampling. Luckily there is an alternative importance sampling estimator that is as easy to compute as the original one:

$$\mathrm{E}_f[h(X)] \approx \frac{\sum_{i=1}^{n} h(X_i)\frac{f(X_i)}{g(X_i)}}{\sum_{i=1}^{n}\frac{f(X_i)}{g(X_i)}}.$$

In this estimator, the normalizing constants of both $f(x)$ and $g(x)$ cancel out and the denominator converges to $\int \frac{f(x)}{g(x)}g(x)\mathrm{d}x = \int f(x)\mathrm{d}x = 1$ by the SLLN again.

As illustrated by the next example, the importance sampling can be useful even if we can easily simulate from $f(x)$, because importance sampling can be used to reduce the Monte Carlo variance.

Example: Estimating the tail of the standard normal distribution
Let $Z \sim \mathcal{N}(0,1)$. We would like to estimate the tail probability $\Pr(Z > c)$, where c is large (e.g., $c = 4.5$).

Naive Monte Carlo: simulate $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i > c\}} \approx \mathrm{E}\left(1_{\{Z>c\}}\right) = \Pr(Z > c).$$

This estimator will most likely give you 0 even for $n = 10,000$. The problem is the large variance of the integrand:

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}(1_{\{Z_1>c\}}) = \frac{1}{n}\Pr(Z_1 > c)[1-\Pr(Z_1 > c)] = \mathbf{3.4 \times 10^{-10}} \text{ for } n = 10,000 \text{ and } c = 4.5.$$

This variance is huge, because the quantity of interest is $\Pr(Z_1 > c) = 3.39 \times 10^{-6}$ and the standard deviation of our estimator is 1.84×10^{-5}.

Importance sampling: Simulate $Y_1, ..., Y_n \overset{\text{iid}}{\sim} \text{Exp}(c, 1)$ from a shifted exponential with density

$$g(y) = e^{-(y-c)} 1_{\{y>c\}}.$$

Generating such random variables is very easy: just simulate a regular exponential Exp(1) and add c to the simulated value. Then the importance sampling estimator becomes

$$\tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(Y_i)}{g(Y_i)} 1_{\{Y_i>c\}},$$

where $\phi(x)$ is the standard normal density. The variance of this estimator amounts to

$$\begin{aligned} \text{Var}(\tilde{\mu}) &= \frac{1}{n} \text{Var} \left[\frac{\phi(Y)}{g(Y)} 1_{\{Y>c\}} \right] = \frac{1}{n} \left\{ \text{E}_g \left[\frac{\phi^2(Y)}{g^2(Y)} 1_{\{Y>c\}} \right] - \left[\text{E}_g \left(\frac{\phi(Y)}{g(Y)} 1_{\{Y>c\}} \right) \right]^2 \right\} \\ &= \frac{1}{n} \left[\int_c^{\infty} \frac{\phi^2(y)}{g(y)} dy - \Pr(Z > c)^2 \right] = \mathbf{1.9474 \times 10^{-15}} \text{ for } n = 10,000 \text{ and } c = 4.5. \end{aligned}$$

This means that we reduced Monte Carlo variance roughly by a factor of 10^5 using importance sampling.

Now let's code this example up during the practical/lab session.

In conclusion, we point out that the most difficult aspect of classical Monte Carlo is generating iid samples. Even importance sampling has severe limitations in high dimensions. In such difficult cases, Markov chain Monte Carlo (MCMC) can come to rescue. Before we master this numerical integration technique we need to refresh our knowledge of Markov chains.

2.2 Elementary Markov chain theory

In this section we will cover some basic results for Markov chains. For a more detailed treatment, see for example [Brémaud, 1998].

2.2.1 Definitions and examples

Definition. A stochastic process is a family of ordered random variables X_t, where t ranges over a suitable index set T, e.g. $T_1 = [0, \infty)$, $T_2 = \{1, 2, ...\}$.

Definition. A discrete time stochastic process $\{X_n\}_{n=0}^{\infty}$ is called a Markov chain if for all $n \geq 0$ and for all $i_0, i_1, ..., i_{n-1}, i, j$,

$$\text{P}(X_{n+1} = j \,|\, X_n = i, X_{n-1} = i_{n-1}, ..., X_0 = i_0) = \text{P}(X_{n+1} = j \,|\, X_n = i).$$

We call X_n a homogeneous Markov chain if $\Pr(X_{n+1} = j \mid X_n = i)$ is independent of n, and inhomogeneous otherwise. We also define 1-step transition probabilities

$$p_{ij} = \Pr(X_1 = j \mid X_0 = i) \quad \sum_j p_{ij} = 1, p_{ij} \geq 0 \text{ for all } i, j$$

and n-step transition probabilities

$$p_{ij}^{(n)} = \Pr(X_n = j \mid X_0 = i), \quad \sum_j p_{ij}^{(n)} = 1, p_{ij}^{(n)} \geq 0 \text{ for all } i, j$$

and collect them into transition probability matrix $\mathbf{P} = \{p_{ij}\}$ and n-step transition probability matrix $\mathbf{P}^{(n)} = \{p_{ij}^{(n)}\}$.

Note. A Markov chained is fully specified by its transition probability matrix $\mathbf{P}$ and an initial distribution $\boldsymbol{\nu}$.

Note. It is easy to show that n-step transition probabilities can be obtained by repeatedly multiplying transition probability matrix by itself. More precisely, $\mathbf{P}^{(n)} = \mathbf{P}^n$. This observation makes it easy to compute the marginal distribution of X_n, $\boldsymbol{\nu}^{(n)} = (\nu_1^{(n)}, \ldots, \nu_s^{(n)})$, where $\nu_i^{(n)} = \Pr(X_n = i)$. Then

$$\boldsymbol{\nu}^{(n)} = \boldsymbol{\nu}\mathbf{P}^n.$$

Definition. A Markov chain with transition probability matrix $\mathbf{P}$ is called irreducible if for any pair of states (i, j) there exists $n > 0$ such that $p_{ij}^{(n)} > 0$ and reducible otherwise. In other words, an irreducible Markov chain can get from any state to any state in a finite number of steps with positive provability.

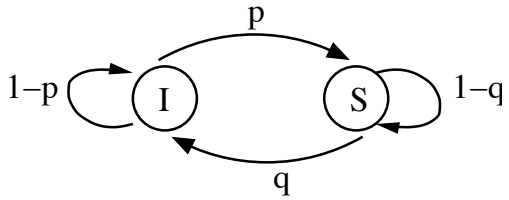
Example: SIS model

Suppose we observe an individual over a sequence of days $n = 1, 2, \ldots$ and classify this individual each day as

$$X_n = \begin{cases} I & \text{if infected} \\ S & \text{if susceptible.} \end{cases}$$

We would like to construct a stochastic model for the sequence $\{X_n\}_{n=1}^{\infty}$. One possibility is to assume that X_ns are independent and $\mathrm{P}(X_n = I) = 1 - \mathrm{P}(X_n = S) = p$. However, this model is not very realistic since we know from experience that the individual is more likely to stay infected if he or she is already infected. Since Markov chains are the simplest models that allow us to relax independence, we proceed by defining a transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$



The directed graph with labeled edges, shown next to the matrix, graphically encodes the same information contained in the transition probability matrix. Such graphs are called transition graphs of Markov chains. If p and q are strictly positive, then the Markov chain is irreducible.

2.2.2 Stationary distribution and long term behavior

Definition. Any probability distribution $\boldsymbol{\pi}$ on state space E that satisfies $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{P}$ (also called the global balance equation) is called a stationary (or equilibrium) distribution of the corresponding homogeneous Markov chain.

Note. $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T P$ if and only if $\pi(i) = \sum_{j\in E} \pi_j p_{ji}$ for all $i \in E$.

Example: SIS model continued

Let's assume that $0 < p < 1$ and $0 < q < 1$ in the SIS model. Then global equations become

$$(\pi_1, \pi_2)\begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = (\pi_1, \pi_2) \Rightarrow \begin{cases} \pi_1(1-p) + \pi_2 q & = \pi_1 \\ \pi_1 p + (1-q)\pi_2 & = \pi_2 \end{cases} \Rightarrow \pi_1 = \frac{q}{p}\pi_2.$$

Adding the constraint $\pi_1 + \pi_2 = 1$, we obtain the unique solution

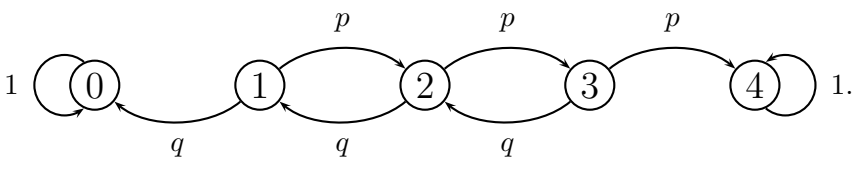
$$\pi_1 = \frac{q}{p+q} \qquad \text{and} \qquad \pi_2 = \frac{p}{p+q}.$$

Not all Markov chains have a stationary distribution and if a stationary distribution exists, it may be not unique as illustrated by the following example.

Example: Gambler's ruin

In this example, we assume that a gambler can increase or decrease his/her fortune by one with corresponding probabilities p and $q = 1-p$. The game ends as soon the gambler runs out of money or reaches a predefined fortune, 4 in our example. The transition matrix and the corresponding transition graph are shown below.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ 0 & q & 0 & p & 0 \\ 0 & 0 & q & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$



The chain is reducible, because it is impossible to get out of states 0 and 4. Such states are called absorbing states. It is easy to show that vector $\boldsymbol{\pi}^T = (\alpha, 0, 0, 0, 1-\alpha)$ satisfies $\boldsymbol{\pi}^T\mathbf{P} = \boldsymbol{\pi}$ for any $\alpha \in [0, 1]$.

Global balance equations can be hard to check in practice when the Markov chain state space is large. However, there is an easier set of equations that one can check to ensure that a stationary distribution exists.

Definition. A probability vector $\boldsymbol{\pi}$ is said to satisfy detailed balance equations with respect to stochastic matrix $\mathbf{P}$ if

$$\pi_i p_{ij} = \pi_j p_{ji} \text{ for all } i, j.$$

Proposition. *(detail balance $\Rightarrow$ global balance) Let $\mathbf{P}$ be a transition probability matrix of X_n on E and let $\boldsymbol{\pi}$ be a probability distribution on E. If $\boldsymbol{\pi}$ satisfies detailed balance equations, then $\boldsymbol{\pi}$ also satisfies global balance equations.*

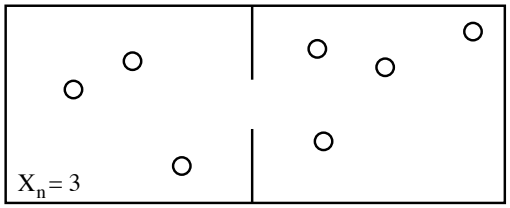
Proof: $\pi_i p_{ij} = \pi_j p_{ji} \Rightarrow \sum_{j\in E} \pi_i p_{ij} = \pi_i \cdot 1 = \sum_{j\in E} \pi_j p_{ji}$. $\square$

Note. Markov chains with a stationary distribution that satisfies detailed balance equations are often called reversible Markov chains. However, there is some disagreement among textbook authors about this term. For example, some authors require reversible chains to have initial distribution being equal to the stationary distribution. Irreducibility is also often added to the list of requirements for reversible Markov chains.

Example: Ehrenfest model of diffusion

Imagine a two dimensional rectangular box with a divider in the middle. The box contains N balls (gas molecules) distributed somehow between the two halves. The divider has a small gap, through which balls can go through one at a time. We assume that at each time step we select a ball uniformly at random and force it go through the gap to the opposite side of the divider. Letting X_n denote the total number of balls in the left half of the box, our Markov process is described by the following transition probabilities.

$$p_{ij} = \begin{cases} \frac{i}{N}. & \text{for } j = i-1, \\ 1 - \frac{i}{N}, & \text{for } j = i+1, \\ 0, \text{ otherwise.} & \end{cases}$$



If we want to derive a stationary distribution of the system, we can solve the global balance equations $\boldsymbol{\pi}^T \mathbf{P} = \boldsymbol{\pi}^T$. Alternatively, we may "guess" that at equilibrium $X_n \sim \text{bin}(\frac{1}{2}, N)$ and verify this candidate stationary distribution via detailed balance. Notice we do not know whether the Ehrenfest chain is reversible, but we'll go ahead with the detailed balance check anyway. First, notice that entries of our candidate vector are

$$\pi_i = \binom{N}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{N-i} = \binom{N}{i} \frac{1}{2^N}$$

Since X_n can only increase or decrease by one at each time step, we need to check detailed balance only for i and $j = i + 1$.

$$\pi_i p_{i,i+1} = \frac{1}{2^N}\binom{N}{i}\frac{N-i}{N} = \frac{1}{2^N}\frac{N!}{i!(N-i)!}\frac{N-i}{N} = \frac{1}{2^N}\frac{N!}{(i+1)!(N-i-1)!}\frac{i+1}{N}$$
$$= \binom{N}{i+1}\frac{1}{2^N}\frac{i+1}{N} = \pi_{i+1}p_{i+1,i},$$

confirming our guess.

Definition. An irreducible Markov chain is called recurrent if starting from any state the chain returns this state eventually with probability one. The recurrent chain is called positive recurrent if all expected return times are finite.

Proposition. *If a Markov chain is irreducible and positive recurrent, then there exists a stationary distribution and this distribution is unique.*

Note. Irreducible Markov chains on finite state spaces are always positive recurrent.

Proposition. *An irreducible Markov chain is positive recurrent if and only if the chain possesses a stationary distribution.*

Theorem. *(Ergodic Theorem) Let $\{X_n\}$ be an irreducible positive recurrent Markov chain with stationary distribution $\boldsymbol{\pi}$ and let $f : E \to \mathbb{R}$ be an arbitrary function that maps Markov chain states to real numbers satisfying $\sum_{i\in E}|f(i)|\pi_i < \infty$. Then for any initial distribution*

$$\underbrace{\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} f(X_k)}_{time\ average} = \sum_{i\in E} f(i)\pi_i = \underbrace{E_{\boldsymbol{\pi}}[f(X)]}_{space\ average}.$$

Example: Ehrenfest model of diffusion (continued)
Separate practical.

Note. Just as the strong law of large numbers is the key behind Monte Carlo simulations, the ergodic theorem for Markov Chains is the reason why Markov chain Monte Carlo (MCMC) works. This remark naturally leads us to the next section.

2.3 Markov chain Monte Carlo

Before we dive into MCMC, let's ask ourselves why we are not happy with classical Monte Carlo and if there is any need to invent something more complicated. The main motivation for developing MCMC is the fact that classical Monte Carlo is very hard to implement in high dimensional spaces. MCMC also often experiences difficulties in high dimensions. However, for almost any

high dimensional integration, it is fairly straightforward to formulate an MCMC algorithm, while the same is not true for classical Monte Carlo.

Recall that our objective in MCMC is the same as in classical Monte Carlo: to estimate expectations of the form

$$\mathrm{E}_{\boldsymbol{\pi}}(h(\mathbf{x}) = \sum_{\boldsymbol{x}\in E} \pi_{\boldsymbol{x}} h(\mathbf{x}).$$

Notice that here we assume that our state space is discrete so the above expectation is a finite sum. However we assume that the size of E is so large that carrying out this summation even on fastest computers is impractical. We also assume that we do not know how to produce iid samples from $\boldsymbol{\pi}$. The general MCMC strategy then is to construct an ergodic Markov chain $\{X_n\}$ with stationary distribution $\boldsymbol{\pi}$. Then from the ergodic theorem and N realizations from the Markov chain, we get

$$\mathrm{E}_{\boldsymbol{\pi}}[h(\mathbf{x})] \approx \frac{1}{N}\sum_{i=1}^{N} h(X_i).$$

The question is how to construct such a Markov chain, $\{X_n\}$.

2.3.1 Metropolis-Hastings algorithm

As always in MCMC, we start with a target distribution $\boldsymbol{\pi}$. Given some initial value $X_0 = x_0$, we construct a Markov chain according to the following set of rules.

Algorithm 2.1 Metropolis-Hastings Algorithm: approximate $\mathrm{E}_{\boldsymbol{\pi}}[h(\mathbf{x})]$

1: Start with some initial value $X_0 = x_0$.
2: **for** $n = 0$ to N **do**
3: Simulate a candidate value $Y \sim q(j \mid X_n = i)$. Suppose $Y = j$.
4: Compute the Metropolis-Hastings acceptance probability

$$a_{ij} = \min\left\{\frac{\pi_j q(i \mid j)}{\pi_i q(j \mid i)}, 1\right\}$$

5: Generate $U \sim \mathrm{Unif}[0, 1]$.
6: Accept the candidate $Y = j$ if $U \leq a_{ij}$, otherwise set $X_{n+1} = X_n$. More specifically, set

$$X_{n+1} = \begin{cases} Y & \text{if } U \leq a_{ij} \\ X_n & \text{if } U > a_{ij} \end{cases}$$

7: **end for**
8: **return** $\frac{1}{N}\sum_{i=1}^{N} h(X_i)$.

