

Bayesian Hierarchical Modeling:

From 0 to 90 in 90 minutes

SAMSI

Program on Environmental Sensor Networks

January 13, 2008

Jennifer A. Hoeting

Associate Professor

Department of Statistics

Colorado State University

Outline

1. Introduction to Bayesian inference	Page 3–10
2. Bayesian computation	12–53
(a) Brief overview of Markov chains	14–15
(b) Metropolis-Hastings algorithm	17– 27
(c) Gibbs sampling	28–30
(d) Example: population size estimation	31–40
(e) MCMC Implementation issues	41–53
3. Bayesian Hierarchical Modeling	54–56
4. Example: Bayesian Models for Capture-Recapture Data	57–73

These notes were developed from the book

COMPUTATIONAL STATISTICS
by G. H. Givens and J. A. Hoeting
available from Wiley

Book web page: www.stat.colostate.edu/computationalstatistics/

Bayesian Statistics

Bayesian inference

- **Basic Idea:** Fit a probability model to data
- **Goal:** probability distribution on observable and unobservable quantities

Some history

- **1763:** Paper by Rev. Thomas Bayes
- **1950's:** Researchers advocated Bayesian methods to remedy deficiencies with classical inference (Savage, de Finetti, Lindley)
- **1950's–1980's:** Controversy and computational roadblocks
- **1980's:** Markov chain Monte Carlo (MCMC) methods were recognized to be useful for Bayesian inference
- **1990's:** Considerable development in MCMC and software
- **Now:** Everybody's doing it

Some notation and review

- **observed data:** $\mathbf{y} = (y_1, \dots, y_n)$
- **parameters:** $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ (unobservable)
- **likelihood function:** If \mathbf{y} is fixed and $\boldsymbol{\theta}$ are the parameters, then call $f(\mathbf{y}|\boldsymbol{\theta})$ the likelihood function

Bayes' Theorem

$$f(\theta|\mathbf{y}) = \frac{f(\theta, \mathbf{y})}{f(\mathbf{y})} \quad (1)$$

$$= \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})} \quad (2)$$

$$= \frac{f(\mathbf{y}|\theta)f(\theta)}{\int f(\mathbf{y}|\theta)f(\theta)d\theta} \quad (3)$$

$$f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)f(\theta) \quad (4)$$

Prior distributions

The prior distribution should include all possible values of a parameter.

Some examples of types of prior distributions:

- conjugate: the posterior distribution is a member of the same distributional family as the prior.
- noninformative: Contains no information about the parameter. For example, “ $\theta \sim \text{Uniform}(-\infty, \infty)$ ”. Noninformative priors are popular, but care needs to be taken to insure that the posterior is a proper distribution.
- Proper distribution with large variance.

A common approach for priors for multiple parameters: assume a univariate prior for each parameter and assume independence between parameters *a priori*

Example: Sex ratio for wildlife population

Goal: Estimate the sex ratio for a newly discovered population of prairie dogs

y is the observed number of females in n recorded births

θ is the proportion of female births

Assume that y follows a Binomial distribution, so

$$f(y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$$

Prior for θ : Assume $\theta \sim \text{Uniform}(0,1)$

Applying Bayes' rule we see

$$f(\theta|y) \propto \theta^y (1 - \theta)^{n-y}$$

Can recognize that this the unnormalized form of the beta density, so $\theta|y \sim \text{Beta}(y + 1, n - y + 1)$

Posterior mean $E(\theta|y) = \frac{y+1}{n+2}$ is a compromise between the sample proportion $\frac{y}{n}$ and the prior mean $E(\theta) = \frac{1}{2}$

Bayesian inference

Posterior distribution

- The posterior summarizes everything we know about the model parameters in light of the data.
- Posterior distribution represents a compromise between the prior information and the data (likelihood function). For a given prior distribution, the prior should become less influential as sample size increases.

Point estimation: Use the posterior mean $E(\theta|\mathbf{y})$ and variance $\text{Var}(\theta|\mathbf{y})$.

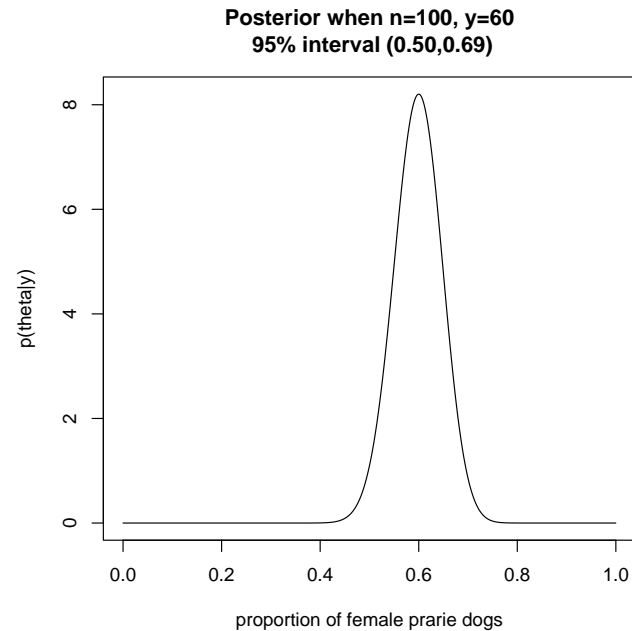
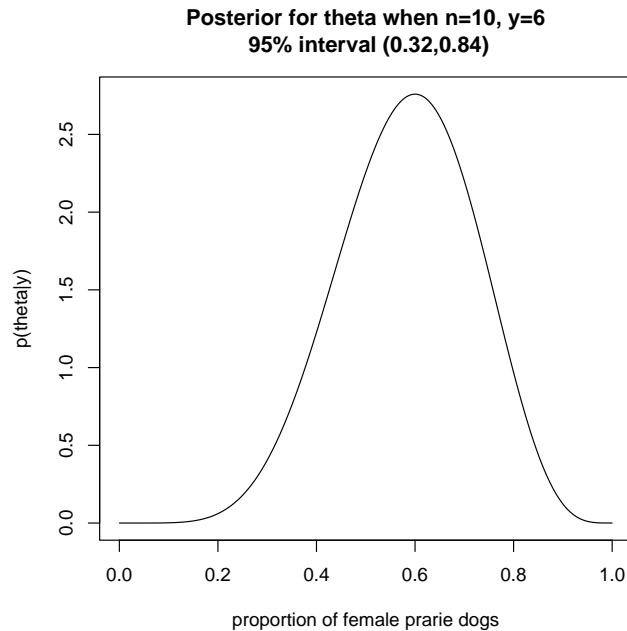
Interval estimation: For a symmetric unimodal posterior distribution, can construct a $(1-\alpha)$ interval as follows:
Find the points a and b such that

$$\int_{-\infty}^a f(\theta|\mathbf{y})d\theta = \alpha/2 \quad \text{and} \quad \int_b^{\infty} f(\theta|\mathbf{y})d\theta = 1 - \alpha/2$$

Interpretation:

The probability that θ lies in the interval (a, b) is $(1-\alpha)$.

Example, continued



When $n=100$, $y=60$, the probability is 0.95 that the proportion of females in the population, θ , is between 0.50 and 0.69.

One approach to Bayesian data analysis

1. Do exploratory data analysis (EDA)
2. Choose prior and likelihood
3. Derive and/or draw samples from the posterior distribution
4. Evaluate the fit of the model
 - Do model diagnostics (does the model fit the data?)
 - Do MCMC diagnostics (has your sampler converged?)
 - Do sensitivity analyses (is prior influential?)
5. Iteratively repeat steps 1–4, as needed
6. Make inferences

Outline

1. Introduction to Bayesian inference	Page 3–10
2. Bayesian computation	12–53
(a) Brief overview of Markov chains	14–15
(b) Metropolis-Hastings algorithm	17– 27
(c) Gibbs sampling	28–30
(d) Example: population size estimation	31–40
(e) MCMC Implementation issues	41–53
3. Bayesian Hierarchical Modeling	54–56
4. Example: Bayesian Models for Capture-Recapture Data	57–73

A note about notation

The following notes on MCMC will use different notation than used above. MCMC methods aren't just for Bayesian methods!

