Bayesian estimation of bivariate regression models via MCMC

Applied Bayesian Statistics
Winter Term 2018

Susumu Shikano
GSDS

Two approaches to obtain posterior

- Conjugacy analysis
  - Conjugacy: The property that the prior and posterior have the same probability form depending on the form of the distribution used to calculate the likelihood.
  - Posterior can be obtained analytically.

Linear Regression Model

Two approaches to obtain posterior

- Conjugacy analysis
  - Conjugacy: The property that the prior and posterior have the same probability form depending on the form of the distribution used to calculate the likelihood.
  - Posterior can be obtained analytically.
- Deriving posterior per Gibbs Sampling
  - Conjugacy is not must.
  - Use of MCMC.

Today's session

- An alternative approach to obtain posterior
  - Derive conditional posterior
  - Run Gibb sampling
- Gibbs sampling and its alternatives
  - Gibbs sampling
  - Metropolis-Hasting algorithm
  - Slice Sampling

An alternative approach

Conditional posterior for the slope parameter β_1

- Prior:

$$f(\beta_1) = f_N(\mu_1^0, V_1^0) = \frac{1}{\sqrt{2\pi V_1^0}} \exp\left\{-\frac{(\beta_1 - \mu_1^0)'(\beta_1 - \mu_1^0)}{2V_1^0}\right\}$$
$$= \frac{1}{\sqrt{2\pi V_1^0}} \exp\left\{-\frac{(\beta_1 - \mu_1^0)^2}{2V_1^0}\right\}$$

- Likelihood:

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \boldsymbol{X}\boldsymbol{\beta})'(y_i - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right\}$$
$$= \prod_i^n \frac{1}{\sqrt{\pi\sigma^2}} \exp\left\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\}$$

An alternative approach

Conditional posterior for the slope parameter β_1

$$f(\beta_1|y, \beta_0, \sigma^2) \quad \propto \quad f(\beta_1)f(y|\beta_0, \beta_1, \sigma^2)$$

An alternative approach

Conditional posterior for the slope parameter β_1

$$
\begin{aligned}
f(\beta_1|y,\beta_0,\sigma^2) &\propto f(\beta_1)f(y|\beta_0,\beta_1,\sigma^2) \\
&\propto \frac{1}{\sqrt{2\pi V_1^0}}\exp\left\{-\frac{(\beta_1-\mu_1^0)^2}{2V_1^0}\right\}\times \\
&\quad \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right\}
\end{aligned}
$$

An alternative approach

Conditional posterior for the slope parameter β_1

$$\begin{aligned}
f(\beta_1|y,\beta_0,\sigma^2) &\propto f(\beta_1)f(y|\beta_0,\beta_1,\sigma^2)\\
&\propto \frac{1}{\sqrt{2\pi V_1^0}}\exp\left\{-\frac{(\beta_1-\mu_1^0)^2}{2V_1^0}\right\}\times\\
&\quad \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right\}\\
&\propto \exp\left\{-\left(\frac{1}{2V_1^0}+\frac{\sum_i^n x_i^2}{2\sigma^2}\right)\beta_1^2+\right.\\
&\qquad \left.\left(\frac{\mu_1^0}{V_1^0}+\frac{\sum_i^n(y_i-\beta_0)x_i}{\sigma^2}\right)\beta_1+const\right\}
\end{aligned}$$

An alternative approach

If a pdf can be expressed as…

$$f(\theta) \propto \exp(a\theta^2 + b\theta + const)$$

The corresponding distribution is a normal distribution: $N\left(-\frac{b}{2a}, -\frac{1}{2a}\right)$

An alternative approach

If a pdf can be expressed as…

$$f(\theta) \propto \exp(a\theta^2 + b\theta + const)$$

The corresponding distribution is a normal distribution: $N\left(-\frac{b}{2a}, -\frac{1}{2a}\right)$

Conditional posterior for β_1: $N(\mu_1^*, V_1^*)$

$$\begin{aligned}
\mu_1^* &= V_1^* \left(\frac{\mu_1^0}{V_1^0} + \frac{\sum_i^n (y_i - \beta_0) x_i}{\sigma^2} \right) \\
V_1^* &= \left(\frac{1}{V_1^0} + \frac{\sum_i^n x_i^2}{\sigma^2} \right)^{-1}
\end{aligned}$$