Proposition. *The Metropolis-Hastings algorithm generates a Markov chain with stationary distribution* $\boldsymbol{\pi}$.

Proof: Let $\mathbf{P} = \{p_{ij}\}$ be the transition matrix for X_n. Then for $i \neq j$,

$$p_{ij} = \Pr(X_{n+1} = j \,|\, X_n = i) = \Pr(X_1 = j \,|\, X_0 = i) = a_{ij} q(j \,|\, i).$$

Again, for $i \neq j$,

$$\pi_i p_{ij} = \pi a_{ij} q(j \,|\, i) = \begin{cases} \pi_i q(j \,|\, i) \dfrac{\pi_j q(i \,|\, j)}{\pi_i q(j \,|\, i)} & \text{if } \dfrac{\pi_j q(i \,|\, j)}{\pi_i q(j \,|\, i)} \leq 1 \\ \pi_i q(j \,|\, i) \cdot 1 & \text{otherwise} \end{cases} = \begin{cases} \pi_j q(i \,|\, j) & \text{if } \dfrac{\pi_j q(i \,|\, j)}{\pi_i q(j \,|\, i)} \leq 1 \\ \pi_i q(j \,|\, i) & \text{otherwise} \end{cases}$$

and

$$\pi_j p_{ji} = \pi_j a_{ji} q(i \,|\, j) = \begin{cases} \pi_j q(i \,|\, j) \cdot 1 & \text{if } \dfrac{\pi_j q(i \,|\, j)}{\pi_i q(j \,|\, i)} \leq 1 \\ \pi_j q(i \,|\, j) \dfrac{\pi_i q(j \,|\, i)}{\pi_j q(i \,|\, j)} & \text{otherwise} \end{cases} = \begin{cases} \pi_j q(i \,|\, j) & \text{if } \dfrac{\pi_j q(i \,|\, j)}{\pi_i q(j \,|\, i)} \leq 1 \\ \pi_i q(j \,|\, i) & \text{otherwise} \end{cases}$$

So we have shown $\pi_i p_{ij} = \pi_j p_{ji}$. We require $\pi_i > 0$ for all i and $q(i|j) > 0 \Leftrightarrow q(j|i) > 0$. Since we have detailed balance, we conclude that $\boldsymbol{\pi}$ is a stationary distribution. □

Note. If we choose $\{q(i,j)\}$ so that $\{X_n\}$ is irreducible, then $\{X_n\}$ is positive recurrent by the stationary distribution criterion. Therefore, we can use the Ergodic theorem.

Note. We do not need a normalizing constant of $\boldsymbol{\pi}$ in order to execute the Metropolis-Hastings algorithm.

Example: Toric Ising model on a circle

We model ferromagnetism with a set of n electron spins, $\mathbf{x}$. We assume that spins are arranged on a circles and have two directions, denoted by 1 and -1. The Gibbs distribution of configuration $\mathbf{x}$ is

$$\pi(\mathbf{x}) = \frac{1}{Z} e^{\beta \sum_{i=1}^{n} x_i x_{i+1}},$$

where the normalizing constant

$$Z = \sum_{\mathbf{x} \in \{1,-1\}^n} e^{\beta \sum_{i=1}^{n} x_i x_{i+1}}$$

is called a partition function. In this particular example, Z can be computed using a transfer matrix method, but we will pretend that Z is not available to us.

To set up a Metropolis-Hastings algorithm, we need a proposal mechanism to move from one configuration to another. At each step, let's choose a site uniformly at random and change the direction of the spin. This translates to the proposal probabilities

$$q(\mathbf{y} \,|\, \mathbf{x}) = q(\mathbf{x} \,|\, \mathbf{y}) = \begin{cases} \frac{1}{n} & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ at exactly one location,} \\ 0 & \text{otherwise.} \end{cases}$$

If $\mathbf{x}^{(t)}$ is the current state of the Markov chain and $\mathbf{x}'$ is a proposed state with the jth site changed to the opposite direction, then

$$a_{\mathbf{x}^{(t)},\mathbf{x}'} = \frac{\pi(\mathbf{x}')\frac{1}{n}}{\pi(\mathbf{x}^{(t)})\frac{1}{n}} = \frac{e^{\beta\sum_{i\notin\{j,j-1\}} x_i^{(t)}x_{i+1}^{(t)}} e^{\beta(-x_{j-1}^{(t)}x_j^{(t)} - x_j^{(t)}x_{j+1}^{(t)})}}{e^{\beta\sum_{i\notin\{j,j-1\}} x_i^{(t)}x_{i+1}^{(t)}} e^{\beta(x_{j-1}^{(t)}x_j^{(t)} + x_j^{(t)}x_{j+1}^{(t)})}} = e^{-2\beta x_j^{(t)}(x_{j-1}^{(t)} + x_{j+1}^{(t)})}.$$

Clearly, this proposal mechanism makes it possible to get from any state to any other state of spin configurations, so the Metropolis-Hastings chain is irreducible.

Variants of Metropolis-Hastings:

1. $q(i \,|\, j) = q(j \,|\, i)$ - symmetric proposal. This is the original Metropolis algorithm. Here, the acceptance probability simplifies to $a_{ij} = \min\left\{\frac{\pi_j}{\pi_i}, 1\right\}$. So we move to a more probable state with probability 1, and move to less probable states sometimes (more rarely if the candidate is much less probable).

2. Independence sampler: $q(j \,|\, i) = q(j)$. Note this is *not* the same as iid sampling. Independence sampler is still a Markov chain, since the sampler can stay in the same place with some probability at each step of the algorithm.

Metropolis-Hastings algorithm can be executed without any difficulties on continuous state spaces. This requires defining Markov chains on continuous state spaces.

Definition. A sequence of r.v.s $X_0, X_1, \ldots$ is called a Markov chain on a state space E if $\forall t$ and $\forall A \subset E$

$$\Pr(X_{n+1} \in A \,|\, X_n, X_{n-1}, X_0) = \Pr(X_{n+1} \in A \,|\, X_n) = [\text{in homogeneous case}] = \Pr(X_1 \in A \,|\, X_0).$$

A family of functions $\Pr(X_1 \in A \,|\, x) = K(x, A)$ is called <u>transition kernel</u>.

If there exists $f(x, y)$ such that

$$\Pr(X_1 \in A \,|\, x) = \int_A f(x, y)dy,$$

then $f(x, y)$ is called transition kernel density. This is a direct analog of a transition probability matrix in discrete state spaces.

A lot of notions transfer from discrete to continuous state spaces: irreducibility, periodicity, etc. Chapman-Kolmogorov, for example takes the following form:

$$K^{m+n}(x, A) = \int_E K^n(y, A)K^m(x, dy),$$

where $K^n(x, A) = \Pr(X_n \in A \,|\, x)$.

Definition. A probability distribution π on E is called a stationary distribution of a Markov process with transition kernel $K(x, A)$ if for any Borel set B in E

$$\pi(B) = \int_E K(x, B)\pi(dx).$$

If transition kernel density is available, then global balance equation can be re-written

$$\pi(y) = \int_E \pi(x)f(x, y)dx.$$

Using the introduced terminology, we define a Metropolis-Hastings algorithm for continuous state spaces Let $f(\mathbf{x})$ be a target density, where $\mathbf{x}$ is a vector in $\mathbb{R}^n$ now. Then we simply can replace proposal probabilities $q(j \mid i)$ with proposal densities $q(\mathbf{y} \mid \mathbf{x})$ so that Metropolis-Hastings acceptance ratio becomes

$$a(\mathbf{x}, \mathbf{y}) = \min\left\{\frac{f(\mathbf{y})q(\mathbf{x} \mid \mathbf{y})}{f(\mathbf{x})q(\mathbf{y} \mid \mathbf{x})}, 1\right\} \tag{2}$$

The rest of the algorithm remains intact. As before, we need to ensure that the resulting Markov chain is irreducible. One way to do this is to require that $q(\mathbf{y} \mid \mathbf{x}) > 0$ for all $\mathbf{x}, \mathbf{y} \in E$. Alternately, a less restrictive assumption is that there exists some fixed δ and ϵ so that $q(\mathbf{y} \mid \mathbf{x}) > \epsilon$ if $|\mathbf{x}-\mathbf{y}| < \delta$.

A common example of a proposal scheme is a random walk. The proposal is given by

$$Y = X_n + \epsilon_n \tag{3}$$

where ϵ_n is some random perturbation independent of X_n with $E(\epsilon_n) = 0$. By convention, random walk proposals are always taken to be symmetric and have the following form

$$q(y \mid x) = q(|y - x|). \tag{4}$$

Example: Approximating standard normal distribution
Separate practical

2.3.2 Combining Markov kernels

Suppose we have constructed m transition kernels with stationary distribution $\boldsymbol{\pi}$. In discrete state spaces, this means that we have m transition matrices, $\mathbf{P}_1, \ldots, \mathbf{P}_m$, where $\boldsymbol{\pi}^T\mathbf{P}_i = \boldsymbol{\pi}$ for all $i = 1, \ldots, m$. There are two simple ways to combine these transition kernels. First, we can construct a Markov chain, where at each step we sequentially generate new states from all kernels in a predetermined order. The transition probability matrix of this new Markov chain is

$$\mathbf{S} = \mathbf{P}_1 \times \cdots \times \mathbf{P}_m.$$

It is easy to show that $\boldsymbol{\pi}^T\mathbf{S} = \boldsymbol{\pi}$. So as long the new Markov chain is irreducible, we can use the Ergodic theorem applied to the new Markov chain. In the second method of combining Markov

kernels, we first create a probability vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$. Next, we first randomly select kernel i with probability α_i and then use this kernel to advance the Markov chain. The corresponding transition kernel is

$$\mathbf{R} = \sum_{i=1}^{m} \alpha_i \mathbf{P}_i.$$

Again, $\boldsymbol{\pi}^T \mathbf{R} = \boldsymbol{\pi}$, so this MCMC sampling strategy is valid as long as we can guarantee irreducibility.

2.3.3 Gibbs sampling

Suppose now that our state space is a Cartesian product of smaller subspaces, $\mathbf{E} = E_1 \times \cdots \times E_m$. The target distribution or density is $f(\mathbf{x})$ and we still want to calculate $\mathrm{E}_f[h(\mathbf{x})]$. We assume that we can sample from full conditional distributions $x_i \,|\, \mathbf{x}_{-i}$, where the notation $\mathbf{x}_{-i}$ means all elements of $\mathbf{x}$ except the ith component. It turns out that if keep iteratively sampling from these full conditionals, we will form a Markov chain with the required target distribution or density $f(\mathbf{x})$. More formally, let's look at the sequential scan Gibbs sampling algorithm below.

Algorithm 2.2 *Sequential Scan* Gibbs Sampling Algorithm: approximate $\mathrm{E}_{\mathbf{f}}[h(\mathbf{x})]$

1: Start with some initial value $\mathbf{x}^{(0)}$.
2: **for** $t = 0$ to N **do**
3: Sample $x_1^{(t+1)} \sim f_1\left(x_1 \,|\, \mathbf{x}_{-1}^{(t)}\right)$
4: Sample $x_2^{(t+1)} \sim f_2\left(x_2 \,|\, x_1^{(t+1)}, x_3^{(t)}, ..., x_p^{(t)}\right)$
$\vdots$
5: Sample $x_p^{(t+1)} \sim f_p\left(x_p \,|\, \mathbf{x}_{-p}^{(t+1)}\right)$
6: **end for**
7: **return** $\frac{1}{N}\sum_{t=1}^{N} h(\mathbf{x}^{(t)})$.

The question remains why the Gibbs sampling algorithm actually works. Consider one possible move in the Gibbs sampling procedure from $\mathbf{x}^{\text{cur}} \to \mathbf{x}^{\text{new}}$, where $\mathbf{x}^{\text{new}}$ is obtained by replacing the ith component in $\mathbf{x}^{\text{cur}}$ with a draw from the full conditional $f_i\left(x_i \,|\, \mathbf{x}_{-i}^{\text{cur}}\right)$. Now, let's view this "move" in light of the Metropolis-Hastings algorithm. Our proposal density will be the full conditional itself. Then the Metropolis-Hastings acceptance ratio becomes

$$a(\mathbf{x}^{\text{cur}}, \mathbf{x}^{\text{new}}) = \min\left\{\frac{f\left(x_i^{\text{new}}, \mathbf{x}_{-i}^{\text{cur}}\right) f_i\left(x_i^{\text{cur}} \,|\, \mathbf{x}_{-i}^{\text{cur}}\right)}{f\left(x_i^{\text{cur}}, \mathbf{x}_{-i}^{\text{cur}}\right) f_i\left(x_i^{\text{new}} \,|\, \mathbf{x}_{-i}^{\text{cur}}\right)}, 1\right\} = \min\left\{\frac{f\left(\mathbf{x}_{-i}^{\text{cur}}\right)}{f\left(\mathbf{x}_{-i}^{\text{cur}}\right)}, 1\right\} = 1. \tag{5}$$

So when we use full conditionals as our proposals in the Metropolis-Hastings step, we always accept. This means that drawing from a full conditional distribution produces a Markov chain with stationary distribution $f(\mathbf{x})$. Clearly, we can not keep updating just the ith component,

because we will not be able to explore the whole state space this way. Therefore, we update each component in turn. This is not the only way to execute Gibbs sampling. We can also randomly select an component to update. This is called a random scan Gibbs sampling.

Algorithm 2.3 *Random Scan* Gibbs Sampling Algorithm: approximate $\mathrm{E}_{\mathbf{f}}[h(\mathbf{x})]$

1: Start with some initial value $\mathbf{x}_0$.
2: **for** $t = 0$ to N **do**
3: Sample index i by drawing a random variable with probability mass function $\{\alpha_1, \ldots, \alpha_m\}$.
4: Sample $x_i^{(t+1)} \sim f_i\left(x_i \mid \mathbf{x}_{-i}^{(t)}\right)$
5: **end for**
6: **return** $\frac{1}{N}\sum_{t=1}^{N} h(\mathbf{x}^t)$.

Note. Although it is not obvious, but in many cases sampling from full conditional distribution does not require knowing the normalizing constant of the target distribution.

Example: Ising model (continued)
Recall that in the Ising model

$$\pi(\mathbf{x}) = \frac{1}{Z} e^{\beta \sum_{i=1}^{k} x_i x_{i+1}},$$

where $\mathbf{x} = (x_1, \ldots, x_k)$. The full conditional is

$$\pi(x_j \mid \mathbf{x}_{-j}) = \frac{\pi(\mathbf{x})}{\pi(\mathbf{x}_{-j})} = \frac{\pi(\mathbf{x})}{\sum_{y \in \{-1,1\}} \pi(y, \mathbf{x}_{-j})} = \frac{\frac{1}{Z} e^{\beta \sum_{i=1}^{k} x_i x_{i+1}}}{\frac{1}{Z} e^{\beta \sum_{i \notin \{j, j-1\}} x_i x_{i+1}} \left[e^{\beta(x_{j-1} + x_{j+1})} + e^{-\beta(x_{j-1} + x_{j+1})}\right]}$$

$$= \frac{e^{\beta(x_{j-1} x_j + x_j x_{j+1})}}{e^{\beta(x_{j-1} + x_{j+1})} + e^{-\beta(x_{j-1} + x_{j+1})}}.$$

2.3.4 Combining Gibbs and Metropolis-Hastings samplers

Our discussion of combining Markov kernels suggests that it it is possible to combine Gibbs and Metropolis-Hastings steps in MCMC sampler.

Example: Beta-binomial hierarchical model
Separate practical

2.3.5 Variance of MCMC estimators

Let $X_1, X_2, \ldots$ be an ergodic Markov chain and

$$\hat{h} = \frac{1}{N} \sum_{i=1}^{N} h(X_i)$$

be the corresponding estimate of $\mathrm{E}_f[h(X)]$, where f is the stationary distribution of the chain. Estimating the variance of this estimator is complicated by the dependence among $X_1, X_2, \ldots, X_N$.