	Previously	MCMC notes
Parameters	$(\theta_1, \dots, \theta_p)$	$\mathbf{X} = (X_1, \dots, X_p)$
Random variable		X
Realized values		x
Posterior distribution	$f(\theta y)$	Often just $f(x)$ Sometimes more precise

Recall that if X is a random variable with distribution function f , then the expectation of X is given by

$$E\{X\} = \int_{-\infty}^{\infty} x f(x) dx$$

Monte Carlo Integration: To approximate $\mu = E\{\mathbf{X}\} = \int \mathbf{x} f(\mathbf{x}) d\mathbf{x}$, obtain an i.i.d. random sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ from f and then approximate μ via the sample average:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$$

Markov chain Monte Carlo

- Markov chain Monte Carlo (MCMC) methods introduced below can be used to generate a draw from a distribution that approximates some target distribution $f(\cdot)$
- In Bayesian applications of MCMC, the target distribution is usually the posterior distribution.
- The basic idea is that we know that $\mathbf{X} \sim f(\mathbf{x})$, but we can't get $f(\mathbf{x})$ in closed form or we can't simulate easily from $f(\mathbf{x})$.
- The MCMC methods introduced below can be used to generate a draw from a distribution that approximates f , but they are more properly viewed as methods for generating a sample from which expectations of functions of \mathbf{X} can reliably be estimated.

Brief overview of Markov chains

Let $\{X^{(t)}\}$, $t = 0, 1, \dots$ be a sequence of random variables where t indexes time (iteration number)

A complete probabilistic specification of $X^{(0)}, \dots, X^{(n)}$ would be to write their joint distribution as the product of conditional distributions of each random variable given its history, or

$$P \left[X^{(0)}, \dots, X^{(n)} \right] = P \left[X^{(n)} \mid x^{(0)}, \dots, x^{(n-1)} \right] P \left[X^{(n-1)} \mid x^{(0)}, \dots, x^{(n-2)} \right] \times \dots \\ \times P \left[X^{(1)} \mid x^{(0)} \right] P \left[X^{(0)} \right]$$

The *Markov property* states that

$$P \left[X^{(t)} \mid x^{(0)}, \dots, x^{(t-1)} \right] = P \left[X^{(t)} \mid x^{(t-1)} \right]$$

This allows a simpler form of the joint distribution:

$$P \left[X^{(0)}, \dots, X^{(n)} \right] = P \left[X^{(n)} \mid x^{(n-1)} \right] P \left[X^{(n-1)} \mid x^{(n-2)} \right] \dots P \left[X^{(1)} \mid x^{(0)} \right] P \left[X^{(0)} \right]$$

Brief overview of Markov chains: Ergodic theorem

If $X^{(1)}, X^{(2)}, \dots$ are realizations from an irreducible and aperiodic Markov chain with stationary distribution π , then

1. $X^{(n)}$ converges in distribution to the distribution given by π
2. For any function h

$$\frac{1}{n} \sum_{t=1}^n h(X^{(t)}) \rightarrow E_{\pi}\{h(X)\}$$

almost surely as $n \rightarrow \infty$, provided $E_{\pi}\{|h(X)|\}$ exists.

This is one form of the *ergodic theorem*, which is a generalization of the strong law of large numbers.

Basic summary: If $X^{(1)}, X^{(2)}, \dots$ are realizations from a Markov chain with stationary distribution f , then the mean of these realizations approximates the mean of the target distribution

Markov chain Monte Carlo

- Let $\{\mathbf{X}^{(t)}\}$ denote a Markov chain for $t = 0, 1, 2, \dots$, where $\mathbf{X}^{(t)} = \left(X_1^{(t)}, \dots, X_p^{(t)} \right)$
- The MCMC sampling strategy is to construct (an irreducible, aperiodic) Markov chain for which the stationary distribution equals the target distribution, f .
- There are various strategies for constructing a MCMC algorithm including the Metropolis-Hastings algorithm, the Gibbs sampler, etc.
- A popular application of MCMC methods is to facilitate Bayesian inference where f is a Bayesian posterior distribution for parameters \mathbf{X}

Metropolis-Hastings algorithm

The method begins at $t = 0$ with the selection of $\mathbf{X}^{(0)} = \mathbf{x}^{(0)}$ drawn at random from some starting distribution, g , with the requirement that $f(\mathbf{x}^{(0)}) > 0$.

Given $\mathbf{X}^{(t)} = \mathbf{x}^{(t)}$, the algorithm generates $\mathbf{X}^{(t+1)}$ as follows.

1. Sample a candidate value, \mathbf{X}^* , from a *proposal distribution* $g(\cdot | \mathbf{x}^{(t)})$.
2. Compute the *Metropolis-Hastings ratio*, $R(\mathbf{x}^{(t)}, \mathbf{X}^*)$, where

$$R(\mathbf{u}, \mathbf{v}) = \frac{f(\mathbf{v}) g(\mathbf{u} | \mathbf{v})}{f(\mathbf{u}) g(\mathbf{v} | \mathbf{u})}. \quad (5)$$

3. Sample a value for $\mathbf{X}^{(t+1)}$ according to the following:

$$\mathbf{X}^{(t+1)} = \begin{cases} \mathbf{X}^* & \text{with probability } \min\{R(\mathbf{x}^{(t)}, \mathbf{X}^*), 1\} \\ \mathbf{x}^{(t)} & \text{otherwise.} \end{cases} \quad (6)$$

4. Increment t and return to step 1.

Metropolis-Hastings algorithm

Important note:

The sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots$ will likely include multiple copies of some points in the state space.

It is important to include these copies in the chain and in any estimates based on the output, since the frequencies of sampled points are used to correct for the fact that the proposal density differs from the target density.

Metropolis-Hastings algorithm

The distribution of realizations from the Metropolis-Hastings chain approximates the stationary distribution of the chain as t progresses, therefore

$$\mathbf{E} \{h(\mathbf{X})\} \approx \frac{1}{n} \sum_{i=1}^n h(\mathbf{x}^{(i)})$$

Some useful quantities that can be estimated include

- means: $\mathbf{E} \{h(\mathbf{X})\}$
- variances: $\mathbf{E} \{ (h(\mathbf{X}) - \mathbf{E} \{h(\mathbf{X})\})^2 \}$
- tail probabilities: $\mathbf{E} \{ 1_{\{h(\mathbf{X}) \leq q\}} \}$ for constant q , where $1_{\{A\}} = 1$ if A is true and 0 otherwise
- Virtually any other statistic that may be of interest!

Can also estimate f itself using density estimation methods

Independence chains

Suppose that the proposal distribution for the Metropolis-Hastings algorithm is chosen such that for some fixed density g

$$g(\mathbf{x}^* | \mathbf{x}^{(t)}) = g(\mathbf{x}^*)$$

This yields an independence chain, where each candidate value is drawn independently of the past.

In this case, the Metropolis-Hastings ratio is

$$R(\mathbf{x}^{(t)}, \mathbf{X}^*) = \frac{f(\mathbf{X}^*) g(\mathbf{x}^{(t)})}{f(\mathbf{x}^{(t)}) g(\mathbf{X}^*)}$$

The resulting Markov chain is irreducible and aperiodic if $g(\mathbf{x}) > 0$ whenever $f(\mathbf{x}) > 0$.

Bayesian inference and MCMC

In the Bayesian paradigm, parameters are assumed to follow some probability distribution. Suppose that data \mathbf{Y} have a distribution parameterized by θ .

Bayes Theorem:

$$\begin{aligned} f(\theta|\mathbf{y}) &= \frac{f(\theta)f(\mathbf{y}|\theta)}{\int f(\theta)f(\mathbf{y}|\theta) d\theta} \\ &= \frac{f(\theta)f(\mathbf{y}|\theta)}{f(\mathbf{y})} \\ &\propto f(\theta)f(\mathbf{y}|\theta) \end{aligned}$$

where

- $f(\theta|\mathbf{y})$ is the *posterior density* of θ
- $f(\mathbf{y}|\theta)$ is the *likelihood* or *sampling distribution*
- $f(\theta)$ is the *prior density*

In Bayesian applications we use MCMC to simulate from the posterior distribution $f(\theta|\mathbf{y})$, so the target distribution is typically the posterior distribution.

Bayesian inference and MCMC

Simple strategy: use the prior as a proposal distribution in an independence chain.