An alternative approach

Conditional posterior for the intercept β_0: $N(\mu_0^*, V_0^*)$

Analogously to the conditional posterior for β_1:

$$
\begin{aligned}
\mu_0^* &= V_0^* \left(\frac{\mu_0^0}{V_0^0} + \frac{\sum_i^n (y_i - \beta_1 x_i)}{\sigma^2} \right) \\
V_0^* &= \left(\frac{1}{V_0^0} + \frac{n}{\sigma^2} \right)^{-1}
\end{aligned}
$$

An alternative approach

Conditional posterior for β_0 and β_1

An alternative approach

Conditional posterior for β_0 and β_1

- It is also possible to derive conditional posterior for β_0 and β_1.

An alternative approach

Conditional posterior for β_0 and β_1

- It is also possible to derive conditional posterior for β_0 and β_1.
- Prior: multivariate normal.

An alternative approach

Conditional posterior for β_0 and β_1

- It is also possible to derive conditional posterior for β_0 and β_1.
- Prior: multivariate normal.
- For more details see Shikano (2014) Bayesian estimation of regression models,...

An alternative approach

Conditional posterior for σ^2

$$
\begin{aligned}
f(\sigma^2|y,\beta_0,\beta_1) &\propto f_{\Gamma^{-1}}(\sigma^2) f(y|\beta_0,\beta_1,\sigma^2) \\
&\propto \frac{a0^{d0}}{\Gamma(d0)} \left(\sigma^2\right)^{-d0-1} \exp\left(-\frac{a0}{\sigma^2}\right) \\
&\quad \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right\} \\
&\propto \left(\sigma^2\right)^{-d0-1-\frac{n}{2}} \\
&\quad \exp\left\{-\frac{1}{\sigma^2}\left(a0+\frac{1}{2}\sum_i^n (y_i-\beta_0-\beta_1 x_i)^2\right)\right\}
\end{aligned}
$$

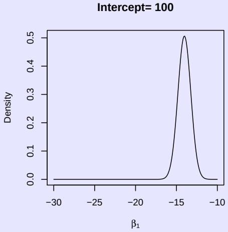
An alternative approach

Conditional posterior for σ^2

$$\sigma^2 | y, \beta_0, \beta_1 \sim IG\left(a0 + \frac{1}{2}\sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2, d0 + \frac{n}{2} \right)$$

Conditional posterior

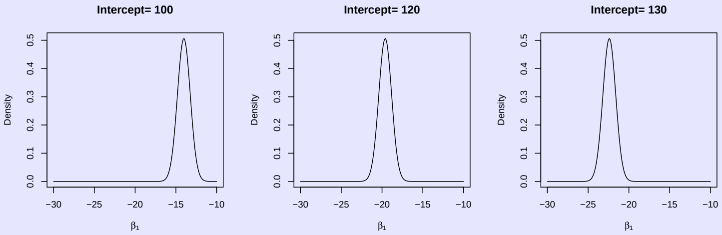
$\beta_1 | \beta_0, \sigma^2$



σ^2 is set to 253.11. Prior for β_1 is $N(0, 10000)$.

Conditional posterior

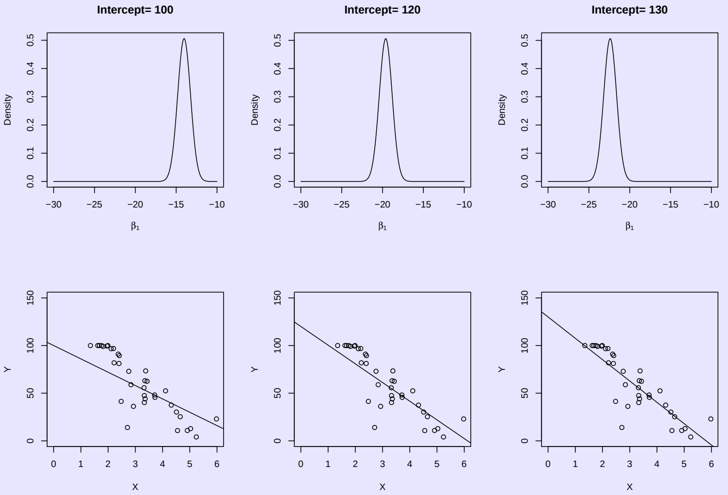
$\beta_1 | \beta_0, \sigma^2$



σ^2 is set to 253.11. Prior for β_1 is $N(0, 10000)$.