One simple way to get around it is to subsample the Markov chain output so that the resulting sample is approximately iid. Then, the variance can be approximated as before with

$$\hat{v} = \frac{1}{N^2}\sum_{i=1}^{N}[h(X_i) - \hat{h}]^2.$$

Subsampling can be wasteful and impractical for slow mixing chains. One way to quantify the loss of efficiency due to dependence among samples is to compute the effective sample size,

$$\hat{N}_{eff} = \frac{N}{\kappa_h},$$

where

$$\kappa_h = 1 + 2\sum_{i=1}^{\infty}\text{corr}[h(X_0), h(X_i)]$$

is the autocorrelation time that can be estimated using spectral analysis for time series. After $\hat{N}_{eff}$ is obtained, the variance of $\hat{h}$ is computed as

$$\tilde{v} = \frac{1}{N}\frac{1}{N_{eff}}\sum_{i=1}^{N}[h(X_i) - \hat{h}]^2.$$

2.3.6 Convergence diagnostics

Although there is no definitive way to tell whether one ran a Markov chain long enough, several useful diagnostic tools can illuminate problems with the sampler, bugs in the code, and suggest ways to improve the design of the MCMC sampler. We organize these tools into the following categories:

1. Visualizing MCMC output. Trace plots provide a useful method for detecting problems with MCMC convergence and mixing. Ideally, trace plots of unnormalized log posterior and model parameters should look like stationary time series. Slowly mixing Markov chains produce trace plots with high autocorrelation, which can be further visualized by autocorrelation plots at different lags. Slow mixing does not imply lack of convergence.

2. Comparing batches. We take two vectors from MCMC output: $(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{T/2})$ and $(\boldsymbol{\theta}^{(T/2+1)}, \ldots, \boldsymbol{\theta}^{T})$. If MCMC achieved stationarity at the time of collecting these batches, then both vectors follow the same stationary distribution. To test this hypothesis, we can apply Kolmogorov-Smirnov test, for example.

3. Renewal theory methods. Monitor return times of the Markov chain to a particular state and check whether these return times are iid. Care is needed on continuous state-spaces. See [Mykland et al., 1995] for details.

4. Comparing multiple chains, started from random initial conditions. There are many ways of performing such a comparison. One popular method is called Potential Scale Reduction Factor (PSRF) due to Gelman and Rubin [1992].

Many useful diagnostic tools are implemented in R package CODA [Plummer et al., 2006]. Cowles and Carlin [1995] and Brook and Roberts [1998] review many of the methods in depth.

2.3.7 Special topics

1. Perfect sampling. Strictly speaking perfect sampling is a Monte Carlo, not Markov chain Monte Carlo method. However, the algorithm relies on running Markov chains. Coupling these Markov chains in a certain way (coupling from the past), allows one to generate a sample from the stationary distribution exactly [Propp and Wilson, 1996].

2. Green [1995] formally introduced a Metropolis-Hastings algorithm for sampling parameter spaces with variable dimensions. This class of MCMC is called reversible jump MCMC (rjMCMC). Newton et al. [1992] and Arjas and Gasbarra [1994] have developed reversible jump procedure before Peter Green popularized these algorithms with his now classical 1995 paper.

3. Simulated tempering. Simulated tempering, proposed by Geyer and Thompson [1995], constructs a multivariate Markov chain $(X^{(1)}, \ldots, X^{(n)}$ to sample from the vector-valued function $(f(\mathbf{x}), f^{1/\tau_1}(\mathbf{x}), \ldots, f^{1/\tau_n}(\mathbf{x}))^T$, The auxiliary "heated" chains allow for better exploration of multimodal targets. The idea is similar in spirit to simulated annealing.

4. Sequential importance sampling and particle filters. These methods are useful for sequential building of instrumental densities in high dimensions. The main idea is to use the following representation:

$$f(x_1, \ldots, x_n) = f(x_1 \mid x_2, \ldots, x_n) f(x_2 \mid x_3, \ldots, x_n) \cdots f(x_n).$$

   Using specific structure of the problem at hand, conditioning often simplifies due to conditional independences [Liu and Chen, 1998, Chen et al., 2005].

References

E. Arjas and D. Gasbarra. Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statistica Sinica*, 4:505–524, 1994.

P. Brémaud. *Markov Chains: Gibbs fields, Monter Carlo Simulation, and Queues*. Springer-Verlag, New York, USA, 1998.

S.P. Brook and G.O. Roberts. Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, 8:319–335, 1998.

Y. Chen, J. Xie, and J.S. Liu. Stopping-time resampling for sequential Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 67:199–217, 2005.

M.K. Cowles and B.P. Carlin. Markov chain Monte Carlo diagnostics: a comparative review. *Journal of the American Statistical Association*, 91:883–904, 1995.

R. Durret. *Probability: Theory and Examples*. Duxbury Press, third edition, 2004.

A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

C. Geyer and E. Thompson. Annealing Markov chain Monter Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90:909–920, 1995.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

J.S. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

P. Mykland, L. Tierney, and B. Yu. Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, 90:233–241, 1995.

M.A. Newton, P. Guttorp, and J.L. Abkowitz. Bayesian inference by simulation in a stochastic model from hematology. In *24th Symposium on the Interface: Computing Science and Statistics*, pages 449–455. Fairfax Station: Interface Foundation, 1992.

M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6:7–11, 2006.

J.G. Propp and D.B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, USA, second edition, 2004.

Model checking, hierarchical modeling and combined M-H and Gibbs

SISMID/July 13–15, 2015

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin

Outline

- The chain binomial model for household outbreaks of measles
  - Bayesian analysis of incompletely observed data, using data augmentation (cf. the earlier lecture and computer lab)
  - Checking the model fit through comparison of predictive data with the observed data of the final number infected
- Model extension by allowing heterogeneity across households $\rightarrow$ a hierarchical model
- Implementation of posterior sampling in the hierarchical model by a combined Gibbs and Metropolis algorithm

The observed outbreak sizes

Recall the observed data in the chain binomial model:

Chain	Chain probability	Frequency	Observed frequency	Final number infected
1	q_j^2	n_1	34	1
$1 \to 1$	$2q_j^2 p_j$	n_{11}	25	2
$1 \to 1 \to 1$	$2q_j p_j^2$	n_{111}	not observed	3
$1 \to 2$	p_j^2	n_{12}	not observed	3
Total	1	N	334	

- If the final number infected is 1 or 2, the actual chain is observed
- If the final number infected is 3, the actual chain data are not observed
  - We still know that $N_3 \equiv n_{111} + n_{12} = 275$
- In the previous analysis, we assumed $q_j = q$ for $j = 1, \ldots, 334$, i.e., for all 334 households

Prediction

- Recall that new (predictive) data y^{pred} can be generated by drawing from the posterior predictive distribution $f(y^{\text{pred}}|y)$
- Posterior predictive distribution because
  - conditioning on the observed data y
  - predicting a future observable y^{pred}
- Predictive data can be compared with the observed data to assess the fit of the model
- In this example, we compare the predictive and observed frequencies of chains 1 and $1 \to 1$

Posterior predictive distribution

- Denote the model parameters by θ. Then

$$f(y^{\text{pred}}|y) = \int f(y^{\text{pred}}, \theta|y)d\theta = \int f(y^{\text{pred}}|\theta, y)f(\theta|y)d\theta$$
$$= \int f(y^{\text{pred}}|\theta)f(\theta|y)d\theta$$

- Samples from the posterior predictive distribution can be realised as follows:
  - [1] Draw an MCMC sample θ_k from the posterior $f(\theta|y)$
  - [2] Draw a sample y_k^{pred} from $f(y^{\text{pred}}|\theta_k)$
  - [3] Repeat steps [1] and [2] K times $(k = 1, \ldots, K)$

Model checking

- The posterior predictive distribution of frequencies $(n_1, n_{11}, n_{111}, n_{12})$ is now

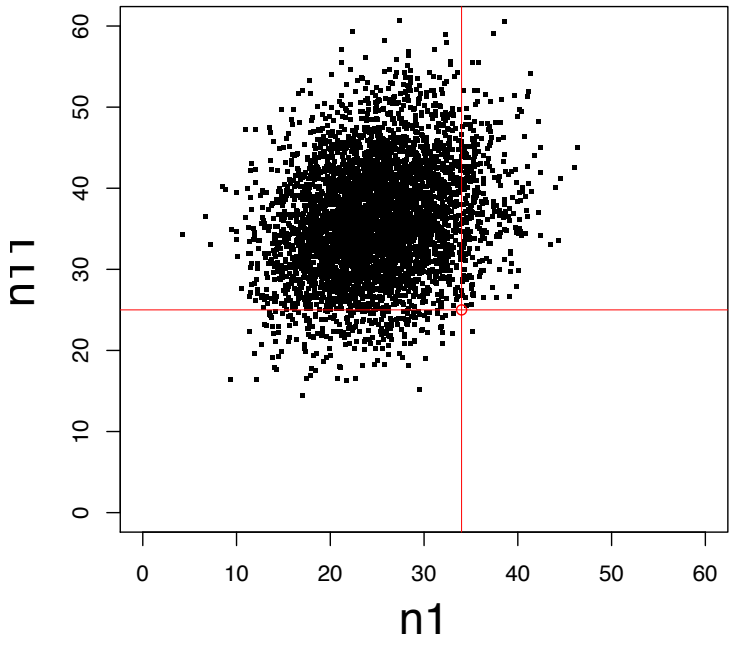
$$f(n_1^{\text{pred}}, n_{11}^{\text{pred}}, n_{111}^{\text{pred}}, n_{12}^{\text{pred}}|n_1, n_{11}, N_3)$$
$$= \int_0^1 f(n_1^{\text{pred}}, n_{11}^{\text{pred}}, n_{111}^{\text{pred}}, n_{12}^{\text{pred}}|q)f(q|n_1, n_{11}, N_3)dq$$

- Samples from the posterior predictive distribution:
  - [1] Draw an MCMC sample $q^{(k)}$ from the posterior $f(q|n_1, n_{11}, N_3)$
  - [2] Draw a sample $(n_1^{(k)}, n_{11}^{(k)}, n_{111}^{(k)}, n_{12}^{(k)})$ from Multinomial$(334, (q^{(k)}, 2(q^{(k)})^2 p^{(k)}, 2q^{(k)}(p^{(k)})^2, p^{(k)}))$
  - [3] Repeat steps [1] and [2] K times $(k = 1, \ldots, K)$

Model checking continues

- Comparison of a sample from the joint predictive posterior of $(n_1^{\text{pred}}, n_{11}^{\text{pred}})$ with the actually observed point (34,25) reveals a poor model fit (next page)
- The model did not take into account possible heterogeneity across households in the escape probability
- Therefore, we'll consider model extension through allowing such heterogeneity

Model checking continues

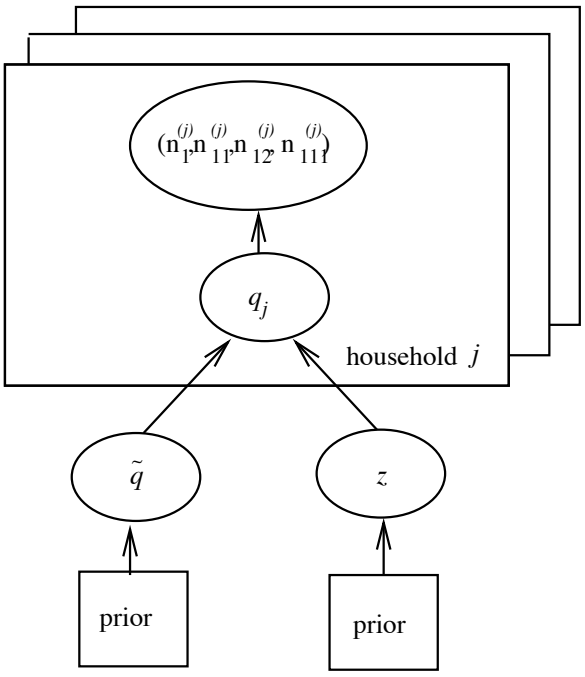


A hierarchical model

- In household j, the observation $(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)})$ follows a multinomial distribution with size 1 and probability vector $(q_j^2, 2q_j^2 p_j, 2q_j p_j^2, p_j^2)$, $j = 1, \ldots, 334$
- The *household-specific* escape probabilities q_j follow a Beta$(\tilde{q}/z, (1-\tilde{q})/z)$ distribution
- Assuming uniform and gamma priors for $\tilde{q}$ and z, respectively, the hierarchical model becomes fully defined:

$$
\begin{aligned}
(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)}) | q_j &\sim \text{Multinomial}(1, (q_j^2, 2q_j^2 p_j, 2q_j p_j^2, p_j^2)) \\
q_j | \tilde{q}, z &\sim \text{Beta}(\tilde{q}/z, (1-\tilde{q})/z)) \\
\tilde{q} &\sim \text{Uniform}(0, 1) \\
z &\sim \text{Gamma}(1.5, 1.5)
\end{aligned}
$$

A hierarchical model continues



The joint distribution

- The joint distribution of the parameters $\tilde{q}$ and z, the household-specific escape probabilities q_j $(j = 1, \ldots, 334)$, and the chain frequencies is

$$\prod_{j=1}^{334} \left(f(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)} | q_j) f(q_j | \tilde{q}, z) \right) f(\tilde{q}) f(z),$$

- The model unknowns are parameters $\tilde{q}$ and z, frequencies $n_{111}^{(j)}$ for all 275 household with outbreak size 3, as well as all 334 household-specific escape probabilities q_j

Sampling from the posterior

- Notation: $\alpha^{(k)} = \tilde{q}^{(k)}/z^{(k)}$, $\beta^{(k)} = (1 - \tilde{q}^{(k)})/z^{(k)}$, k refers to iteration, j refers to household
- A sketch of the steps in kth iteration of the sampling algorithm:

$q_j^{(k)} | \alpha^{(k-1)}, \beta^{(k-1)} \sim \text{Beta}(2 + \alpha^{(k-1)}, \beta^{(k-1)})$, j=1,..., 34

$q_j^{(k)} | \alpha^{(k-1)}, \beta^{(k-1)} \sim \text{Beta}(2 + \alpha^{(k-1)}, 1 + \beta^{(k-1)})$, j=35,..., 59

$q_j^{(k)} | \alpha^{(k-1)}, \beta^{(k-1)}, n_{111}^{(j,k-1)} \sim \text{Beta}(n_{111}^{(j,k-1)} + \alpha^{(k-1)}, 2 + \beta^{(k-1)})$,

j=60,..., 334

$n_{111}^{(j,k)} | q_j^{(k)} \sim \text{Binom}(1, 2q_j^{(k)}/(2q_j^{(k)} + 1))$, j=60,...,334

$\tilde{q}^{(k)} | z^{(k-1)}, q_1^{(k)}, ..., q_{334}^{(k)}$ using a Metropolis-Hastings step

$z^{(k)} | \tilde{q}^{(k)}, q_1^{(k)}, ..., q_{334}^{(k)}$ using a Metropolis-Hastings step

Sampling from the posterior cont.

- In each household, the full conditional (Beta) distribution of $q_j^{(k)}$ depends on the current iterates of the numbers of escapes ($e_j^{(k-1)}$) and infections ($d_j^{(k-1)}$) *in that household* and the prior parameters $\alpha^{(k-1)}$ and $\beta^{(k-1)}$
- The numbers of escapes and infections: see Table
- So, $q_j^{(k)} \sim \text{Beta}(e_j^{(k-1)} + \alpha^{(k-1)}, d_j^{(k-1)} + \beta^{(k-1)})$

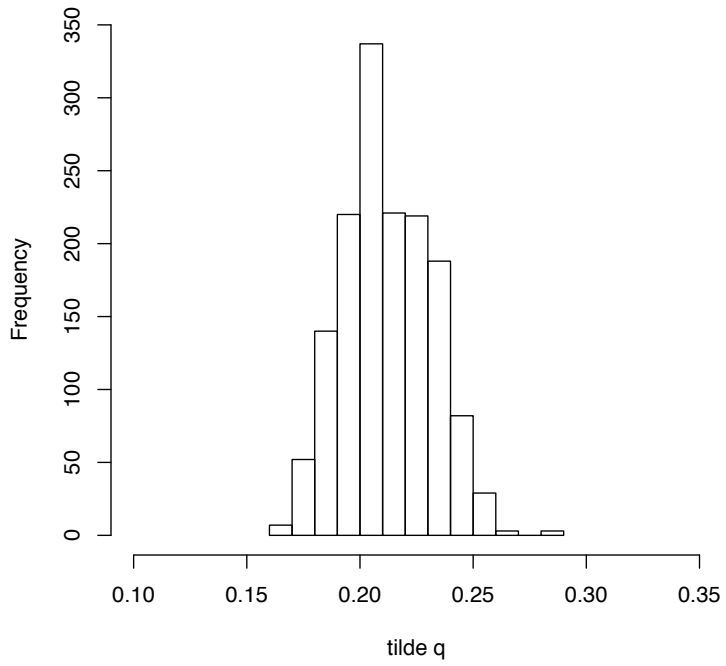
Chain	Number of escapes $e_j^{(k-1)}$	Number of infections $d_j^{(k-1)}$
1	2	0
1→1	2	1
1→1→1	$1 = n_{111}^{(j,k-1)}$	2
1→2	$0 = n_{111}^{(j,k-1)}$	2

Sampling from the posterior cont.