In our Metropolis-Hastings notation,

$$f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{y}) \quad \text{target distribution}=\text{posterior distribution}$$

and

$$g(\boldsymbol{\theta}^*) = f(\boldsymbol{\theta}^*) \quad \text{proposal distribution}=\text{prior distribution}$$

Then the Metropolis-Hastings ratio is given by:

$$R(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^*) = \frac{f(\mathbf{y}|\boldsymbol{\theta}^*)}{f(\mathbf{y}|\boldsymbol{\theta}^{(t)})}$$

Comments:

- If proposal distribution= prior distribution, the Metropolis-Hastings ratio equals the likelihood ratio.
- More sophisticated algorithms often have better performance.

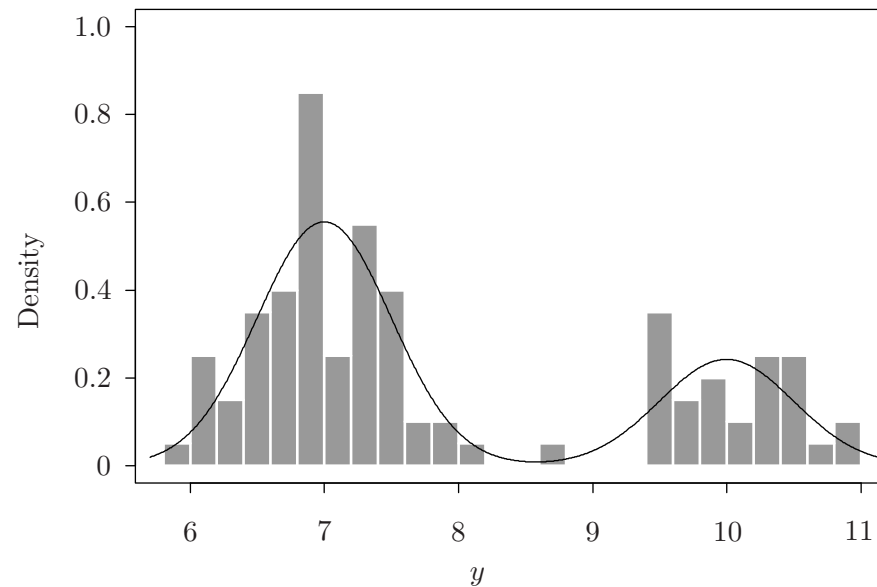
MCMC Example: Estimating a mixture parameter

Suppose we have observed data y_1, y_2, \dots, y_{100} sampled independently and identically distributed from the mixture distribution

$$\delta N(7, 0.5^2) + (1 - \delta)N(10, 0.5^2) \quad (7)$$

- Mixture densities are common in real-life applications where, for example, the data may come from more than one population
- We will use MCMC techniques to construct a chain whose stationary distribution equals the posterior density of δ assuming a $\text{Unif}(0,1)$ prior distribution for δ

MCMC Example: Estimating a mixture parameter



Histogram of 100 observations simulated from the mixture distribution (7). The data were generated with $\delta = 0.7$, so we should find that the posterior density is concentrated in this area.

MCMC Example: Estimating a mixture parameter

Try two different MCMC proposal densities:

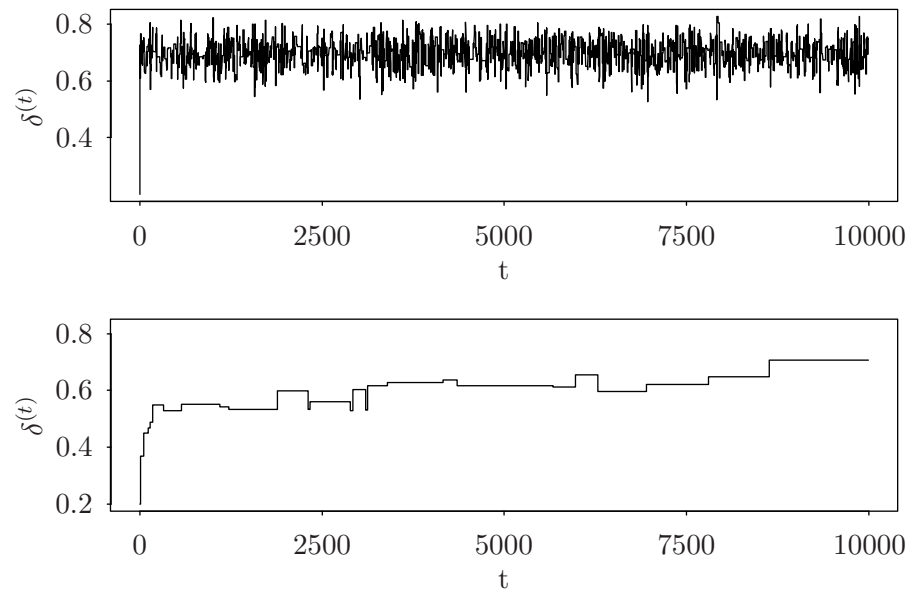
1. Beta(1,1): equivalent to a Unif(0,1) distribution
2. Beta(2,10):
 - skewed right with mean of 0.167
 - values of δ near 0.7 are unlikely to be generated from the proposal distribution

Generated 10,000 iterations of Metropolis-Hastings algorithm for each proposal.

Some output:

- Sample paths:
 - a plot of the chain realizations, $\delta^{(t)}$, against the iteration number, t
 - useful for investigating the behavior of the Markov chain
- Histogram of realizations

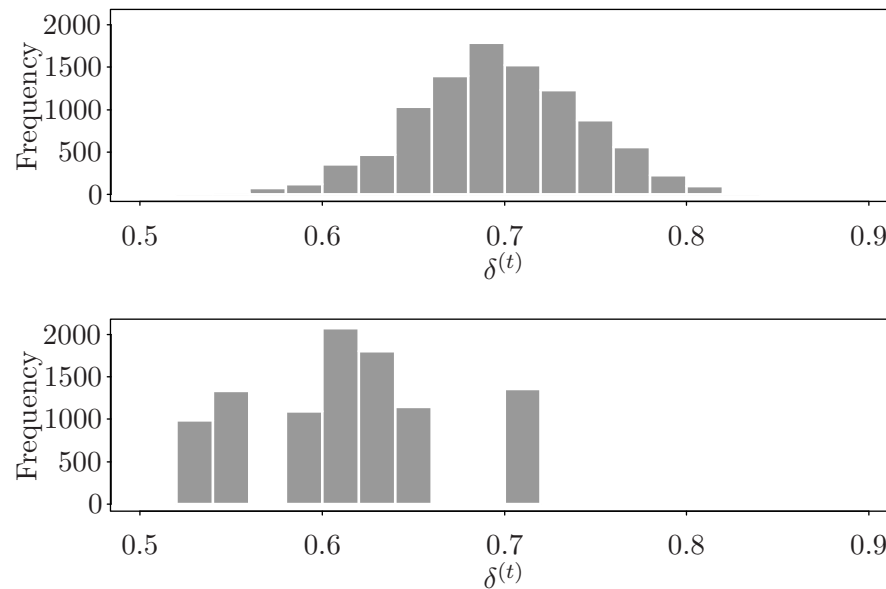
MCMC Example: Estimating a mixture parameter



Sample paths for δ from independence chains with proposal densities Beta(1,1) (top) and Beta(2,10) (bottom)

Burn-in period: Since initial iterates retain strong dependence on the starting value, eliminate them in estimation

MCMC Example: Estimating a mixture parameter



Histograms of $\delta^{(t)}$ for iterations 201-10,000 of independence chains with proposal densities Beta(1,1) (top) and Beta(2,10) (bottom)

Gibbs Sampling

- Specifically adapted for multidimensional target distributions.
- Goal: still to construct a Markov chain whose stationary distribution—or some marginalization thereof—equals the target distribution, f
- Sequentially samples from univariate conditional distributions of f which are often available in closed form

Basic Gibbs sampler

Suppose it is easy to sample from the univariate conditional distributions:

$$X_i | X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_p \sim f(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p)$$

A basic Gibbs sampler:

1. Select starting values $\mathbf{x}^{(0)}$ and set $t = 0$.
2. Generate, in turn,

$$X_1^{(t+1)} | \cdot \sim f(x_1 | x_2^{(t)}, \dots, x_p^{(t)})$$

$$X_2^{(t+1)} | \cdot \sim f(x_2 | x_1^{(t+1)}, x_3^{(t)}, \dots, x_p^{(t)})$$

⋮

$$X_{p-1}^{(t+1)} | \cdot \sim f(x_{p-1} | x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{p-2}^{(t+1)}, x_p^{(t)})$$

$$X_p^{(t+1)} | \cdot \sim f(x_p | x_1^{(t+1)}, x_2^{(t+1)}, \dots, x_{p-1}^{(t+1)})$$

where $|\cdot$ denotes conditioning on the most recent updates to all other elements of \mathbf{X} .

3. Increment t and go to step 2.

Hybrid Gibbs sampling

A *hybrid MCMC* algorithm might proceed with the following sequence of updates with $p = 6$:

1. Update $X_1^{(t+1)} \mid \left(x_2^{(t)}, x_3^{(t)}, x_4^{(t)}, x_5^{(t)}, x_6^{(t)} \right)$ with a Gibbs step.
2. Update $\left(X_2^{(t+1)}, X_3^{(t+1)} \right) \mid \left(x_1^{(t+1)}, x_4^{(t)}, x_5^{(t)}, x_6^{(t)} \right)$ with a Metropolis step.
3. Update $X_4^{(t+1)} \mid \left(x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}, x_5^{(t)}, x_6^{(t)} \right)$ with a step from a random walk chain.
4. Update $\left(X_5^{(t+1)}, X_6^{(t+1)} \right) \mid \left(x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t+1)}, x_4^{(t+1)} \right)$ with a Gibbs step.

The Metropolis-Hastings steps within a Gibbs algorithm are typically useful when the univariate conditional density for one or more elements of \mathbf{X} is not available in closed form.

Example: Fur seal pup capture-recapture analysis

- After centuries of severe population reductions due to commercial and subsistence hunting, the abundance of fur seals in New Zealand has been increasing in recent years.
- Our goal is to estimate the number of pups in a fur seal colony using a capture-recapture approach.
- Estimation: can estimate population size and survival rates, etc.
- Consider, for example, high recapture rates suggest that the true population size does not greatly exceed the total number of unique individuals ever captured.

Example: Fur seal pups

Let

- N be the unknown population size to be estimated
- I be the number of census attempts
- $\mathbf{c} = (c_1, \dots, c_I)$ be the total number of captures (including recaptures) at each attempt
- r be the total number of distinct animals captured during the study

We assume that the population is closed during the period of the sampling which means that deaths, births and migrations are inconsequential during this period.

Example: Fur seal pups

- We consider a model with separate, unknown capture probabilities for each census effort, $\alpha = (\alpha_1, \dots, \alpha_I)$.
- This model assumes that all animals are equally catchable on any one capture occasion, but capture probabilities may change over time.

The likelihood for this model is

$$L(N, \boldsymbol{\alpha} | \mathbf{c}, r) \propto \frac{N!}{(N-r)!} \prod_{i=1}^I \alpha_i^{c_i} (1 - \alpha_i)^{N-c_i} \quad (8)$$

This model is sometimes called the $M(t)$ model

Example: Fur seal pups

- In a study conducted on the Otago Peninsula on the South Island of New Zealand, fur seal pups were marked and released during $I = 7$ census attempts during one season.
- It is reasonable to assume the population of pups was closed during the study period.
- $r = \sum_{i=1}^7 m_i = 84$ unique fur seals were observed during the sampling period.

Number of pups captured and recaptured during 7 census efforts in one season

		Census attempt, i						
		1	2	3	4	5	6	7
Number captured	c_i	30	22	29	26	31	32	35
Number newly caught	m_i	30	8	17	7	9	8	5

Example: Fur seal pups

For estimation, one might adopt a hierarchical Bayesian framework where N and α are assumed to be a priori independent with the following priors:

- N : noninformative Jeffreys prior $f(N) \propto 1/N$
- capture probabilities: $f(\alpha_i | \theta_1, \theta_2) = \text{Beta}(\theta_1, \theta_2)$ for $i = 1, \dots, 7$, assumed to be a priori exchangeable

Previous analyses with the $M(t)$ model have indicated sensitivity to the prior distribution for the capture probabilities. To mitigate this sensitivity, we introduce a hyperprior for (θ_1, θ_2) :

$$f(\theta_1, \theta_2) \propto \exp \{ -(\theta_1 + \theta_2)/1000 \},$$

with (θ_1, θ_2) assumed to be a priori independent of the remaining parameters.

Example: Fur seal pups

Goal:

$$f(N, \alpha_1, \dots, \alpha_7, \theta_1, \theta_2 | \mathbf{c}, r) \propto f(\mathbf{c}, \mathbf{m} | N, \alpha_1, \dots, \alpha_7, \theta_1, \theta_2) \\ f(N) \left[\prod_{i=1}^7 f(\alpha_i | \theta_1, \theta_2) \right] f(\theta_1, \theta_2)$$

A Gibbs sampler can then be constructed by simulating from the conditional posterior distributions

$$N - 84 | \cdot \sim \text{NegBin} \left(84, 1 - \prod_{i=1}^7 (1 - \alpha_i) \right) \\ \alpha_i | \cdot \sim \text{Beta}(c_i + \theta_1, N - c_i + \theta_2) \quad \text{for } i = 1, \dots, 7 \\ \theta_1, \theta_2 | \cdot \sim k \left[\frac{\Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} \right]^7 \prod_{i=1}^7 \alpha_i^{\theta_1} (1 - \alpha_i)^{\theta_2} \exp\{-(\theta_1 + \theta_2)/1000\}$$

where $|\cdot$ denotes conditioning on the remaining parameters and the data, and k is an unknown constant.

Example: Fur seal pups

Problem: it is difficult to produce a chain for (θ_1, θ_2) with adequate mixing and convergence behavior.

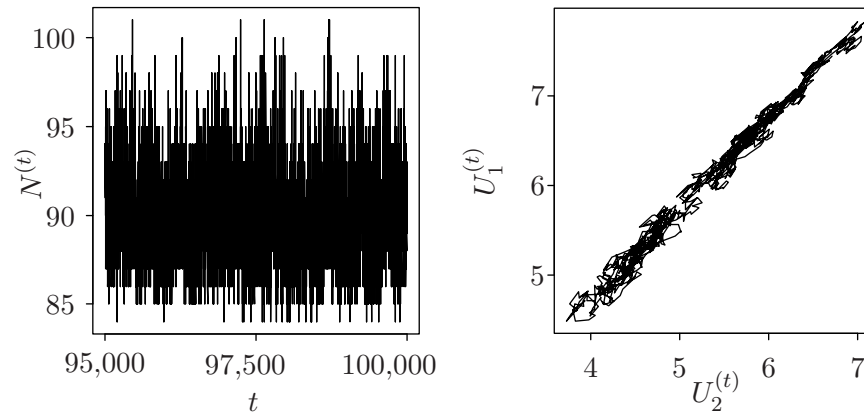
To improve performance, transform (θ_1, θ_2) to

$$\mathbf{U} = (U_1, U_2) = (\log \theta_1, \log \theta_2).$$

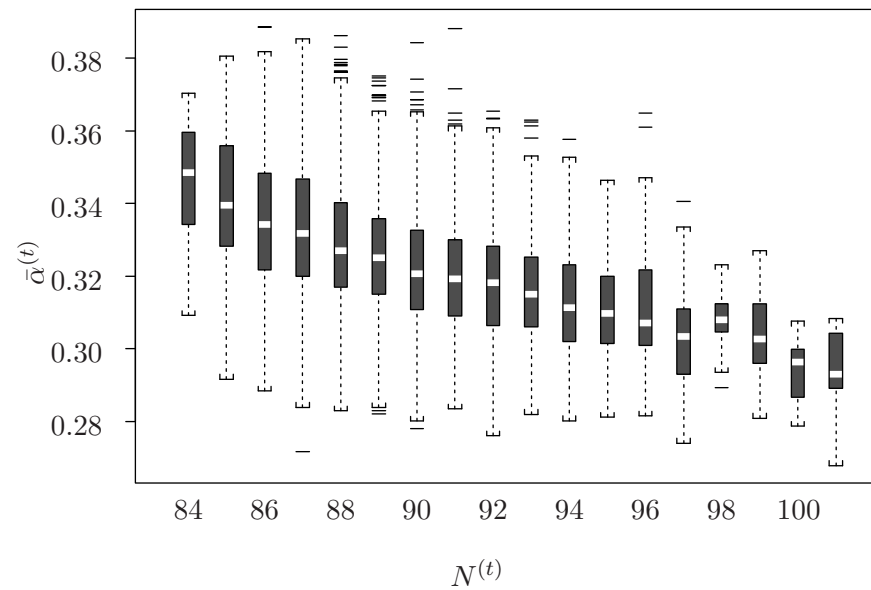
This requires transformation of the conditional densities.