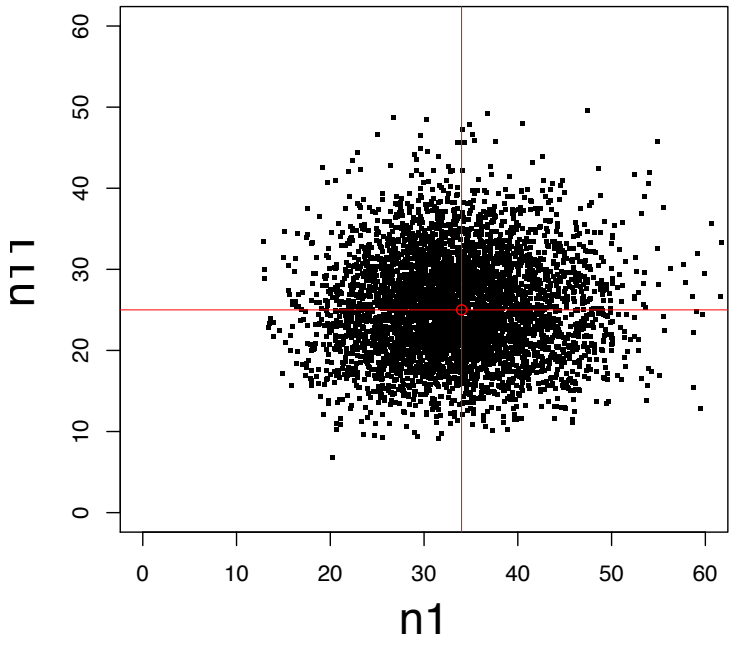
- Parameters $\tilde{q}$ and z require a Metropolis-Hastings step
- For $\tilde{q}$, if the current iterate is $\tilde{q}^{(k-1)}$, a new value $\bar{q}$ is first proposed (e.g.) uniformly about the current iterate (this is a symmetric proposal)
- The proposal is then accepted, i.e., $\tilde{q}^{(k)} := \bar{q}$, with probability

$$\min\{1, \frac{\prod_{j=1}^{334} f(q_j^{(k)}|\bar{q}, z^{(k-1)}) f(\bar{q})}{\prod_{j=1}^{334} f(q_j^{(k)}|\tilde{q}^{(k-1)}, z^{(k-1)}) f(\tilde{q}^{(k-1)})}\}$$

Posterior distribution of $\tilde{q}$



Checking the hierarchical model



An alternative approach

- In this example, it is possible to marginalise q_j over its prior distribution
- This means calculating the chain probabilities under as expectations of the respective probabilities in the previous table, with respect to Beta$(\tilde{q}/z, (1-\tilde{q})/z)$:

Chain	Chain probability	Frequency	Observed frequency	Final number infected
1	$\tilde{q}(\tilde{q}+z)/(1+z)$	n_1	34	1
1→1	$2\tilde{p}\tilde{q}(\tilde{q}+z)/((1+z)(1+2z))$	n_{11}	25	2
1→1→1	$2\tilde{p}\tilde{q}(\tilde{p}+z)/((1+z)(1+2z))$	n_{111}	missing	3
1→2	$\tilde{p}(\tilde{p}+z)/(1+z)$	n_{12}	missing	3

Alternative approach continues

- The following identity helps to calculate the expectations:

$$\mathrm{E}(p_j^u q_j^v) = \frac{\tilde{q}(\tilde{q}+z)\ldots(\tilde{q}+z(u-1))\tilde{p}(\tilde{p}+z)\ldots(\tilde{p}+z(v-1))}{(1+z)\ldots(1+z(u+v-1))}$$

- Using the probabilities as given in the table, it is straightforward to implement a Metropolis-Hastings algorithm to draw samples from the posterior of parameters $\tilde{q}$ and z

[1] Bailey T.J.N. The Mathematical Theory of Infectious Diseases. Charles Griffiths and Company, London 1975.

[2] O'Neill Ph. and Roberts G. Bayesian inference for partially observed stochastic processes. Journal of the Royal Statistical Society, Series A, 1999; 162: 121–129.

[3] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.

[4] O'Neill Ph. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences 2002; 180:103-114.

Data augmentation in the continuous-time SIR model

SISMID/July 13–15, 2015

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin

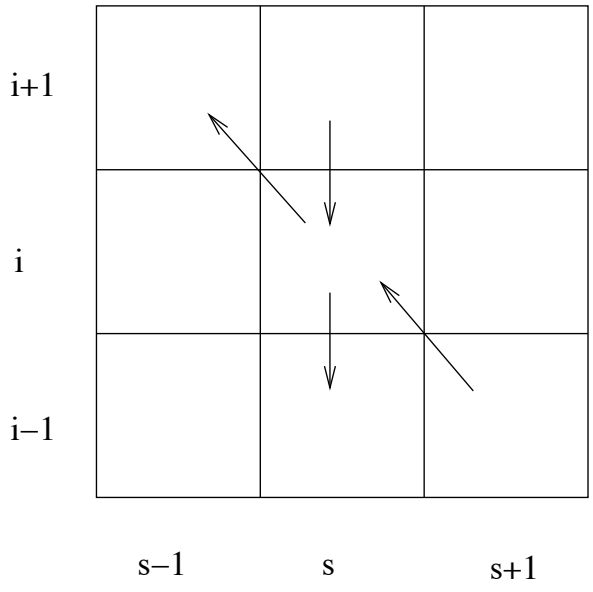
Outline

- The general epidemic model
  - A simple Susceptible–Infected–Removed (SIR) model of an outbreak of infection in a closed population
- Poisson likelihood for infection and removal rates
  - Complete data: both infection and removal times are observed
  - Under Gamma priors for the infection and removal rates, their full conditionals are also Gamma, so Gibbs updating steps can be used
- Incomplete data: only removal times are observed
  - Augment the unknown infection times
  - Additional Metropolis-Hastings steps for sampling infection times, requiring explicit computation of the complete data likelihood

The SIR model

- Consider a closed population of M individuals
- One introductory case (infective) introduces the infection into a population of initially susceptible individuals, starting an outbreak
- Once the outbreak has started, the hazard of infection for a still susceptible individual depends on the number of infectives in the population: $(\beta/M)I(t)$
- If an individual becomes infected, the hazard of clearing infection (and stopping being infective) is γ, i.e., he/she remains infective for an exponentially distributed period of time. He/she then becomes *removed* and does not contribute to the outbreak any more
- There is no latency

Transitions in the state space



The complete data

- Assume one introductory case whose infection takes place at time $t = 0$ (i.e. this fixes the time origin)
- For M individuals followed from time 0 until the end of the outbreak at time T (after which time the number of infectives $I(t) = 0$), the *complete data* record all event times
- This is equivalent to observing $n - 1$ infection times and n removal times, and the fact the $M - n$ individuals escaped infection throughout the outbreak

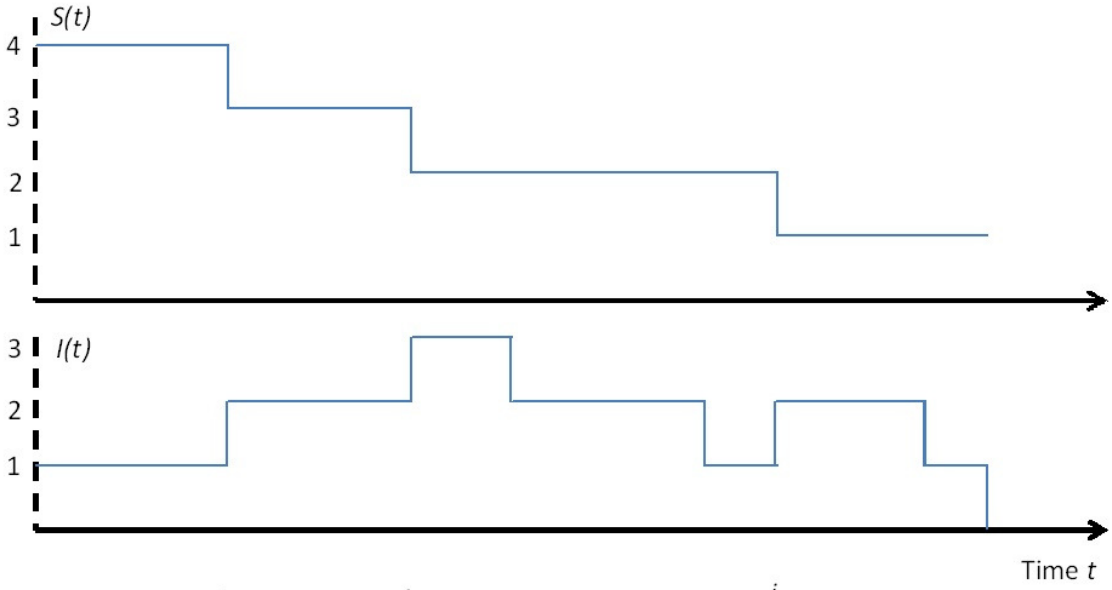
$$\overbrace{\{0 = i_1 < i_2 < ... < i_n\}}^{\text{infection times}} \text{ and } \overbrace{\{r_1 < ... < r_{n-1} < r_n = T\}}^{\text{removal times}}$$

  - N.B. Here, the i_k and r_k need not correspond to the same individual

Counting infectives and susceptibles

- Denote the ordered event times $i_1, \ldots, i_n$ and $r_1, \ldots, r_n$ jointly as $0 = u_1 < u_2 < \ldots < u_{2n} = T$
- Denote the indicators of time u_k being an infection or removal time by D_k and R_k, respectively
- Denote the number of infectives at time t by $I(t)$
  - it is a piecewise constant (left-continuous) function, assuming values in the set $\{0, 1, \ldots, M\}$
  - it jumps at times $u_2 < \ldots < u_{2n}$
- Denote the number of susceptibles at time t by $S(t)$
  - it is a piecewise constant (left-continuous) function, jumping at times $i_2 < \ldots < i_n$
- Both $I(t)$ and $S(t)$ are determined by the complete data

Example



	$i_1 = 0$		i_2		i_3			i_4	
				r_1		r_2		r_3	$r_4 = T$
	$u_1 = 0$		u_2	u_3	u_4	u_5	u_6	u_7	$u_8 = T$
D_k	0		1	1	0	0	1	0	0
R_k	0		0	0	1	1	0	1	1

The process of infections

- The model of new infections is a non-homogeneous Poisson process with rate $\beta I(t)S(t)/M$
  - the rate is a piecewise constant (left-continuous) function
  - it jumps at times $u_2 < \ldots < u_{2n}$, with levels $\beta I(u_2)S(u_2)/M, \beta I(u_3)S(u_3)/M, \ldots, \beta I(u_{2n})S(u_{2n})/M$
- The probability density of the infection events is thus proportional to

$$\prod_{k=2}^{2n} \left[\left((\beta/M) I(u_k) S(u_k) \right)^{D_k} \exp^{-(\beta/M) I(u_k) S(u_k)(u_k - u_{k-1})} \right]$$

$$\propto \prod_{k=2}^{2n} \left(\beta I(u_k) S(u_k) \right)^{D_k} \times \exp^{-(\beta/M) \overbrace{\sum_{k=2}^{2n} I(u_k) S(u_k)(u_k - u_{k-1})}^{\text{total time for "infectious pressure"}}}$$

The process of removals

- The model of removals is a non-homogeneous Poisson process with rate $\gamma I(t)$
  - the rate is a piecewise constant (left-continuous) function
  - it jumps at times $u_2 < \ldots < u_{2n}$, with levels $\gamma I(u_2), \gamma I(u_3), \ldots, \gamma I(u_{2n})$
- The probability density of the removal events is thus proportional to

$$\prod_{k=2}^{2n} \left[(\gamma I(u_k))^{R_k} \exp^{-\gamma I(u_k)(u_k - u_{k-1})} \right]$$

$$= \prod_{k=2}^{2n} (\gamma I(u_k))^{R_k} \times \exp^{-\gamma \overbrace{\sum_{k=2}^{2n} I(u_k)(u_k - u_{k-1})}^{\text{total time spent infective}}}$$

The complete data likelihood

- The joint likelihood of parameters β and γ, based on the complete data:

$$\overbrace{L(\beta, \gamma; \boldsymbol{i}, \boldsymbol{r})}^{f(\boldsymbol{i}, \boldsymbol{r} | \beta, \gamma)} = \prod_{k=2}^{2n} (\beta I(u_k) S(u_k))^{D_k} \prod_{k=2}^{2n} (\gamma I(u_k))^{R_k}$$

$$\times \quad \exp^{-\sum_{k=2}^{2n} ((\beta/M) I(u_k) S(u_k) + \gamma I(u_k))(u_k - u_{k-1})}$$

$$= \quad \prod_{k=2}^{n} \{\beta I(i_k) S(i_k)\} \prod_{k=1}^{n} \{\gamma I(r_k)\}$$

$$\times \quad \exp^{-\sum_{k=2}^{2n} ((\beta/M) I(u_k) S(u_k) + \gamma I(u_k))(u_k - u_{k-1})}$$

Simplifying the notation

- Note that $\sum_k I(u_k)S(u_k)(u_k - u_{k-1}) = \int_0^T I(u)S(u)du$
- Similarly $\sum_k I(u_k)(u_k - u_{k-1}) = \int_0^T I(u)du$
- The likelihood function can thus be written as

$$\prod_{k=2}^{n}\{\beta I(i_k)S(i_k)\}\prod_{k=1}^{n}\{\gamma I(r_k)\} \times \exp\left(-\int_0^T\{(\beta/M)I(u)S(u)+\gamma I(u)\}du\right)$$

Poisson likelihood and Gamma priors

- This above likelihood is the so called Poisson likelihood for parameters β and γ
- In particular, Gamma distributions can be used as conjugate priors for β and γ
- It follows that the full conditional distributions of β and γ are also Gamma and can be updated by Gibbs steps

Gamma prior distributions

- Rate parameters β and γ are given independent Gamma priors

$$f(\beta) \propto \beta^{\nu_\beta - 1} \exp(-\lambda_\beta \beta)$$

$$f(\gamma) \propto \gamma^{\nu_\gamma - 1} \exp(-\lambda_\gamma \gamma)$$

- This allows easy updating of these parameters using Gibbs sampling (the next two pages)

The full conditional of β

- Parameter β can be updated through a Gibbs step

$$f(\beta|\boldsymbol{i}, \boldsymbol{r}, \gamma) \propto f(\beta, \gamma, \boldsymbol{i}, \boldsymbol{r}) \propto f(\boldsymbol{i}, \boldsymbol{r}|\beta, \gamma) f(\beta)$$

$$\propto \beta^{n-1} \exp\left(-(\beta/M) \int_0^T I(u) S(u) du\right) \beta^{\nu_\beta - 1} \exp(-\lambda_\beta \beta)$$

- This means that

$$\beta|(\boldsymbol{i}, \boldsymbol{r}, \gamma) \sim \Gamma\left(n - 1 + \nu_\beta, (1/M) \int_0^T I(u) S(u) du + \lambda_\beta\right)$$

The full conditional of γ

- Parameter γ can be updated through a Gibbs step:

$$f(\gamma|\boldsymbol{i},\boldsymbol{r},\beta) \propto f(\beta,\gamma,\boldsymbol{i},\boldsymbol{r}) \propto f(\boldsymbol{i},\boldsymbol{r}|\beta,\gamma)f(\gamma)$$

$$\propto \gamma^n \exp\left(-\gamma \int_0^T I(u)du\right) \gamma^{\nu_\gamma - 1} \exp(-\lambda_\gamma \gamma)$$

- This means that

$$\gamma|(\boldsymbol{i},\boldsymbol{r},\beta) \sim \Gamma\left(n+\nu_\gamma, \int_0^T I(u)du + \lambda_\gamma\right)$$

Computation of the integral terms

- In practice, the integral terms can be calculated as follows:

$$\overbrace{\int_0^T I(u)du}^{\text{total time spent infective}} = \sum_{k=1}^{n}(r_k - i_k)$$

$$\overbrace{\int_0^T I(u)S(u)du}^{\text{total time for ``infectious pressure''}} = \sum_{k=1}^{n}\sum_{j=1}^{M}(\min(r_k, i_j) - \min(i_k, i_j))$$

where $i_j = \infty$ for $j > n$, i.e., for those never infected

- These expressions are invariant to choice of which r_k corresponds to which i_k