Example: Fur seal pups

- The results below are based on a chain of 100,000 iterations with the first 1,000 iterations discarded for burn-in.
- Based on five runs of 100,000 iterations each, the Gelman and Rubin statistic for N is equal to 1.00047 which suggests the $N^{(t)}$ chain is roughly stationary.
- Sample paths for N (left panel) and U (right panel) for final 5,000 iterations in the seal pup example.

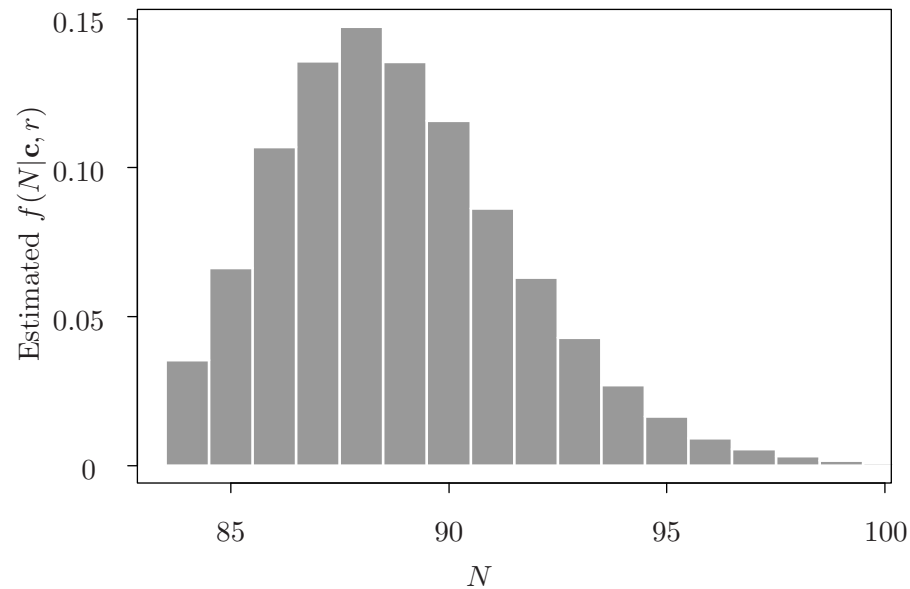


Example: Fur seal pups



Split boxplot of $\bar{\alpha}^{(t)}$ against $N^{(t)}$ for the seal pup example.

Example: Fur seal pups



Estimated marginal posterior probabilities for N for the seal pup example.

The posterior mean of N is 90 with a 95% highest posterior density interval of (84, 95).

MCMC algorithm implementation

All MCMC algorithms described above have the correct limiting stationary distribution.

Mixing:

- How quickly does the chain forget its starting value?
- How quickly does the chain fully explore the support of the target distribution?
- How far apart in a sequence do observations need to be before they can be considered to be approximately independent?

Stationarity: has the chain has run sufficiently long so that

- it is reasonable to believe that the output adequately represents the target distribution?
- the output can be used reliably for estimation?

Ensuring good mixing and convergence

Strategies and diagnostics

1. Choice of proposal
2. Number of chains
3. Reparameterization
4. Burn-in and run length
5. Simple graphs to assess mixing and convergence

1. Choice of proposal

Gibbs: performance is often enhanced when components of \mathbf{X} are as independent as possible.

Metropolis-Hastings: we wish the proposal distribution, g , to approximate the target distribution, f , very well

- The tail behavior of g is more important than the resemblance of f in regions of high density.
- If f/g is bounded, the convergence of the Markov chain to its stationary distribution is faster overall.
- Choose a proposal distribution that is somewhat more diffuse than f .

2. Number of chains

A difficult problem to diagnose: chain has become stuck in one or more modes of the target distribution.

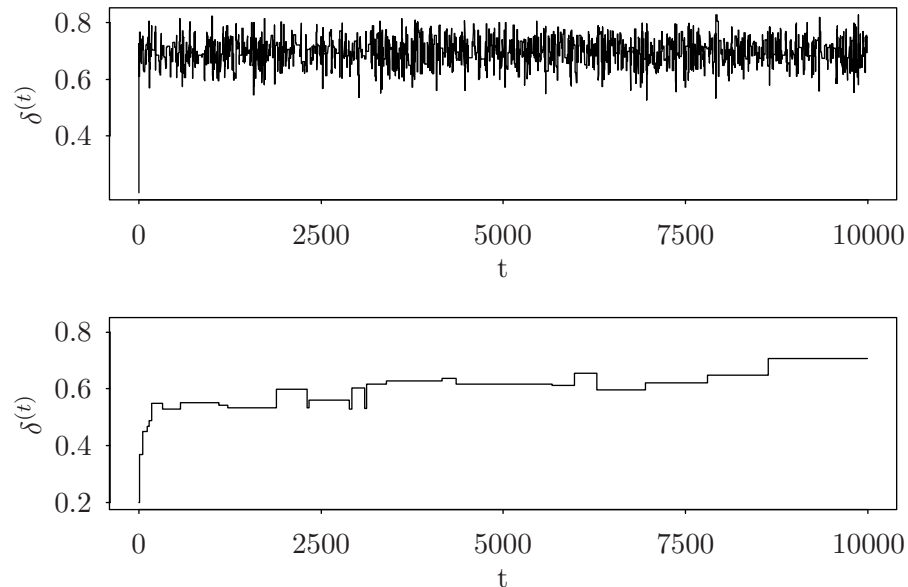
Partial solution: run multiple chains from diverse starting values and then compare the within- and between-chain behavior.

Diagnostics: sample path plots and Gelman-Rubin statistic

3. Simple graphs to assess mixing and convergence

Sample paths: plot of the iteration number versus the realizations of $X^{(t)}$ for $t = 0, 1, \dots$

- also called trace or history plots
- Want to see a very wiggly line that (perhaps) moves away from the starting value.



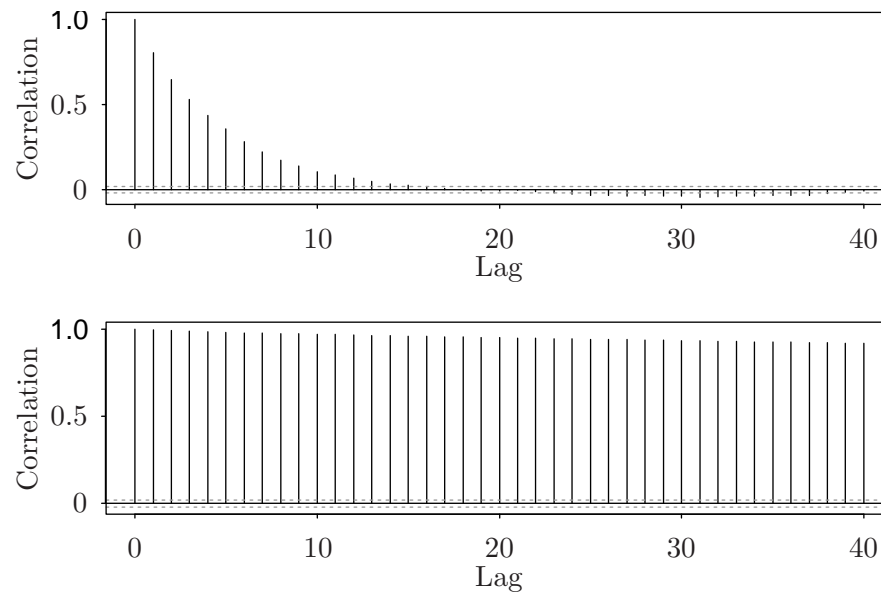
Sample paths for δ from independence chains with proposal densities Beta(1,1) (top) and Beta(2,10) (bottom)

3. Simple graphs to assess mixing and convergence

autocorrelation function (ACF) plot:

- A plot of the lag versus the correlation between iterates at that lag.
- Slow decay in the acf suggests poor mixing.
- Multi-parameter problems:
 - examine cross-correlations between parameters that might be related
 - High cross-correlations may indicate poor mixing of the chain

MCMC Example: Estimating a mixture parameter



Autocorrelation plots for example on estimating a mixture parameter with with proposal densities Beta(1,1) (top) and Beta(2,10) (bottom).

4. Reparameterization

- The mixing rates of both the Gibbs sampler and Metropolis-Hastings algorithms can be improved via reparameterization of the model.
- High correlation between the elements of \mathbf{X} can lead to slow convergence. Reparameterization of the model can reduce the correlation and thus speed convergence.

Reparameterization approaches are typically adapted for specific models:

- If there are continuous covariates in a linear model, center and scale the covariates to reduce correlations between the parameters in the model.
- Hierarchical centering: parameters in a hierarchical Bayesian model are moved down the hierarchy. Particularly useful in models with random effects.

5. Burn-in and run length

Burn-in

- Initial MCMC iterates will not have exactly the correct marginal distribution, and the dependency on the initial point (or distribution) from which the chain was started may remain strong.
- To reduce the severity of this problem, the first several thousand values from the chain are typically discarded.

Gelman and Rubin statistic, R

- Compares the output of several chains
- If within and between chain variances are similar in magnitude, suggests that the chains are stationary
- Roughly: want $\sqrt{R} < 1.2$
- If not: increase run-length, increase burn-in, reparameterize the model, or change the MCMC strategy

MCMC software

WinBUGS:

- WinBUGS is free software to carry out Gibbs sampling (www.mrc-bsu.cam.ac.uk/bugs/)
- WinBUGS can carry out Gibbs sampling for many models, but difficult problems can require alternative software

Other software available in R and other languages, but WinBUGS is the current popular choice

Bottom line: Programming your own MCMC algorithm is the best way to learn what is going on in MCMC, but WinBUGS makes life a lot easier.

Output of MCMC procedures

Marginalization: If $\{\mathbf{X}^{(t)}\}$ represents a p -dimensional Markov chain, then the limiting distribution of $\{X_i^{(t)}\}$ is the i th marginal of f . If you are focused only on a property of this marginal, discard the rest of the simulation and analyze the realizations of $X_i^{(t)}$.

Mean: The most commonly used estimator is based on an empirical average. Discard the burn-in; then calculate the desired statistic by taking

$$\frac{1}{L} \sum_{t=D}^{D+L-1} h(\mathbf{X}^{(t)})$$

Standard deviation: sample standard deviation of realizations (after burn-in)

Probability estimates: The probability of any event can be estimated by the frequency of that event in the chain.

Output of MCMC procedures

Monte Carlo standard error: There are several ways to compute this.

Batch Method:

1. Separate the realizations into batches with, say, 50 consecutive iterations in each batch.
2. Compute the mean of each batch.
3. The estimated standard error is the standard deviation of these means divided by the square root of the number of batches.

Density/histogram:

- Histogram of the raw iterates.
- A kernel density estimate of the marginal posterior distributions.

Quantiles: Compute empirical quantiles

Other summary statistics: Be creative!

Practical implementation advice

- **Mixing:** For a Metropolis algorithm with normal target and proposal distributions, it has been suggested that a user should aim for acceptance rates of approximately
 - 45% for one- or two-dimensional problems
 - 23% for higher-dimensional problems
- **Burn-in lengths:** 0-50,000
- **Chain lengths:** 5,000-5,000,000
 - **One rule of thumb:** the simulation should be run until the Monte Carlo error for each parameter of interest is less than 5% of the sample standard deviation.
 - More samples means more accurate estimates of the posterior distribution.
 - Chain lengths will continue to increase as computing power grows.

Outline

1. Introduction to Bayesian inference	Page 3–10
2. Bayesian computation	12–53
(a) Brief overview of Markov chains	14–15
(b) Metropolis-Hastings algorithm	17– 27
(c) Gibbs sampling	28–30
(d) Example: population size estimation	31–40
(e) MCMC Implementation issues	41–53
3. Bayesian Hierarchical Modeling	54–56
4. Example: Bayesian Models for Capture-Recapture Data	57–73

Bayesian Hierarchical Modeling

What if information is available on several levels?

Example: I want to model Chronic Wasting disease for deer in Colorado. I have

- **Data model:** likelihood function linking observations of disease presence/absence (y) to observation model parameters (θ) and process model parameters (ϕ)
- **Process model:** physical model about process of disease transmission depending on process model parameters (ϕ)
- **Parameter model(s):** model for parameters from data model and process model

This leads to a posterior distribution about the parameters of interest:

$$p(\theta, \phi | y) \propto p(y | \phi, \theta) p(\phi | \theta) p(\theta)$$

Note: This view of Bayesian statistics is often attributed to M. Berliner (1996)

One Example of a Process Model

For the chronic wasting disease problem, we might adopt a matrix population projection models (e.g., Casewell, 2001, Oli et al., 2006). Let

$$N_t = \begin{bmatrix} \# \text{ susceptibles at time } t \\ \# \text{ infected at time } t \\ \# \text{ recovered at time } t \end{bmatrix} = \begin{bmatrix} S_t \\ I_t \\ R_t \end{bmatrix}$$

Then the total population size at time t is given by $S_t + I_t + R_t$. The discrete-time dynamic population model can be written:

$$\begin{aligned} N_{t+1} &= AN_t \\ &= \begin{bmatrix} F_S + (1 - \beta)p_S & F_I & F_R \\ p_S\beta & (1 - \gamma)p_I & 0 \\ 0 & \gamma p_I & p_R \end{bmatrix} \begin{bmatrix} S_t \\ I_t \\ R_t \end{bmatrix} \end{aligned}$$

where

- F_S , F_I , and F_R are the fertility rates for each group
- Susceptible individuals survive with probability p_S and become infective with probability β
- Infective individuals survive with probability p_I and become infective with probability γ

Outline

1. Introduction to Bayesian inference	Page 3–10
2. Bayesian computation	12–53
(a) Brief overview of Markov chains	14–15
(b) Metropolis-Hastings algorithm	17– 27
(c) Gibbs sampling	28–30
(d) Example: population size estimation	31–40
(e) MCMC Implementation issues	41–53
3. Bayesian Hierarchical Modeling	54–56
4. Example: Bayesian Models for Capture-Recapture Data	57–73

Example: Autoregressive Models for Capture-Recapture Data

Model and methods developed by Jennifer Hoeting and Devin Johnson

Paper: Johnson, D. S. and J. A. Hoeting (2003) “Autoregressive Models for Capture-Recapture Data: A Bayesian Approach,” *Biometrics*, vol 59, 340-349.

Paper and software available at www.stat.colostate.edu/~jah

What factors affect animal survival?

- Historically, survival probabilities were modeled as as fixed constants.
- More recently Burnham and others have begun to consider survival probabilities as realizations of a random process.
- We model survival as a function of environmental factors and allow for a time series correlation structure between survival probabilities.
- We develop methodology for
 1. Open population mark-recapture models (Cormack-Jolly-Seber Model)
 2. Band recovery models

Band Recovery Model

- Animals are captured, banded and released.
- Hunters report when bands are recovered.
- Some notation:

$R_i =$ # banded animals released at time t_i

$m_{ij} =$ # animals recovered at time t_j
out of the R_i animals released at time t_i

$I =$ # capture occasions when banding
is performed

$J =$ # occasions when bands are recovered

$(m_{i1}, \dots, m_{iJ}) \sim \text{Multinomial}(R_i, p = f(\text{survival, recovery}))$

Likelihood for Band Recovery Model

$$\mathcal{L}(\phi, \lambda; \mathbf{R}, \mathbf{m}) = \prod_{i=1}^I \binom{R_i}{m_{ii}, \dots, m_{iJ}} \xi_i^{v_i} \prod_{j=i}^J \left\{ \lambda_j \prod_{k=i}^{j-1} \phi_k \right\}^{m_{ij}}$$

where

- ξ_i is the probability that an animal is never recovered after release at t_i
- v_i is the number of animals captured at t_i and never subsequently recovered
- λ_j is the probability that a marked animal, alive at t_j , is harvested between time t_j and t_{j+1} and reported to the banding agency.
- ϕ_k is the probability that an animal survives from time t_k to t_{k+1} given that it is alive at time t_k

A Random Effects Model for Survival Probabilities

We consider a generalized linear model for the probability that an animal survives from time t_j to time t_{j+1} of the form

$$g(\phi_j) = \mathbf{X}'_j \boldsymbol{\beta} + \epsilon_j, \quad j = 1, \dots, J,$$

where

- g is an appropriate link function to constrain survival between 0 and 1
- \mathbf{X}_j is a $P \times 1$ matrix of environmental covariates for capture occasion j
- $\boldsymbol{\beta}$ is a $P \times 1$ vector of regression coefficients
- $(\epsilon_1, \dots, \epsilon_J)' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$

Covariance Matrix for the Survival Model

The covariance matrix, Σ , can be any general form.