Incomplete data

- Assume that only the removal times $\boldsymbol{r} = (r_1, \ldots, r_n)$ have been observed
- Augment the set of unknowns (β and γ) with infection times $\boldsymbol{i} = (i_2, \ldots, i_n)$
- The aim is to do statistical inference about rates β and γ (and times $\boldsymbol{i}$), based on their posterior distribution $f(\beta, \gamma, \boldsymbol{i}|\boldsymbol{r})$
- The posterior distribution is proportional to the joint distribution of all model quantities:

$$f(\beta, \gamma, \boldsymbol{i}|\boldsymbol{r}) \propto f(\beta, \gamma, \boldsymbol{i}, \boldsymbol{r}) = \overbrace{f(\boldsymbol{i}, \boldsymbol{r}|\beta, \gamma)}^{\text{complete data likelihood}} \overbrace{f(\beta)f(\gamma)}^{\text{prior}},$$

Updating infection times

- The full conditional distributions of β and γ are as above
- The unknown infection times require a Metropolis–Hastings step, including explicit evaluations of the Poisson likelihood
- If the current iterate of i_k is $i_k^{(j)}$, a new value $\tilde{i}_k$ is first proposed (e.g.) from a uniform distribution on $[0, T]$
- The proposal is then accepted, i.e., $i_k^{(j+1)} := \tilde{i}$, with probability

$$\min\{1, \frac{f(\tilde{\boldsymbol{i}}, \boldsymbol{r}|\beta, \gamma)}{f(\boldsymbol{i}, \boldsymbol{r}|\beta, \gamma)}\}$$

- Here $\tilde{\boldsymbol{i}}$ is $\boldsymbol{i}$ except for the kth entry which is $\tilde{i}_k$ (instead of $i_k^{(j)}$)

Augmenting individual histories

- The likelihood above was constructed for the aggregate processes, i.e., to count the total numbers of susceptibles and infectives
- In such case, the corresponding augmentation model must not consider individuals
  - In particular, times $i_2, \ldots, i_n$ must not be tied to particular removal times, i.e., individual event histories must not be reconstructed
- If one considers individual event histories as pairs of times (i_k, r_k) for individuals $k = 1, \ldots, M$, the appropriate complete data likelihood is (cf. above)

$$\gamma^n \prod_{k=2}^{n} \{\beta I(i_k)\} \exp\left(- \int_0^T (\gamma I(u) + (\beta/M) I(u) S(u)) du \right)$$

Example: a smallpox outbreak

- The Abakaliki smallpox outbreak
  - A village of $M = 120$ inhabitants
  - One introductory case
  - 29 subsequent cases; this means that $n = 1 + 29 = 30$
- The observations are given as time intervals between detection of cases (removals) (0 means that symptoms occurred at the same day):

$$13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5$$

- The problem: to estimate rates β and γ from these outbreak data
- See the computer class exercise

References

[1] O'Neill Ph. and Roberts G. Bayesian inference for partially observed stochastic processes. Journal of the Royal Statistical Society, Series A, 1999; 162: 121–129.

[2] O'Neill Ph. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences 2002; 180:103-114.

[3] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.

[4] Andersen et al. Statistical models based on counting processes. Springer Verlag, New York, 1993.

SIS models for recurrent infections

SISMID/July 13–15, 2015

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin

Outline

- Background: recurrent infections
- Binary Markov processes and their generalizations
- Counting process likelihood
- Incomplete observations
  - discrete-time transition models
  - Bayesian data augmentation and reversible jump MCMC
- A computer class exercise

Background

- Many infections can be considered recurrent, i.e., occurring as an alternating series of presence and absence of infection
  - Nasopharyngeal carriage of *Streptococcus pneumoniae* (Auranen et al.; Cauchemez et al.; Melegaro et al.)
  - Nasopharyngeal carriage of *Neisseria menigitidis*
  - multi-resistant *Staphylococcus aureus* (Cooper et al.)
  - some parasitic infections (e.g. Nagelkerke et al.)
- Observation of these processes requires active sampling of the underlying epidemiological states
- Acquisition and clearance times often remain unobserved $\Rightarrow$ incompletely observed data

A binary Markov process

A simple model for a recurrent infection is the binary Markov process:

- The state of the individual alternates between "susceptible" (state 0) and "infected" (state 1)
- The hazard of acquiring infection is λ:

$$\mathrm{P}(\text{acq. in } [t, t+dt[\mid \text{ susceptible at time } t-) = \lambda dt$$

- The hazard of clearing infection is μ:

$$\mathrm{P}(\text{clearance in } [t, t+dt[\mid \text{infected at time } t-) = \mu dt$$

The complete data

- For each individual i, the complete data include the times of acquisition and clearance during the observation period $[0, T]$:
  - Denote the ordered acquisition times for individual i during $]0, T[$ by $\boldsymbol{t}^{(i)} = (t_{i1}, \ldots, t_{iN_{01}^{(i)}})$
  - Denote the ordered clearance times for individual i during $]0, T[$ by $\boldsymbol{r}^{(i)} = (r_{i1}, \ldots, r_{iN_{10}^{(i)}})$
  - Denote the ordered acquisition and clearance times together as $u_{i1} = 0, u_{i2}, \ldots, u_{i,N^{(i)}} = T$
    - Note: these include times 0 and T (so that $N^{(i)} = N_{01}^{(i)} + N_{10}^{(i)} + 2$)

Keeping track who is susceptible

- The indicators for individual i to be susceptible or infected at time t are denoted by $S_i(t)$ and $I_i(t)$, respectively
  - Both indicators are taken to be *predictable*, i.e., they values at time t are determined by the initial value $S_i(0)$ and the complete data observed up to time $t-$
  - Note that $I_i(t) = 1 - S_i(t)$ for all times $t \geq 0$

The process of acquisitions

- In each individual, acquisitions occur with intensity $\lambda S_i(t)$
  - The intensity is λ when the individual is in state 0 (susceptible) and 0 when the individual is in state 1 (infected)
- The probability density of the acquisition events is proportional to

$$\prod_{k=1}^{N^{(i)}} \left[\beta^{1\left(u_k \text{ is time of acq.}\right)} \exp^{-\beta S_i(u_k)(u_k - u_{k-1})} \right]$$

$$\propto \beta^{N_{01}^{(i)}} \times \exp^{-\beta \overbrace{\sum_{k=1}^{N^{(i)}} S_i(u_k)(u_k - u_{k-1})}^{\text{total time susceptible}}}$$

The process of clearances

- In each individual, the clearances occur with intensity $\mu I_i(t)$
  - The intensity is μ when the individual is in state 1 (infected) and 0 when then individual is in state 0 (susceptible)
- The probability density of the clearance events is proportional to

$$\prod_{k=1}^{N^{(i)}} \left[\mu^{1\left(u_k \text{ is time of clearance}\right)} \exp^{-\mu I_i(u_k)(u_k - u_{k-1})} \right]$$

$$= \mu^{N_{10}^{(i)}} \times \exp^{-\mu \overbrace{\sum_{k=1}^{N^{(i)}} I_i(u_k)(u_k - u_{k-1})}^{\text{total time infectd}}}$$

The complete data likelihood

- The likelihood function of parameters β and μ, based on the complete data from individual i:

$$\overbrace{L_i(\beta,\mu;\boldsymbol{t}^{(i)},\boldsymbol{r}^{(i)})}^{f(\boldsymbol{t}^{(i)},\boldsymbol{r}^{(i)}|\beta,\mu)}$$
$$= \beta^{N_{01}^{(i)}}\mu^{N_{10}^{(i)}} \times \exp\ -\sum_{k=1}^{N^{(i)}}(\beta S_i(u_k)+\mu I_i(u_k))(u_k-u_{k-1})$$
$$= \beta^{N_{01}^{(i)}}\mu^{N_{10}^{(i)}} \times \exp\left(-\int_0^T\{\beta S_i(u)+\mu I_i(u)\}du\right)$$

- Likelihood for all M individuals is $\prod_{i=1}^{M} L_i(\beta,\mu;\boldsymbol{t}^{(i)},\boldsymbol{r}^{(i)})$

More complex models

- In the following six slides, the binary model is formulated as a process of counting transitions "0 $\rightarrow$ 1" (acquisitions) and "1 $\rightarrow$ 0" (clearances)
- More complex models can then be defined, allowing e.g.
  - different (sero)types/strains of infection
  - taking into account exposure from other individuals in the relevant mixing groups, e.g., modelling transmission in households

A counting process formulation

- For individual i, the binary process can be described in terms of two counting processes (jump processes):
  - $N_{01}^{(i)}(t)$ counts the number of acquisitions for individual i from time 0 up to time t
  - $N_{10}^{(i)}(t)$ counts the number of clearances for individual i from time 0 up to time t
- Specify the initial state: (e.g.) $N_{01}^{(i)}(0) = N_{10}^{(i)}(0) = 0$
- Denote $H_t^{(i)}$ the history of the processes up to time t: $H_t^{(i)} = \{N_{01}^{(i)}(s), N_{10}^{(i)}(s); 0 \leq s \leq t\}$

Stochastic intensities

- The two counting processes can be specified in terms of their stochastic intensities:

$$\mathrm{P}(dN_{01}^{(i)}(t) = 1 | H_{t-}^{(i)}) = \alpha_{01}^{(i)}(t) Y_0^{(i)}(t) dt$$

$$\mathrm{P}(dN_{10}^{(i)}(t) = 1 | H_{t-}^{(i)}) = \alpha_{10}^{(i)}(t) Y_1^{(i)}(t) dt$$

- Here, $Y_j^{(i)}(t)$ is indicator for individual i being in state j at time $t-$
- In the simple Markov model, $\alpha_{01}^{(i)}(t) = \lambda$, $\alpha_{10}^{(i)}(t) = \mu$, $Y_0^{(i)}(t) = S_i(t)$, and $Y_1^{(i)}(t) = I_i(t)$

Several types of infection

- The infection can involve a "mark", e.g. the serotype of the infection
  - $N_{0j}^{(i)}(t)$ counts the number of times that individual i has acquired infection of type j from time 0 up to time t
  - $N_{j0}^{(i)}(t)$ counts the number of times that individual i has cleared infection of type j from time 0 up to time t
  - Stochastic intensities can be defined accordingly for all possible transitions between the states. For example, for K serotypes, $\alpha_{rs}^{(i)}(t)Y_r^{(i)}(t), \quad r, s = 0, \ldots, K$

Modelling transmission

- The hazard of infection may depend on the presence of infected individuals in the family, day care group, school class etc.
- The statistical unit is the relevant mixing group
- Denote $H_t^{(i,fam)}$ the joint history of all members in the mixing group (e.g. family) of individual i:

$$\mathrm{P}(dN^{(i)}(t) = 1 | H_{t-}^{(i,\mathrm{fam})}) = \alpha_{01}^{(i)}(t)S_i(t)dt \equiv \frac{\beta C^{(i)}(t)}{M_{\mathrm{fam}}^{(i)} - 1} S_i(t)dt$$

where $C^{(i)}(t) = \sum_{j=1}^{M_{fam}^{(i)}} I_j^{(i)}(t)$ is the number of infected individuals in the family of individual i at time $t-$

The counting process likelihood

- For M individuals followed from time 0 to time T, the *complete data* record all transitions between states 0 and 1 (equivalent to observing all jumps in the counting processes):

$$y_{\text{complete}} = \{ T_{rs}^{(ik)};\ r,s = 0,1\ (r \neq s),\ k = 1,\ldots,N_{rs}^{(i)}(T),\ i = 1,\ldots,M \}$$

- The likelihood of the rate parameters θ, based on the complete (event-history) data

$$\overbrace{L(\theta; y_{\text{complete}})}^{f(y_{\text{complete}}|\theta)} = \prod_{i}^{N} \prod_{r \neq s} \prod_{k}^{N_{rs}^{(i)}(T)} \left[\alpha_{rs}^{(i)}(T_{rs}^{(ik)}) \times \exp\left(-\int_0^T \alpha_{rs}^{(i)}(u) Y_r^{(i)}(u) du \right) \right]$$

Remarks

- The likelihood is valid even when the individual processes are dependent on the histories of *other* individuals, e.g. in the case of modelling transmission (cf. Andersen et al)
- The likelihood is correctly normalized with respect to any number of events occurring between times 0 and T (cf. Andersen et al)
  - This is crucial when performing MCMC computations through data augmentation with an unknown number of events

Incomplete observations

- Usually, we do not observe complete data
- Instead, the status $y_j^{(i)}$ of each individual is observed at pre-defined times $t_j^{(i)}$
  - This creates *incomplete data*: the process is only observed at discrete times (panel data)
  - The observed data likelihood is now a complicated function of the model parameters
- How to estimate the underlying continuous process from discrete observations?
  - a discrete-time Markov transition model
  - Bayesian data augmentation

Markov transition models

- Treat the problem as a discrete-time Markov transition model
- This is parameterized in terms of transition probabilities $\mathrm{P}(X^{(i)}(t) = s | X^{(i)}(u) = r)$ for all r, s in the state space χ, and for all times $t \geq u \geq 0$
- In a time-homogeneous model the transition probabilities depend only on the time difference:

$$p_{rs}(t) = \mathrm{P}(X^{(i)}(t) = s | X^{(i)}(0) = r)$$

- This defines a transition probability matrix P_t with entries $[P_t]_{rs} = p_{rs}(t)$, where $\sum_s p_{rs}(t) = 1$ for all r and all $t \geq 0$

The likelihood

- When observations $y_j^{(i)}$ are made at equal time intervals (Δ), the likelihood is particularly simple

$$L(P_\Delta) = \prod_{r,s} [p_{rs}(\Delta)]^{N_{rs}(T)} = \prod_{r,s} [P_\Delta]_{rs}^{N_{rs}(T)}$$

- When observation are actully made at intervals $k\Delta$ only (e.g. Δ = day and $k = 28$), the likelihood is

$$L(P_\Delta) = \prod_{r,s} [P_\Delta^k]_{rs}^{N_{rs}(T)}$$

Modeling transmission

- In a mixing group of size M, the state space is $\chi_1 \times \chi_2 \times \cdots \chi_M$
  - For example, in a family of three the states then are: (0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)
  - For M individuals, the dimension of the state space is 2^M
- Application to pneumococcal carriage in families (Melegaro et al.)
  - The transition probability matrix in a family of 3 (next page), assuming the same probabilities (per day) for each family member
  - Notation: $q_{ii} = 1$ - the sum of the ith row

Transition probability matrix

$$P_\Delta = \begin{array}{c} \\ (0,0,0) \\ (1,0,0) \\ (0,1,0) \\ (0,0,1) \\ (1,1,0) \\ (1,0,1) \\ (0,1,1) \\ (1,1,1) \end{array} \begin{array}{c} \begin{array}{cccccccc} (0,0,0) & (1,0,0) & (0,1,0) & (0,0,1) & (1,1,0) & (1,0,1) & (0,1,1) & (1,1,1) \end{array} \\ \begin{pmatrix} q_{11} & \kappa & \kappa & \kappa & 0 & 0 & 0 & 0 \\ \mu & q_{22} & 0 & 0 & \lambda/2+\kappa & \lambda/2+\kappa & 0 & 0 \\ \mu & 0 & q_{33} & 0 & \beta/2+\kappa & 0 & \beta/2+\kappa & 0 \\ \mu & 0 & 0 & q_{44} & 0 & \lambda/2+\kappa & \lambda/2+\kappa & 0 \\ 0 & \mu & \mu & 0 & q_{55} & 0 & 0 & \lambda+\kappa \\ 0 & \mu & 0 & \mu & 0 & q_{66} & 0 & \lambda+\kappa \\ 0 & 0 & \mu & \mu & 0 & 0 & q_{77} & \lambda+\kappa \\ 0 & 0 & 0 & 0 & \mu & \mu & \mu & q_{88} \end{pmatrix} \end{array}$$

Potential problems

- The dimension of the state space
  - With M individuals and $K+1$ types of infection, the dimension of the state space is $(K+1)^M$
  - With 13 serotypes and 25 individuals (see Hoti et al.), the dimension is $\sim 4.5 \times 10^{28}$
- Non-Markovian sojourn times
  - e.g. a Weibull duration of infection may be more realistic than the exponential one
- Handling of varying observation intervals and individuals with completely missing data are still cumbersome

Bayesian data augmentation

- Retaining the continuous-time model formulation, the unknown event times are taken as additional model unknowns (parameters)
- Statistical inference on all model unknowns (θ and y_{complete})

$$\overbrace{f(y_{\text{observed}}|y_{\text{complete}})}^{\text{observation model}} \overbrace{f(y_{\text{complete}}|\theta)}^{\text{complete data likelihood}} \overbrace{f(\theta)}^{\text{prior}}$$

- The observation model often only ensures agreement with the observed data (as an indicator function)
- The computational problem:
  how to sample from $f(y_{\text{complete}}|y_{\text{observed}}, \theta)$?

The sampling algorithm

- Initialize the model parameters and the latent processes
- For each individual, update the latent processes
  - Update the event times using standard MH
  - Add/delete episodes using reversible jump MH
    - with 0.5 probability propose to add a new episode
    - with 0.5 probability propose to delete an existing episode
- Update the model parameters using single-step MH
- Iterate the updating steps for a given number of MCMC iterations
  - See the computer class exercise

Adding/deleting episodes

- Choose one interval at random from among the K sampling intervals (see page+2)
- Choose to add an episode (delete an existing episode) within the chosen interval with probability $\pi_{\text{add}} = 0.5$ ($\pi_{\text{delete}} = 0.5$)
  - If 'add', choose random event times $\bar{t}_1 < \bar{t}_2$ uniformly from Δ (= the length of the sampling interval). These define the new episode.
  - If 'delete', delete the two event times
- The 'add' move is accepted with probability ("acceptance ratio")

$$\min\left(\frac{f(y_{\text{observed}}|y^*_{\text{complete}})f(y^*_{\text{complete}}|\theta)q(y_{\text{complete}}|y^*_{\text{complete}})}{f(y_{\text{observed}}|y_{\text{complete}})f(y_{\text{complete}}|\theta)q(y^*_{\text{complete}}|y_{\text{complete}})}, 1\right)$$

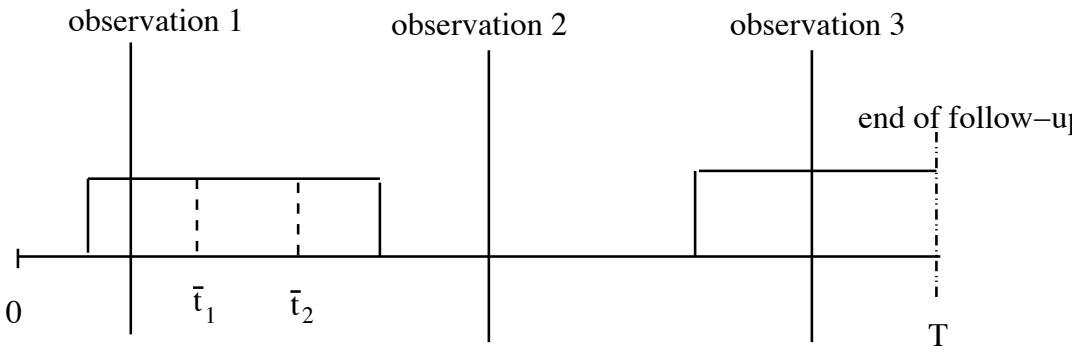
Adding/deleting episodes cont.

- The ratio of the proposal densities is

$$\frac{q(y_{\text{complete}}|y^*_{\text{complete}})}{q(y^*_{\text{complete}}|y_{\text{complete}})} = \frac{\pi_{\text{delete}}\frac{1}{K}\frac{1}{L}}{\pi_{\text{add}}\frac{1}{K}\frac{1}{L}\frac{2}{\Delta^2}} = \frac{\Delta^2}{2}$$

- The ratio of the proposal densities in the 'delete' move is the inverse of the expression above
- Technically, the add/delete step relies on so called reversible jump MCMC (see page+2)
- Reversible jump types should be devised to assure irreducibility of the Markov chain
- For a more complex example, see Hoti et al.

Adding/deleting latent processes cont.



The number of sampling intervals K= 4
The number of 'sub–episodes' within the second interval L = 2

Reversible jump MCMC

- "When the number of things you don't know is one of the things you don't know"
- For example, under incomplete observation of the previous (Markov) processes, the exact number of events is not observed
- This requires a joint model over 'sub-spaces' of different dimensions
- And a method to do numerical integration (MCMC sampling) in the joint state space

References

[1] Andersen et al. "Statistical models based on counting processes", Springer, 1993
[2] Auranen et al. "Transmission of pneumococcal carriage in families – a latent Markov process model for binary data. J Am Stat Assoc 2000; 95:1044-1053.
[3] Melegaro et al. Estimating the transmission parameters of pneumococcal carriage in families. Epidemiol Infect 2004; 132:433-441.
[4] Cauchemez et al. Streptococcus pneumoniae transmission according to inclusion in cojugate vaccines: Bayesian analysis of a longitudinal follow-up in schools. BMC Infectious Diseases 2006, 6:14.
[5] Nakelkerke et al. Estimation of parasitic infection dynamics when detectability is imperfect. Stat Med 1990; 9:1211-1219.
[6] Cooper et al. "An augmented data method for the analysis of nosocomial infection data. Am J Epidemiol 2004; 168:548-557.
[7] Bladt et al. "Statistical inference for discretely observed Markov jump processes. J R Statist Soc B 2005; 67:395-410.
[8] Andersen et al. Multi-state models for event history analysis. Stat Meth Med Res 2002; 11:91-115.
[9] Hoti et al. Outbreaks of Streptococcus pneumoniae carriage in day care cohorts in Finland – implications to elimination of carriage. BMC Infectious Diseases, 2009 (in press)
[10] Green P. Reversible jump Markov chain Monte Carlo computation and Bayesianmodel determination. Biometrika 1995; 82:711-732.

MCMC Methods for Infectious Diseases
Vladimir Minin, Kari Auranen, M. Elizabeth Halloran
Summer Institute in Statistics and Modeling in Infectious Diseases, July 2015

1 Introduction to R and Bayes programming

1.1 Simple Beta posterior distribution

The goal is here to learn simple R programming commands relevant to introductory Bayesian methods. In this first exercise, we compute the posterior distribution of the transmission probability. The sampling distribution is binomial, the prior distribution is Beta, so the posterior distribution is Beta. You can use the command `help(dbeta)` in R to learn more about this function.

Let's see how the posterior distribution of the transmission probability depends on the amount of data given a uniform prior distribution (Sample mean $y/n = 0.40$).

n, number exposed	y, number infected
5	2
20	8
50	20
1000	400

```
## R program to compute the posterior distribution of the transmission probability
## Beta prior distribution for the binomial likelihood

# We want to evaluate the density of the posterior of p along the interval [0,1]
#Generate a sequence from 0 to 1 in increments of .01
x=seq(0,1,.01)

# Observed data
#Number of trials
n=c(5,20,50,1000)

#Number of successes (infections)
y=c(2,8,20,400)

#Set up noninformative Beta prior information
my.alpha=1
my.beta=1
#Set up a matrix to hold the results for the four posterior densities
```

```
posta=matrix(0,4,length(x))

# plot the posterior densities using different amounts of data
#open pdf (or ps)  file graphics device

#pdf(file="/Users/betz/Documents/Bayesintro/betaunif1.pdf", height=6.5, width=8.9)
#set up to have 4 plots in one figure 2 by 2
par(mfrow=c(2,2))
# loop through 4 graphs
for (i in 1:4){
posta[i,] =dbeta(x,my.alpha+y[i],my.beta+n[i]-y[i])
plot(x,posta[i,], type="l", ylab="Posterior Density ", xlab="p ")
}
# close graphics device if using pdf (or ps)
#dev.off()
# return to 1  plot if need be
par(mfrow=c(1,1))
```

1.2 Summaries of the posterior distribution

After obtaining the posterior distribution, we might be interested in certain summary measures such as the posterior mean or posterior median. We might want the 95% posterior interval, equitailed, or any quantiles of interest. We could also ask what is the posterior probability that $p > 0.5$.

```
#Posterior summaries  using closed form distribution

#prior mean
priormean=my.alpha/(my.alpha+my.beta)
priormean

[1] 0.5

# posterior mean
postmean=(my.alpha+y)/(my.alpha+my.beta+n)
postmean

[1] 0.4285714 0.4090909 0.4038462 0.4001996
```

```
#posterior median
# use qbeta() to get the values at the given quantiles
postmedian=qbeta(0.5, my.alpha+y, my.beta+n-y)
postmedian

[1] 0.4214072 0.4062879 0.4026042 0.4001332

#median, 95%  equitailed posterior or credible interval
# use qbeta to get the values at the given quantiles
# set up matrix
post95int=matrix(0,4,3)
for (i in 1:4){
 post95int[i,] = qbeta(c(0.5, 0.025,0.975), my.alpha+y[i], my.beta+n[i]-y[i])
}
round(post95int,3)

     [,1]  [,2]  [,3]
[1,] 0.421 0.118 0.777
[2,] 0.406 0.218 0.616
[3,] 0.403 0.276 0.539
[4,] 0.400 0.370 0.431
```

1.3 Random sampling from the posterior distribution

The function `rbeta()` is used to generate random draws from a given beta distribution. Here we draw 5000 random samples from the four posterior distributions in the first exercise based on different amounts of data and a uniform Beta prior.

```
# Drawing random samples from the posterior distributions to imitate results
# of an MCMC output

#Set the number of samples
nsamp=5000
#Set up matrix to hold the results
post=matrix(0,4,nsamp)

#pdf(file="/Users/betz/Documents/Bayesintro/betaunif1r.pdf", height=6.5, width=8.9)

par(mfrow=c(2,2))
```

```
for (i in 1:4){
post[i,]=rbeta(nsamp,my.alpha+y[i],my.beta+n[i]-y[i])
hist(post[i,], xlim=c(0,1), xlab = " p", ylab =‘‘Frequency’’)
}

#dev.off()
par(mfrow=c(1,1))
```

1.4 Posterior summary using samples of posterior

Now we can get the posterior means using the samples of the posterior distributions and compare them with the analytic posterior means. We can also use the function summary() to get posterior summaries.

```
#Posterior summaries using random samples
# posterior mean
postmeanr =  apply(post,1,mean)
postmeanr

#Compare with analytic posterior mean
postmean

# Get summary of simulated posterior distributions
# Row by row
summary(post[1,])
summary(post[2,])
summary(post[3,])
summary(post[4,])

# Get all summaries of all four rows at the same time
apply(post, 1, summary)
```

1.5 Using informative conjugate priors

Let's assume we have more prior information, or stronger prior beliefs about the transmission probability p. See how it affects posterior inference about one data set, $n1 = 50$, $y1 = 20$ (sample mean $= 0.40$).

Prior mean $\frac{\alpha}{\alpha+\beta}$	Prior sum $\alpha+\beta$	α	β
0.50	2	1	1
0.50	4	2	2
0.50	100	50	50
0.60	5	3	2
0.60	20	12	8
0.60	100	60	40
0.80	5	4	1
0.80	20	16	4

```
## Now use different informative conjugate priors

alpha1 = c(1,2,50,3,12,60,4,16)
beta1 = c(1,2,50,2,8,40,1,4)

priorsum = alpha1+beta1
priormean = alpha1/(alpha1+beta1)
priorsum
priormean

n1 = 50
y1 = 20

post95int2 = matrix(0,length(alpha1),3)
for (i in 1:length(alpha1)){
 post95int2[i,]  = qbeta(c(0.5,0.025,0.975), alpha1[i]+y1, beta1[i]+n1-y1)
}
round(post95int2,3)
      [,1]       [,2]     [,3]
[1,] 0.403 0.276 0.539
[2,] 0.406 0.281 0.540
[3,] 0.467 0.388 0.547
[4,] 0.417 0.292 0.550
[5,] 0.457 0.343 0.574
[6,] 0.533 0.453 0.612
[7,] 0.436 0.309 0.568
[8,] 0.514 0.398 0.630
```

2 Writing a function in R

Suppose we want to be able to write our own function that can be called repeatedly so that we do not need to write the code every time. This is the standard approach in R programming. Here we will write a simple function to compute the posterior distribution of the transmission probability as an illustration. The inputs to the function will be n, the number of trials, y, the number of infections, and α and β from the Beta prior distribution. The function plots the posterior, and returns the posterior median and 95% posterior interval.

```
## Here is a simple function to compute the posterior distribution
## of the transmission probability. It plots the posterior
## distribution and returns the posterior median and
## 95% equitailed credible interval
## n = number of observations,
## y = number of successes (infections)
## alpha, beta, parameters of the prior Beta distribution

mybeta=function(n,y,alpha,beta){
          x=seq(0,1,.01)
          postbeta =dbeta(x,alpha+y,beta+n-y)
          plot(x,postbeta, type="l", ylab="Posterior Density ", xlab="p ")
          mybeta=qbeta(c(0.5, 0.025,0.975), alpha+y, beta+n-y)
            }
## Now call the function
test1=mybeta(50,20,1,1)
test1

### Alternatively, you can return the values at the end.

mybeta2= function(n,y,alpha,beta){
          x=seq(0,1,.01)
          postbeta =dbeta(x,alpha+y,beta+n-y)
          plot(x,postbeta, type="l", ylab="Posterior Density ", xlab="p ")
          return(qbeta(c(0.5, 0.025,0.975), alpha+y, beta+n-y))
            }

test2 = mybeta2(50,20,1,1)
test2
```

```
### Alternatively, you can return a list at the end.

mybeta3=function(n,y,alpha,beta){
          x=seq(0,1,.01)
          postbeta=dbeta(x,alpha+y,beta+n-y)
          plot(x,postbeta, type="l", ylab="Posterior Density ", xlab="p ")
          mybeta=list(answer=qbeta(c(0.5, 0.025,0.975), alpha+y, beta+n-y))
}

test3=mybeta3(50,20,1,1)
test3
test3$answer
```

Practical session:
Chain binomial model I: Gibbs sampler

Background

In this computer lab, we apply Gibbs sampling to incompletely observed data in a chain binomial model. The observations are based on outbreaks of measles in Rhode Island during the years 1929–1934 [1]. We restrict the analysis to families with 3 susceptible individuals at the onset of the outbreak. This example is based on references [1]-[4].

We assume that there is a single index case that introduces infection to the family. Thus, possible epidemic chains are 1, $1 \to 1$, $1 \to 1 \to 1$ and $1 \to 2$. Denote the probability for a susceptible to escape infection when exposed to one infective in the family by q (and $p = 1 - q$). The following table lists chain probabilities, with the actually observed frequencies of the size of epidemic:

chain	prob.	frequency	observed frequency
1	q^2	n_1	34
$1 \to 1$	$2q^2p$	n_{11}	25
$1 \to 1 \to 1$	$2qp^2$	n_{111}	not observed
$1 \to 2$	p^2	n_{12}	not observed

In this exercise, we assume that frequencies n_{111} and n_{12} have not been observed. Only their sum $N_3 = n_{111} + n_{12} = 275$ is known.

The estimation problem concerns the escape probability q, so that there is basically only one unknown parameter in the model. However, the fact that not all frequencies have been observed creates a computational problem that can be solved by Bayesian data augmentation and Gibbs sampling [2].

Marginal likelihood. The joint probability of the *complete data* $(n_1, n_{11}, N_3, n_{111})$ is proportional to a multinomial probability:

$$\begin{aligned} f(n_1, n_{11}, N_3, n_{111}|q) &= (q^2)^{n_1}(2q^2p)^{n_{11}}(2qp^2)^{n_{111}}(p^2)^{N_3 - n_{111}} \\ &= \text{constant} \times q^{2n_1+2n_{11}+n_{111}}p^{n_{11}+2N_3}. \end{aligned} \tag{1}$$

The marginal likelihood $f(n_1, n_{11}, N_3|q)$ would be obtained by summing up expressions (1) with n_{111} running from 0 to N_3.

The Bayesian approach. Instead of using the marginal likelihood, we will treat frequency n_{111} as a model unknown in addition to parameter q. The joint distribution of the observations

(n_1, n_{11}, N_3) and the model unknowns (n_{111}, q) is

$$f(n_1, n_{11}, N_3, n_{111}, q) = f(n_1, n_{11}, N_3, n_{111}|q)f(q). \tag{2}$$

The first term in is the complete data likelihood (see (1)), based on the augmented data (i.e. the data are augmented with the unknown frequency n_{111}).

The second term is the prior density of probability q. We choose a Beta prior for parameter q: $q \sim \text{Beta}(\alpha, \beta)$ so that $f(q) \propto q^{\alpha-1}(1-q)^{\beta-1}$. With the choice $\alpha = \beta = 1$, this is uniform prior on [0,1].

The joint posterior distribution of the model unknowns is $f(q, n_{111}|n_1, n_{11}, N_3)$.