Here we consider an AR(m) model which implies that the ϵ_j error terms are realizations from the stochastic process

$$\epsilon_j = \sum_{k=1}^m \rho_k \epsilon_{j-k} + z_j, \quad j = 1, \dots, J,$$

where $z_j \sim \text{i.i.d. } N(0, \sigma^2)$ and $\boldsymbol{\rho} = (\rho_1, \dots, \rho_m)$ is a set of parameters.

The stationary AR(m) model

- allows for positive or negative correlation between survival probabilities that decreases with an increasing separation in time
- imposes constraints on $\boldsymbol{\rho}$ (see paper for details)

Parameter Estimation

- Parameter estimation via maximum likelihood or quasi-likelihood is challenging in this context. We adopt a Bayesian approach.
- For the Bayesian approach, we assume that the parameters β , σ^2 , ρ , and λ are independent *a priori*.
- The posterior distribution of the parameters and random effects is then given by

$$\begin{aligned}\pi(\beta, \sigma^2, \rho, \epsilon, \lambda | \mathbf{R}, \mathbf{X}) &\propto \mathcal{L}(\beta, \epsilon, \lambda; \mathbf{R}, \mathbf{X}) \\ &\times \pi(\beta)\pi(\sigma^2)\pi(\rho)\pi(\lambda) \\ &\times |\Sigma|^{-1/2} \exp \left\{ -\frac{\epsilon' \Sigma^{-1} \epsilon}{2} \right\}\end{aligned}$$

A modified Gibbs Sampler

We show that conditional distributions of the parameters are given by

$$f(\beta_l | \boldsymbol{\beta}_{-l}, \sigma^2, \boldsymbol{\rho}, \boldsymbol{\epsilon}, \boldsymbol{\lambda}, D) \propto \prod_{i=1}^I \xi_i^{v_i} \prod_{j=i}^J \left\{ \prod_{k=i+1}^{j-1} \phi_k \right\}^{m_{ij}} \pi(\beta_l)$$

$$f(\lambda_l | \boldsymbol{\lambda}_{-l}, \boldsymbol{\beta}, \boldsymbol{\epsilon}, \sigma^2, \boldsymbol{\rho}, D) \propto \prod_{i=1}^I \xi_i^{v_i} \prod_{j=i+1}^J \lambda_j^{m_{ij}} \pi(\lambda_l)$$

$$f(\epsilon_l | \boldsymbol{\epsilon}_{-l}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}, \boldsymbol{\lambda}, D) \propto \prod_{i=1}^I \xi_i^{v_i} \prod_{j=i}^J \left\{ \prod_{k=i+1}^{j-1} \phi_k \right\}^{m_{ij}} N(\mu_l / \eta_l, \sigma^2 / \eta_l),$$

A modified Gibbs Sampler, cont.

Where for an AR(2) error process, for example,

$$\mu_l = \begin{cases} \rho_1 \epsilon_2 + \rho_2 \epsilon_3 & l = 1 \\ \rho_1(\epsilon_1 + \epsilon_3) + \rho_2(\epsilon_4 - \rho_1 \epsilon_3) & l = 2 \\ \rho_1(1 - \rho_2)(\epsilon_{l-1} + \epsilon_{l+1}) + \rho_2(\epsilon_{l-2} + \epsilon_{l+2}) & l = 3, \dots, J - 2 \\ \rho_1(\epsilon_J + \epsilon_{J-2}) + \rho_2(\epsilon_{J-3} - \rho_1 \epsilon_J) & l = J - 1 \\ \rho_1 \epsilon_{J-1} + \rho_2 \epsilon_{J-2} & l = J \end{cases}$$

$$\eta_l = \begin{cases} 1 & l = 1 \text{ and } J \\ 1 + \rho_1^2 & l = 2 \text{ and } J - 1 \\ 1 + \rho_1^2 + \rho_2^2 & l = 3, \dots, J - 2 \end{cases}$$

$$f(\boldsymbol{\rho} | \boldsymbol{\beta}, \boldsymbol{\epsilon}, \sigma^2, \boldsymbol{\lambda}, D) \propto \left(\prod_{j=1}^m K_j \right)^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^J (\epsilon_j - \nu_j)^2 / K_j \right\} \pi(\boldsymbol{\rho})$$

$$f(\sigma^2 | \boldsymbol{\beta}, \boldsymbol{\epsilon}, \boldsymbol{\rho}, \boldsymbol{\lambda}, D) \propto \sigma^{-J} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^J (\epsilon_j - \nu_j)^2 / K_j \right\} \pi(\sigma^2)$$

where K_j is a function of $\boldsymbol{\rho}$.

Northern Pintails

- Study of female Northern Pintail ducks in California for banding years 1955 – 1983.
- Birds were banded in January of each banding year.
- No yearly environmental covariates were available, but we considered the possibility of a trend in survival probabilities over time via the model

$$\text{logit } \phi_j = \beta_0 + \beta_1(j - 14) + \epsilon_j \quad j = 1, \dots, 27$$

So the covariate vector $\mathbf{X}_j = (1, j - 14)$ for $j = 1, \dots, 27$.

- We also consider the logit model without a slope parameter.

Northern Pintail recovery data for banding years 1955 - 1983. The R_i represent the number of banded ducks released each year. Birds were banded in January of each Banding Year

Banding		Year of Recovery																											
Year	R_i	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82
55	270	7	6	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
56	693	21	10	4	2	3	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
57	1612	32	20	8	5	1	2	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
58	858	26	12	5	6	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
59	1471	21	18	6	5	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60	1051	18	4	6	4	1	2	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
61	796	24	6	4	0	3	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
62	277	10	9	6	6	4	1	2	4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
63	903	15	8	1	8	4	0	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
64	621	6	4	1	6	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
65	584	10	4	3	7	3	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
66	822	25	6	10	4	4	2	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
67	1344	28	27	8	11	3	1	4	1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
68	566	10	13	6	2	2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	481	9	7	3	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	695	11	11	5	2	2	1	1	0	1	1	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
71	632	22	10	2	4	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
72	1114	21	11	8	3	5	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
73	639	9	10	10	2	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
74	926	16	9	9	2	5	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
75	858	14	12	3	5	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
76	369	13	2	4	4	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
77	450	8	3	4	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78	212	6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
79	1680	18	28	8	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
80	421	14	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
81	118	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
82	60	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Choosing the order for the AR process

We chose an AR(2) model for these data, so the error terms follow the stochastic process

$$\epsilon_j = \rho_1 \epsilon_{j-1} + \rho_2 \epsilon_{j-2} + z_j, \quad j = 1, \dots, 27$$

The AR order was chosen based on a correlogram of the maximum likelihood estimates of yearly survival probabilities from the program MARK.

Prior distributions for the parameters

$$\begin{aligned}(\beta_0, \beta_1)^T &\sim N\left(\mathbf{0}, \frac{1}{0.01} \mathbf{I}\right) \\ \sigma^{-2} &\sim \Gamma(0.001, 0.001) \\ \rho_2 &\sim U(-1, 1) \\ \rho_1 | \rho_2 &\sim U(-(1 - \rho_2), 1 - \rho_2) \\ \lambda_j &\sim \text{i.i.d. } U(0, 1) \quad j = 1, \dots, 28\end{aligned}$$

MCMC sampling was carried out using winBUGS.

Northern Pintails: posterior estimates

Posterior means, standard deviations, and 90% highest probability density (HPD) intervals for the AR(2) model parameters.