Gibbs sampling. In the lecture we demonstrated that the joint posterior distribution of the model unknowns n_{111} and q can be investigated by Gibbs sampling. This means making a numerical sample from the posterior distribution by drawing samples of n_{111} and q in turn from their full conditional posterior distributions:

$$f(q|n_1, n_{11}, N_3, n_{111}) \quad \text{and} \quad f(n_{111}|n_1, n_{11}, N_3, q).$$

These were found to be

$$q|n_1, n_{11}, N_3, n_{111} \sim \text{Beta}(2n_1 + 2n_{11} + n_{111} + \alpha, n_{11} + 2N_3 + \beta) \tag{3}$$

and

$$n_{111}|n_1, n_{11}, N_3, q = n_{111}|N_3, q \sim \text{Binomial}(N_3, 2q/(2q+1)). \tag{4}$$

Exercises

1. **Gibbs sampling.** The R program (**chainGibbs.R**) contains a function `chainGibbs(mcmc.size,`α, β`)` that draws samples from the joint posterior distribution of q and n_{111}. The function has this particular data set "hardwired" within the program. Using Gibbs sampling, the program draws samples in turn from distributions (3) and (4). Starting with the initial values $(q^{(1)}, n_{111}^{(1)}) = (0.5, 275 * (2 * 0.5)/(2 * 0.5 + 1))$, it iterates between sampling

   $q^{(i)}|n_1, n_{11}, N_3, n_{111}^{(i-1)}$ and

   $n_{111}^{(i)}|n_1, n_{11}, N_3, q^{(i)},\ i = 2, \ldots, mcmc.size.$

   This creates a sample $(q^{(i)}, n_{111}^{(i)}),\ i = 1, \ldots, mcmc.size.$

2. **Write your own Gibbs sampler** Before running chainGibbs.R, you might like to try writing your own Gibbs sampler for the chain binomial problem. Assume you will run *mcmc.size* iterations.

   (a) Reserve space for the *mcmc.size*-vector of q and n_{111} values.

(b) Initialize the model unknowns q[1] and n111[1] (round the n111[1])

(c) Enter the data n1, n11, N3

(d) Draw the MCMC samples 2:*mcmc.size* using the rbeta() and rbinom() functions

3. **Posterior inferences.** By discarding a number of "burn-in" samples, you can use the rest of the numerical sample to explore the posterior of escape probability q. It is enough to discard a few hundred first samples, say 500, in this simple model.

   (a) Make a histogram of the samples 501:*mcmc.size* of q and n111.

   (b) Use the `summary()` function to get summaries the samples 501:*mcmc.size* of q and n111.

4. **Writing a Gibbs sampler function** You can now convert your R program to a function that can be called. It could be similar to the function in the file **chainGibbs.R** `chainGibbs(mcmc.size,`α, β`)`.

   (a) However, you might prefer to write a function `mychainGibbs(n1,n11,N3,mcmc.size,`α, β`)` that allows you to do inference on other data sets with observed (n_1, n_{11}, N_3).

   (b) If you write such a function, try altering the value of N_3. How do larger and smaller values alter the posterior distribution of q?

5. **Sensitivity to the choice of prior.** Assess how the choice of the prior distribution affects estimation of the escape probability. Use the Beta(α, β) prior with different values of α and β. Note that both parameters can be given as input to the function `chainGibbs(mcmc.size,`α, β`)` in **chainGibbs.R** or hopefully your own new function.

References:

[1] Bailey T.J.N. "The Mathematical Theory of Infectious Diseases", Charles Griffiths and Company, London 1975.

[2] O'Neill Ph. and Roberts G. "Bayesian inference for partially observed stochastic processes", *Journal of the Royal Statistical Society*, Series A, **162**, 121–129 (1999).

[3] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.

[4] O'Neill Ph. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences* 2002; 180:103-114.

Practical: Monte Carlo and Markov chain theory

Instructors: Kari Auranen, Elizabeth Halloran and Vladimir Minin

July 13 – July 15, 2015

Estimating the tail of the standard normal distribution

Let $Z \sim \mathcal{N}(0,1)$. We would like to estimate the tail probability $\Pr(Z > c)$, where c is large (e.g. $c = 4.5$).

Your task

Implement naive and importance sampling Monte Carlo estimates of $\Pr(Z > 4.5)$, where $Z \sim \mathcal{N}(0,1)$. Download 'import_sampl_reduced.R' from the course web page. The code has a couple of things to get you started.

Ehrenfest model of diffusion

Consider the Ehrenfest model with $N = 100$ gas molecules. From our derivations we know that the stationary distribution of the chain is $\text{Bin}(\frac{1}{2}, N)$. The chain is irreducible and positive recurrent (why?). The stationary variance can be computed analytically as $N \times \frac{1}{2} \times \frac{1}{2}$.

Your task

Use ergodic theorem to approximate the stationary variance and compare your estimate with the analytical result. Don't panic! You will not have to write everything from scratch. Download 'ehrenfest_diff_reduced.R' file from the course web page. Follow comments in this R script to fill gaps in the code.

Practical: Metropolis-Hastings Algorithm

Instructors: Kari Auranen, Elizabeth Halloran and Vladimir Minin

July 13 – July 15, 2015

Sampling from the standard normal distribution

Suppose our target is a univariate standard normal distribution with density $f(x) = 1/(\sqrt{2\pi})e^{-x^2/2}$. Given current state $x^{(t)}$, we generate two uniform r.v.s $U_1 \sim U[-\delta, \delta]$ and $U_2 \sim U[0, 1]$. Then set

$$x^{(t+1)} = \begin{cases} x^{(t)} + U_1 & \text{if } U_2 \leq \min\left\{e^{\left[\left(x^{(t)}\right)^2 - (x^{(t)}+U_1)^2\right]/2}, 1\right\} \\ x^{(t)} & \text{otherwise.} \end{cases}$$

δ is a tuning parameter. Large δ leads to small acceptance rate, small δ leads to slow exploration of the state space. The rule of thumb for random walk proposals is to keep acceptance probabilities around 30-40%. If your proposal is close to the target, then higher acceptance rates are favorable.

Your task

Implement the above algorithm. Experiment with the tuning parameter δ and report empirically estimate acceptance probabilities for different values of this parameter.

Distribution of the time of infection

Consider a two state continuous-time Markov SIS model, where the disease status X_t cycles between the two states: 1=susceptible, 2=infected. Let the infection rate be λ_1 and clearance rate be λ_2. Suppose that an individual is susceptible at time 0 ($X_0 = 1$) and infected at time T ($X_T = 2$). We don't know anything else about the disease status of this individual during the interval $[0, T]$. If T is small enough, it is reasonable to assume that the individual was infected only once during this time interval. We would like to obtain the distribution of the time of infection, I conditional on the information we have.

Your task

Implement a Metropolis-Hastings sampler to draw realizations from the distribution

$$\Pr(I \mid X_o = 1, X_t = 2, N_t = 1) \propto \Pr(0 < t < I : X_t = 1, I < t < T : X_t = 2),$$

where N_t is the number of infections. Since X_t is a continuous-time Markov chain, the last probability (it is actually a density) can be written as

$$\Pr(0 < t < I : X_t = 1, I < t < T : X_t = 2) = \underbrace{\lambda_1 e^{-\lambda_1 I}}_{\text{density of waiting time until infection}} \quad \underbrace{e^{-\lambda_2 (T-I)}}_{\text{prob of staying infected}}.$$

To make things concrete, set $\lambda_1 = 0.1$, $\lambda_2 = 0.2$ and $T = 1.0$. For your proposal distribution, use a uniform random walk with reflective boundaries 0 and T. In other words, given a current value of the infection time t_c, generate $u = \text{Unif}_{[t_c-\delta, t_c+\delta]}$ $(2\delta < T)$ and then make a proposal value

$$t_p = \begin{cases} u & \text{if } 0 < u < T, \\ 2T - u & \text{if } u > T, \\ -u & \text{if } u < 0. \end{cases}$$

This is a symmetric proposal, so your M-H ratio will contain only the ratio of target densities:

$$\frac{\lambda_1 e^{-\lambda_1 t_p} e^{-\lambda_2 (T - t_p)}}{\lambda_1 e^{-\lambda_1 t_c} e^{-\lambda_2 (T - t_c)}} = e^{-\lambda_1 (t_p - t_c) - \lambda_2 (t_c - t_p)} = e^{(t_p - t_c)(\lambda_2 - \lambda_1)}.$$

Practical: Combining Gibbs and Metropolis-Hastings Kernels

Instructors: Kari Auranen, Elizabeth Halloran and Vladimir Minin

July 13 – July 15, 2015

1 Beta-binomial hierarchical model

Let $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_i \,|\, \theta_i \sim \text{Bin}(n_i, \theta_i)$ and x_is are independent given θ_is. We further assume that $\theta_i \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$. We group all success probabilities into a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and put a prior distribution on hyper-parameters α and β, $\Pr(\alpha, \beta)$. Under our assumptions, the posterior distribution becomes

$$\Pr(\boldsymbol{\theta}, \alpha, \beta \,|\, \mathbf{x}) \propto \Pr(\mathbf{x} \,|\, \boldsymbol{\theta}, \alpha, \beta) \Pr(\boldsymbol{\theta}, \alpha, \beta) \propto \Pr(\alpha, \beta) \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \prod_{i=1}^{n} \theta_i^{x_i}(1-\theta_i)^{n_i-x_i}.$$

We can compute the posterior up to a proportionality constant, but this does not mean that we can compute expectations with respect to the posterior. We will tackle this problem with Markov chain Monte Carlo.

The full condition distribution of θ_i is

$$\Pr(\theta_i \,|\, \mathbf{x}, \alpha, \beta, \boldsymbol{\theta}_{-i}) \propto \theta^{x_i+\alpha-1}(1-\theta_i)^{n_i-x_i+\beta-1}.$$

Therefore,

$$\theta_i \mid \mathbf{x}, \alpha, \beta, \boldsymbol{\theta}_{-i} \sim \text{Beta}(x_i + \alpha, n_i - x_i + \beta).$$

Sampling from $\Pr(\alpha, \beta \,|\, \mathbf{x}, \boldsymbol{\theta})$ directly is difficult, so we will use two Metropolis-Hastings steps to update α and β. To propose new values of α and β, we will multiply their current values by $e^{\lambda(U-0.5)}$, where $U \sim U[0,1]$ and λ is a tuning constant. The proposal density is

$$q(y_{\text{new}} \,|\, y_{\text{cur}}) = \frac{1}{\lambda y_{\text{new}}}.$$

This proposal is not symmetric, so we will have to include it into the M-H acceptance ratio.

Your task

Download the file "beta_bin_reduced.R" from the module web site. We will go through this R script together at first. After you become familiar with data structures used in the script, you will fill in two gaps, marked by "TO DO" comments in the script. Your first task is to replace the line "cur.theta = rep(0.5, data.sample.size)" in the script with code that implements the Gibbs update. Your second task is to implement the M-H steps to sample α and β. The file "beta_bin_reduced.R" contains functions that implement the described proposal mechanism and all the pieces necessary for the acceptance probability. The full MCMC algorithm is outlined on the next page.

Algorithm 1 MCMC for the beta-binomial hierarchical model

1: Start with some initial values $(\boldsymbol{\theta}^{(0)}, \alpha^{(0)}, \beta^{(0)})$.

2: **for** $t = 0$ to N **do**

3: **for** $i = 0$ to n **do**

4: Sample $\theta_i^{(t+1)} \sim \text{Beta}(x_i + \alpha^{(t)}, n_i - x_i + \beta^{(t)})$

5: **end for**

6: Generate $U_1 \sim U[0,1]$ and set $\alpha^* = \alpha^{(t)} e^{\lambda_\alpha (U_1 - 0.5)}$. Generate $U_2 \sim U[0,1]$ and set

$$\alpha^{(t+1)} = \begin{cases} \alpha^* & \text{if } U_2 \leq \min\left\{ \frac{\Pr(\boldsymbol{\theta}^{(t+1)}, \alpha^*, \beta^{(t)} \mid \mathbf{x}) q(\alpha^{(t)} \mid \alpha^*)}{\Pr(\boldsymbol{\theta}^{(t+1)}, \alpha^{(t)}, \beta^{(t)} \mid \mathbf{x}) q(\alpha^* \mid \alpha^{(t)})}, 1 \right\}, \\ \alpha^{(t)} & \text{otherwise.} \end{cases}$$

7: Generate $U_3 \sim U[0,1]$ and set $\beta^* = \beta^{(t)} e^{\lambda_\beta (U_3 - 0.5)}$. Generate $U_4 \sim U[0,1]$ and set

$$\beta^{(t+1)} = \begin{cases} \beta^* & \text{if } U_4 \leq \min\left\{ \frac{\Pr(\boldsymbol{\theta}^{(t+1)}, \alpha^{(t+1)}, \beta^* \mid \mathbf{x}) q(\beta^{(t)} \mid \beta^*)}{\Pr(\boldsymbol{\theta}^{(t+1)}, \alpha^{(t+1)}, \beta^{(t)} \mid \mathbf{x}) q(\beta^* \mid \beta^{(t)})}, 1 \right\}, \\ \beta^{(t)} & \text{otherwise.} \end{cases}$$

8: **end for**

9: **return** $(\boldsymbol{\theta}^{(t)}, \alpha^{(t)}, \beta^{(t)})$, for $t = 1, \ldots, N$.

Practical: Hierarchical chain binomial model

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin
July 13 – July 15, 2015

Background

In this computer class, we re-analyse the data about outbreaks of measles in households. The analysis is restricted to households with 3 susceptible individuals at the onset of the outbreak. We assume that there is a single index case that introduces infection to the household. The possible chains of infection then are 1, $1 \to 1$, $1 \to 1 \to 1$, and $1 \to 2$.

In this example, the probabilities for a susceptible to escape infection when exposed to one infective in the household are allowed to be different in different households. These probabilities are denoted by q_j (and $p_j = 1 - q_j$), $j = 1, \ldots, 334$. The following table expresses the chain probabilities in terms of the escape probability q_j. The observed frequency is the number of households with the respective chain.

chain	prob.	frequency	observed frequency
1	q_j^2	n_1	34
$1 \to 1$	$2q_j^2 p_j$	n_{11}	25
$1 \to 1 \to 1$	$2q_j p_j^2$	n_{111}	not observed
$1 \to 2$	p_j^2	n_{12}	not observed

The frequencies n_{111} and n_{12} have not been observed. Only their sum $N_3 = n_{111} + n_{12} = 275$ is known.

The hierarchical model was defined in the lecture notes. The joint distribution of parameters $\tilde{q}$ and z, the household-specific escape probabilities and

the chain frequencies is

$$\prod_{j=1}^{334} \Big(f(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)} | q_j) f(q_j | \tilde{q}, z) \Big) \, f(\tilde{q}) f(z),$$

where

$$(n_1^{(j)}, n_{11}^{(j)}, n_{111}^{(j)}, n_{12}^{(j)}) | q_j \sim \text{Multinomial}(1, (q_j^2, 2q_j^2 p_j, 2q_j p_j^2, p_j^2)),$$

$$q_j | \tilde{q}, z \sim \text{Beta}(\tilde{q}/z, (1-\tilde{q})/z)),$$

$$\tilde{q} \sim \text{Uniform}(0,1) \text{ and } z \sim \text{Gamma}(1.5, 1.5).$$

N.B. The household-specific chain frequencies are vectors in which only one of the elements is 1, all other elements being 0.

N.B. The Beta distribution is parametrized in terms of $\tilde{q}$ and z for better interpretation of the two parameters. In particular, the prior expectation of the escape probability, given $\tilde{q}$ and z, is $\tilde{q}$, i.e., $\text{E}(q_j | \tilde{q}, z) = \tilde{q}$.

We index the households with chain 1 as 1,...,34, and households with chain $1 \to 1$ as 35,...,59, and households with chain $1 \to 1 \to 1$ or $1 \to 2$ as 60,...,334. The model unknowns are $\tilde{q}$, z, frequencies $n_{111}^{(j)}$ for $j = 60, \ldots, 334$ (i.e., for all 275 households with the final number of infected 3) and q_j for $j = 1, \ldots, 334$ (all households).

In this exercise we apply a combined Gibbs and Metropolis algorithm to draw samples from the posterior distribution of the model unknowns. Before that, we explore the fit of the simple model with $q_j = q$ for all j.

Exercises

1. The simple chain binomial model. Using R routine **chainGibbs.R** (or **mychainGibbs**), i.e., repeating the earlier exercise, realize an MCMC sample from the posterior distribution of the escape probability q in the simple model, in which this probability is the same across all households.

2. Model checking (simple model). Based on the posterior sample of parameter q, draw samples from the posterior predictive distribution of frequencies (n_1, n_{11}). Compare the sample to the actually observed value (34,25). The algorithm to do this is as follows:

(a) Discard a number of "burn-in" samples in the posterior sample of parameter q, as realised in exercise (1) above.

(b) When the size of the retained sample is K, reserve space for the Kx4 matrix of predicted frequencies for n_1, n_{11}, n_{111} and n_{12}.

(c) Based on the retained part of the posterior sample, take the kth sample $q^{(k)}$.

(d) Draw a sample of frequencies $(n_1^{(k)}, n_{11}^{(k)}, n_{111}^{(k)}, n_{12}^{(k)})$
from Multinomial(334,$((q^{(k)})^2, 2(q^{(k)})^2 p^{(k)}, 2q^{(k)}(p^{(k)})^2, (p^{(k)})^2)$) using the rmultinom() function in R.

(e) Repeat steps (c) and (d) K times, storing the sample of frequencies after each step (d).

(f) Plot the samples of pairs $(n_1^{(k)}, n_{11}^{(k)})$, $k = 1, ..., K$, and compare to the observed point (34,25).

The R routine covering steps (a)-(f) is provided in the script **checkmodel_reduced.R**, except for step (d). Complete step (d) and check the model fit:

```
mcmc.sample = chainGibbs(5000,1,1)
checkmodel_reduced(mcmc.sample,1000)
```

The complete R routine (**checkmodel.R**) will be provided once you have tried writing your own code.

3. A hierarchical chain binomial model. Samples from the joint posterior distribution of the unknowns in the hierarchical (beta-binomial) chain model can be sampled using the following algorithm, applying both Gibbs and Metropolis-Hastings updatings steps (superscript k refers to the kth MCMC step):

(a) Reserve space for all model unknowns (cf. page 2 what these are).

(b) Initialize the model unknowns.

(c) Update all household-specific escape probabilities from their full conditionals, with $\alpha^{(k)} = \tilde{q}^{(k)}/z^{(k)}$ and $\beta^{(k)} = (1-\tilde{q}^{(k)})/z^{(k)}$:

$$q_j^{(k)}|\alpha^{(k-1)},\beta^{(k-1)} \sim \text{Beta}(2+\alpha^{(k-1)},\beta^{(k-1)}),\ j=1,\ldots,34$$

$$q_j^{(k)}|\alpha^{(k-1)},\beta^{(k-1)} \sim \text{Beta}(2+\alpha^{(k-1)},1+\beta^{(k-1)}),\ j=35,\ldots,59$$

$$q_j^{(k)}|\alpha^{(k-1)},\beta^{(k-1)},n_{111}^{(j,k-1)} \sim \text{Beta}(n_{111}^{(j,k-1)}+\alpha^{(k-1)},2+\beta^{(k-1)}),\ j=60,\ldots,334$$

(d) Update the unknown binary variables $n_{111}^{(j)}$ $(j=60,\ldots,334)$ from their full conditionals:

$$n_{111}^{(j,k)}|q_j^{(k)} \sim \text{Binomial}(1, 2q_j^{(k)}/(2q_j^{(k)}+1))$$

(e) Sample $\tilde{q}^{(k)}$ using a Metropolis-Hastings step (cf. the program code)

(f) Sample $z^{(k)}$ using a Metropolis-Hastings step (cf. the program code)

(g) Repeat steps (b)–(e) K times (in the R code, K=mcmc.size).

The above algorithm is written in the R script **chain_hierarchical_reduced.R**, except for parts of step (c). Complete the code and draw a posterior sample of all model unknowns. Note that the data set and the prior distributions are hardwired within the given program code.

The complete routine (**chain_hierarchical.R**) will be provided once you have tried your own solution.

4. Posterior inferences. Draw a histogram of the posterior distribution of parameter $\tilde{q}$. This shows the posterior variation in the average escape probability. Using output from program **chain_hierarchical.R**, this can be done as follows (based on 2000 samples with the first 500 as burn-in samples):

```
mcmc.sample = chain_hierarchical(2000)
hist(mcmc.sample$tildeq[500:2000],xlab='tilde q',xlim=c(0.1,0.35))
```

It is also of interest to check how the posterior predictive distribution of q_j looks like and compare it to the *prior predictive* distribution of q_j. For help, see the programme code.

5. Model checking (hierarchical model). Check the fit of the hierarchical model with the R program **check_hierarchical.R**. The program draws samples from the posterior predictive distribution of the chain frequencies and plots the these samples for frequencies n_1 and n_{11} with the actually observed point (34,25).

```
check_hierarchical(mcmc.sample,mcmc.burnin=500)
```

N.B. Unlike we pretended in the preceding exercises, the original data actually record the frequencies $n_{12} = 239$ and $n_{111} = 36$. You can now check the model fit with respect to these frequencies.

References:

[1] Bailey T.J.N. "The Mathematical Theory of Infectious Diseases", Charles Griffiths and Company, London 1975.

[2] O'Neill Ph. and Roberts G. "Bayesian inference for partially observed stochastic processes", Journal of the Royal Statistical Society, Series A, **162**, 121–129 (1999).

[3] Becker N. Analysis of infectious disease data. Chapman and Hall, New York 1989.

[4] O'Neil Ph. A tutorial introduction to Bayesian inference for stchastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences 2002; 180:103-114.

Practical:
Parameter estimation with data augmentation in the SIR model

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin
July 13 – July 15, 2015

Background

In this exercise we fit the general epidemic model to the Abakaliki smallpox data using Bayesian data augmentation. The data originate from a smallpox outbreak in a community of $M = 120$ initially susceptible individuals. There is one introductory case and 29 subsequent cases so that the total number of cases is $n = 30$. The observed 29 time intervals (Δ) between the n removals, i.e., between the detection of cases are:

$$13, 7, 2, 3, 0, 0, 1, 4, 5, 3, 2, 0, 2, 0, 5, 3, 1, 4, 0, 1, 1, 1, 2, 0, 1, 5, 0, 5, 5 \text{ (days)}.$$

A zero means that symptoms appeared the same day as for the preceding case. After the last removal there were no more cases. To fix the the time origin we assume that the introductory (index) case became infectious at time 0 and was removed at time 14 days (this appears as a long duration of infectiousness but agrees with the interpretation made in [1]). With this assumption, we can calculate the removal times $\boldsymbol{r}$ with respect to the time origin (see exercise 2 below). The total duration of the outbreak is $T = 90$ days ($= 14 + \sum_{i=1}^{29} \Delta_i$).

We explore the joint posterior distribution of the infection rate β and the removal rate γ. The unknown infection times $(i_2, \ldots, i_{30})$ are augmented, i.e., treated as additional model unknowns. All infection times together are denoted by $\boldsymbol{i}$.

The example program is implemented using *individual-based* event histories (see the lectures). The indices thus refer to individuals. In particular, (i_k, r_k) are the infection and removal times for the *same* individual k. This affects the choice of the likelihood function as explained in the lectures. The appropriate expression is:

$$\gamma^n \prod_{k=2}^{n} \{\beta I(i_k)\} \exp\left(- \int_0^T (\gamma I(u) + (\beta/M) I(u) S(u)) du \right).$$

In actual computations, it is more convenient to use the logarithm of the likelihood function:

$$n \log(\gamma) + (n-1) \log(\beta) + \sum_{k=2}^{n} \log I(i_k) - \int_0^T (\gamma I(u) + (\beta/M) I(u) S(u)) du.$$

N.B. The following is not intended to be a comprehensive analysis of the Abakaliki smallpox data. More appropriate analyses are possible. For example, in reference [2], the time of infection of the index case was included in the model unknowns. No adjustments were made to the original data. In [3], heterogeneity across individuals in their susceptibility to infection and a latent period were allowed.

Exercises

1. Download all required source codes by executing **SIRaugmentation_reduced.R**. The complete code will be provided once we have tried to complete the "reduced" version of the sampling routine (see below).

2. **Read the data.** The observed data in the Abakaliki smallpox outbreak include only the time intervals between removal times in the 30 infected individuals (therefore 29 intervals) and the fact that 90 individuals remained uninfected throughout the outbreak. Function **readdata.R** can be used to read in the time intervals of removals:

   ```
   intervals = readdata()
   ```

   The time intervals are in days. Note that the output vector does not include the piece of information that 90 individuals remained uninfected. This has to be input to the estimation routine separately (see below).

3. **Calculate the removal times.** The removal times can be calculated on the basis of the time intervals between them. This requires fixing a time origin. We make the assumption that the index case became infected at time $t = 0$ and was removed at time 14 (see above). These assumptions are "hardwired" in the program **removaltimes.R** (but can be changed easily for other contexts):

```
remtimes = removaltimes(intervals)
```

4. **Implementing the sampling algorithm.** The steps are

   (a) Reserve space for vectors of length K for the two model parameters β and γ (for an MCMC sample of size K; in the actual R code, K = mcmc.size). Samples of the unknown infections times need not be stored but a (vector) variable is needed to store the current iterates.

   (b) Initialise the model unknowns $\beta[1]$ and $\gamma[1]$. The unknown infection times need to be initialized as well. To do this, you can use routine **initializedata.R** which creates a complete data matrix with two columns (infection times and removal times). Each row corresponds to an infected individuals in the data; the index case is on the first row.

```
completedata = initializedata(remtimes)
```

   (c) Update β from its full conditional distribution in a Gibbs step:

   $$\beta[k+1] \mid \boldsymbol{i}[k-1], \boldsymbol{r} \ \sim \ \Gamma(n-1+\nu_\beta, (1/M)\int_0^T I(u)S(u)du + \lambda_\beta)$$

   (d) Update γ from its full conditional distribution in a Gibbs step:

   $$\gamma[k] \mid \boldsymbol{i}[k-1], \boldsymbol{r} \ \sim \ \Gamma(n+\nu_\gamma, \int_0^T I(u)du + \lambda_\gamma)$$

   (e) Update infection times $(i_2, \ldots, i_n)$ using Metropolis-Hastings steps (cf. the lecture). This creates a new vector of infection times $\boldsymbol{i}$[k] (the first element is always fixed by our assumption).

   (f) Repeat steps (c)–(e) K times, storing the samples $(\beta[k], \gamma[k])$, $k = 1, \ldots K$.

The sampling routine is implemented in **sampleSIR_reduced.R**. It requires as input the removal times ($\boldsymbol{r}$), the total number of individuals (M) and the number of iterations (K). The program uses a number of subroutines (with obvious tasks to perform): **initializedata.R**, **update_beta.R**, **update_gamma.R**, **update_inftimes.R**, **loglikelihood.R**, **totaltime_infpressure.R**, and **totaltime_infected.R**.

The subroutines **update_beta.R** and **update_gamma.R** are reduced, so your task is to complete those. These corresponds to steps (c) and (d) above.

5. **Sampling the posterior distribution.** Use the compeleted sampling routine (or **sampleSIR.R**) to realize an MCMC sample from the joint distribution of the model two parameters:

   ```
   mcmc.sample = sampleSIR(remtimes,M=120,mcmc.size=600)
   ```

   Plot the sample paths of the two model parameters (β and γ). For example, for parameter β:

   ```
   plot(mcmc.sample$beta,type="l",xlab="iteration",ylab="beta")
   ```

   Then explore the marginal and joint distributions of the model parameters.

6. **The effect of priors.** The program applied uninformative priors with $(\nu_\beta, \lambda_\beta) = (0.0001, 0.0001)$ and $(\nu_\gamma, \lambda_\gamma) = (0.0001, 0.0001)$ (see functions update_beta.R and update_gamma.R. Try how sensitive the posterior estimates are to a more informative choice of the prior, e.g. $(\nu_\beta, \lambda_\beta) = (10, 100)$ and $(\nu_\gamma, \lambda_\gamma) = (10, 100)$.

7. **The number of secondary cases.** What is the expected number of secondary cases for the index case, that is, calculate the posterior expectation of β/γ.

References:

[1] Becker N. Analysis of infectious diseases data. Chapman and Hall, 1989.

[2] O'Neill Ph. and Roberts G. Bayesian inference for partially observed stochastic processes. Journal of the Royal Statistical Society, Series A, **162**, 121–129 (1999).

[3] O'Neill Ph. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical Biosciences **180**, 103-114 (2002).

Practical: Convergence Diagnostics

Instructors: Kari Auranen, Elizabeth Halloran and Vladimir Minin

July 13 – July 15, 2015

Examining MCMC output in the chain-binomial Gibbs sampler

Here, we will have a look at some diagnostic tools provided in the R package "coda." Download the script "diagnostics.R" that examines convergence of the chain-binomial Gibbs sampler. We will go over this script during the practical.

Your task

Use "coda" package tools to examine convergence of either the beta-binomial (R script "beta_bin.R") or the hierarchical chain-binomial (R script "chain_hierarchical.R") Metropolis-within-Gibbs sampler.

Practical:
Data simulation and parameter estimation from complete data for a recurrent infection

Instructors: Kari Auranen, Elizabeth Halloran, Vladimir Minin
July 13 – July 15, 2015

Background

In the following exercises we try out Markov chain Monte Carlo methods in the Bayesian data analysis for recurrent infections. The model of infection is taken to be a binary Markov process, where at any given time the epidemiological state for an individual is either 0 (susceptible) or 1 (infected). This is the simplest stochastic "SIS" model (susceptible-infected-susceptible).

To familiarize ourselves with the computational approaches, using the Metropolis-Hastings algorithm with reversible jumps to augment unobserved events, we consider (statistically) independent individuals, omitting thus questions about transmission. This makes the likelihood computations easier and faster.

The binary Markov process is considered from time 0 to time T, at which the process is censored. The model has three parameters: (λ, μ, π), where λ is the per capita rate (force) of infection, μ is the rate of clearing infection and π is the proportion of those that are infected at time 0.

For N independent individuals, the *complete data* comprise the times $(T_{sr}^{(ik)})$ of all transitions between states 0 and 1 that occur between time 0 and the censoring time T (see lectures). In more realistic situations, however, we could not hope to observe complete data. Instead, the process can usually only be observed at some pre-defined times. To apply the complete data likelihood, unobserved event times and states should be augmented. The computations then rely on the reversible jump Markov chain Monte Carlo methodology. However, this problem falls outside the scope of the current exercise.

Exercises

1. **Simulation of complete (event-history) data.** Download the source code of an R function **simulateSIS_N.R**. Then simulate complete data from the binary Markov model ("susceptible-infected-susceptible"):

```
complete_data = simulateSIS_N(N=100,la=0.45,mu=0.67,initprob=0.40,T=12)
```

   The function samples binary processes for N=100 individuals from time 0 to time T=12 (time units). The transition rates are $\lambda = 0.45$ (force of infection, per time unit per capita) and $\mu = 0.67$ (rate of clearing infection, per time unit per capita). The proportion of those that are infected at time 0 is π=0.40 (initprob). The output is a list of N arguments, each containing the event times (times of transition) and the epidemiological states (after each transition) for one individual.

   These data might describe a 12 month follow-up of acquisition and clearance of nasopharyngeal carriage of pneumococci (a recurrent asymptomatic infection), with mean duration of carriage $1/\mu = 1.5$ months and the stationary prevalence of $\lambda/(\lambda + \mu) = 0.40$.

2. **Estimation of model parameters from completely observed data.** You can realize numerical samples from the joint posterior distribution of the three model parameters (λ, μ, π) with the R function **MH_SIS.R**. This function applies a component-wise Metropolis-Hastings algorithm to update each of the parameters in turn. It uses subroutines **likelihoodSIS.R** (to calculate values of the log-likelihood from the observed event histories) and **update_parameters.R** (to perform the actual updating). These routines are in the same source file as the main program.

   To perform M=1500 MCMC iterations, the program is called as follows:

```
par = MH_SIS(complete_data,M=1500)
```

   The output `par` is a list of three parameter vectors, each of length M. These are the MCMC samples from the joint posterior distribution of the model parameters.

(a) Plot the sample paths of each of the parameters. Does it appear that the sampling algorithm has converged? For the rate of acquisition, for example:

```
plot(par[[1]],type="l",xlab="iteration",ylab="rate of acquisition
(per mo)")
```

(b) Calculate the posterior mean and the 90% posterior intervals for the three model parameters. For example:

```
la_samples = par[[1]][501:1500]
la_samples2 = sort(la_samples)
mean(la_samples2)
la_samples2[50] # 5% quantile of the marginal posterior
la_samples2[950]# 95% quantile of the marginal posterior
```

(c) Are there any correlation between rates λ and μ in their joint posterior distribution? For a visual inspection, you can draw the scatter plot of the joint posterior:

```
la_samples = par[[1]][501:1500]
mu_samples = par[[2]][501:1500]
plot(la_samples,mu_samples,type='p')
```

(d) The rate parameters were given (independent) Gamma(ν_1, ν_2) priors with $\nu_1 = \nu_2 = 0.00001$ (see the program code in the subroutine update_parameters.R). With the amount of data, the analysis is quite robust to the choice of prior. However, try how the posterior is affected by a more informative choice of the prior distributions (e.g, by choosing hyperparameters $\nu_1 = 1$ and $\nu_2 = 20$) when $N = 10$.