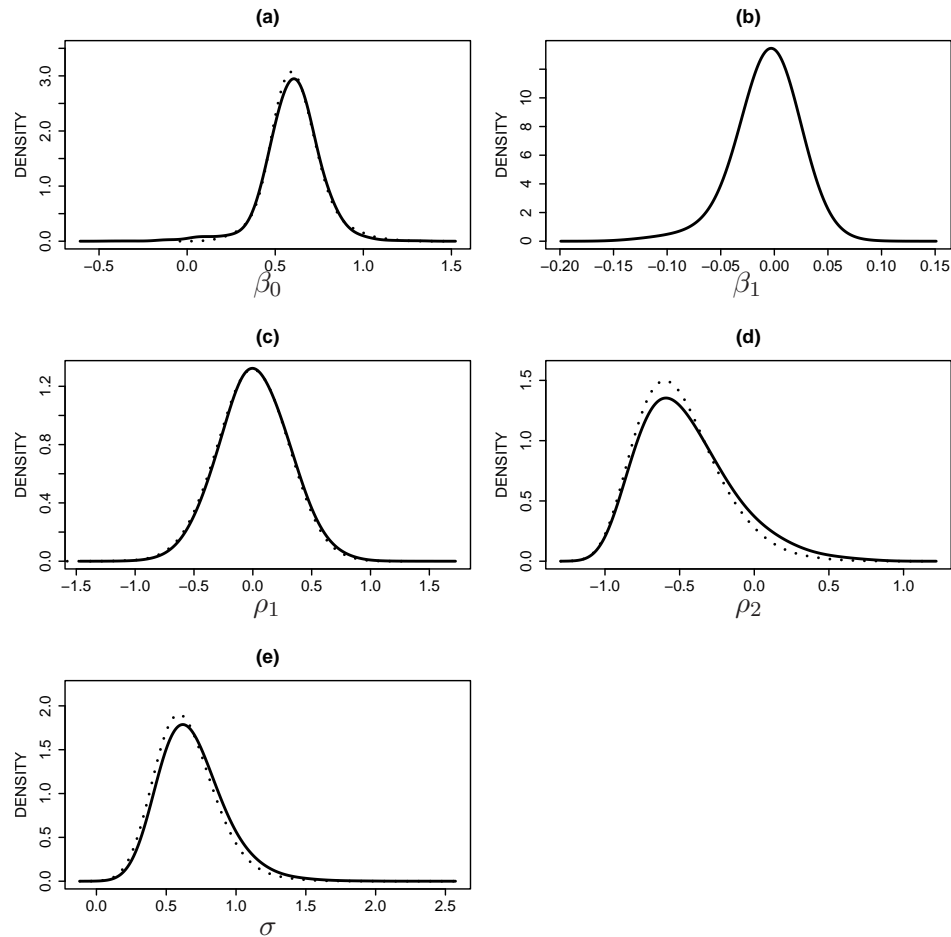
Model	Parameter	Mean	St. Dev.	90% HPD* Interval
Intercept and slope	β_0	0.600	0.159	(0.390, 0.850)
	β_1	-0.007	0.026	(-0.046, 0.033)
	ρ_1	0.014	0.288	(-0.458, 0.485)
	ρ_2	-0.452	0.307	(-0.928, 0.004)
	σ	0.688	0.222	(0.336, 1.015)
Intercept only	β_0	0.612	0.140	(0.409, 0.857)
	ρ_1	0.014	0.288	(-0.483, 0.456)
	ρ_2	-0.452	0.307	(-0.918, -0.109)
	σ	0.644	0.201	(0.330, 0.950)

* Estimated according to the algorithm presented by Chen et al. 2000.

Northern Pintails: Marginal Posterior Densities

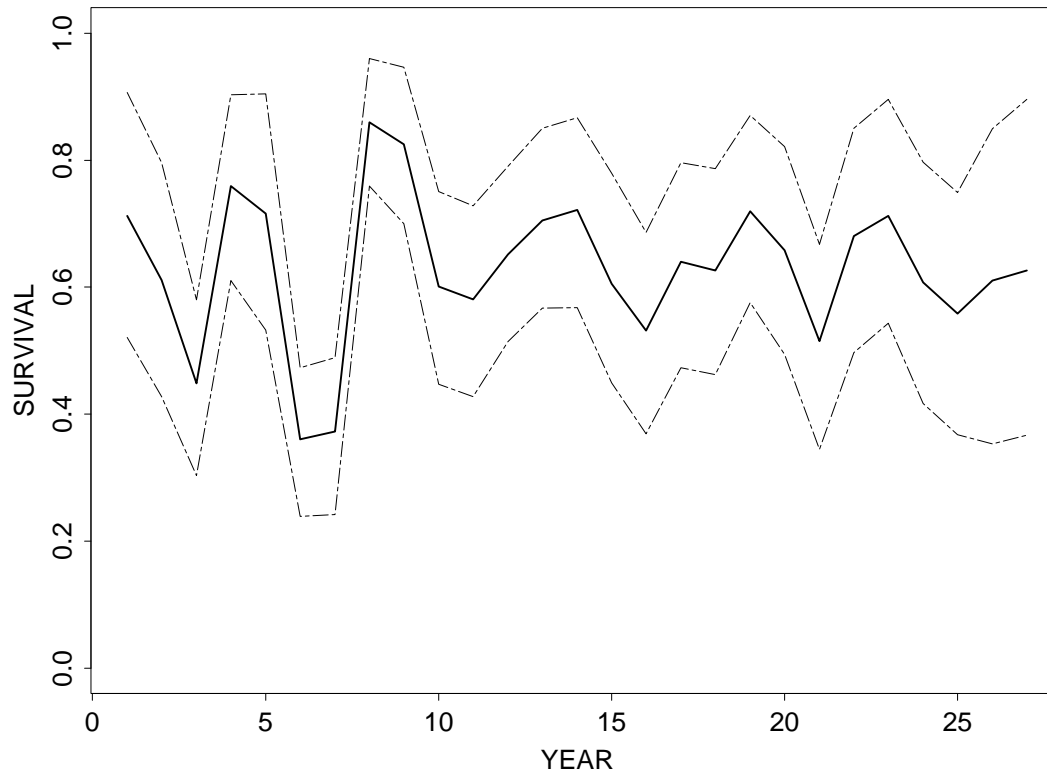
Solid line = linear time trend model

Dotted line = model with no linear time trend



Northern Pintails: Survival

Yearly survival estimates for Northern Pintails (model with no linear time trend). The solid line is the estimated posterior mean survival and the dashed lines represent a 90% HPD interval.



Northern Pintails: Conclusions

The time series modeling of capture-recapture data can provide additional insights to the survival process.

Extensions and related models include

- Random effects models for recovery parameters
- Multivariate AR process

Other issues include model selection and model uncertainty:

- Which covariates?
- Which link function?
- Which order for the AR process?

**GOOD LUCK WITH YOUR
BAYESIAN ANALYSES!**

Some Useful References on Bayesian Statistics

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Rubin (2004) *Bayesian Data Analysis*, Chapman & Hall, London.

Givens, G.H. and J.A. Hoeting (2005) *Computational Statistics*, John Wiley & Sons, New York.

Some Useful References on Spatial Statistics

Banerjee, S., B. Carlin and A. Gelfand (2004) *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall\CRC, New York.

Schabenberger, O. and C. A. Gotway, (2006) *Statistical Methods for Spatial Data Analysis*, Chapman & Hall\CRC, New York.

Other References Used in this Presentation

Berliner, L. M. (1996). "Hierarchical Bayesian time series models" in *Maximum Entropy and Bayesian Methods*, K. M. Hanson and R. N. Silver (eds.). Kluwer Academic Publishers, 15-22.

Caswell, H. (2001). *Matrix Population Models: Construction, Analysis, and Interpretation*. Sinauer, Sunderland, Massachusetts.

Johnson, D. S. and J. A. Hoeting (2003) "Autoregressive Models for Capture-Recapture Data: A Bayesian Approach," *Biometrics*, vol 59, 340-349.

Oli, M. K., M. V., P. A. Kleinb, L. D. Wendlandc, and M. B. Brown. (2006). Population dynamics of infectious diseases: A discrete time model. *Ecological Modelling* 198:183-194.