Conditional posterior

$\beta_1 | \beta_0, \sigma^2$



σ^2 is set to 253.11. Prior for β_1 is $N(0, 10000)$.

Gibbs sampling

Simulation-based MCMC algorithm

- Obtaining the joint distribution from individual conditional distributions.
  - $f(\beta_0, \beta_1, \sigma^2)$ from...

Gibbs sampling

Simulation-based MCMC algorithm

- Obtaining the joint distribution from individual conditional distributions.
  - $f(\beta_0, \beta_1, \sigma^2)$ from...
  - $f(\beta_0|\beta_1, \sigma^2)$, $f(\beta_1|\beta_0, \sigma^2)$ and $f(\sigma^2|\beta_0, \beta_1)$
    1. Choose a starting value $\beta_1^{(0)}$ and $\sigma^{2(0)}$
    2. Draw $\beta_0^{(1)}$ from $f(\beta_0|\beta_1^{(0)}, \sigma^{2(0)})$.
    3. Draw $\beta_1^{(1)}$ from $f(\beta_1|\beta_0^{(1)}, \sigma^{2(0)})$.
    4. Draw $\sigma^{2(1)}$ from $f(\sigma^2|\beta_0^{(1)}, \beta_1^{(1)})$.
    5. $\vdots$
    6. Draw $\beta_0^{(g)}$ from $f(\beta_0|\beta_1^{(g-1)}, \sigma^{2(g-1)})$.
    7. Draw $\beta_1^{(g)}$ from $f(\beta_1|\beta_0^{(g)}, \sigma^{2(g-1)})$.
    8. Draw $\sigma^{2(g)}$ from $f(\sigma^2|\beta_0^{(g)}, \beta_1^{(g)})$.

Gibbs sampling

Bivariate regression model

```
> chain
                    beta0      beta1 sigma.sqr
starting.value         NA   0.000000    1.0000
Iteration 1      59.52691  -2.654881  847.7251
Iteration 2      62.76098  -4.367976  966.1888
Iteration 3      69.56556  -4.469316  877.0900
Iteration 4      63.97678  -4.759496  704.2265
Iteration 5      71.45450  -5.533939  576.5033
Iteration 6      75.19838  -7.264575  511.4502
Iteration 7      81.29282  -9.173630  772.7803
Iteration 8      92.61787 -14.100181  592.2868
Iteration 9     104.85203 -12.859007  395.7271
Iteration 10     99.47174 -13.724145  275.8063
Iteration 11    100.49510 -14.362180  552.9542
Iteration 12    108.28597 -17.344073  311.3477
Iteration 13    111.61257 -17.510766  281.4529
Iteration 14    111.68330 -15.617246  315.2950
Iteration 15    112.97822 -18.594780  358.8411
Iteration 16    115.27897 -19.137953  286.8063
Iteration 17    121.49736 -19.647173  205.9642
```

An example Markov Chain

Gibbs sampling

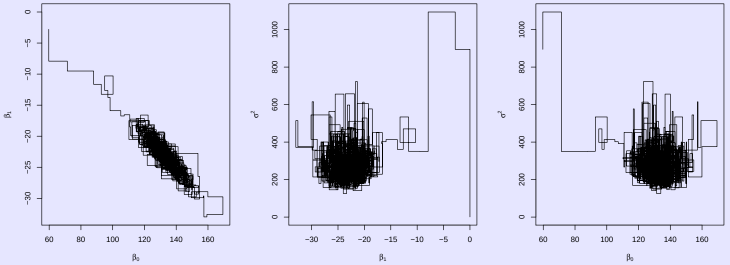
Bivariate regression model

```
> chain
                    beta0      beta1  sigma.sqr
starting.value         NA   0.000000     1.0000
Iteration 1      59.52691  -2.654881   847.7251
Iteration 2      62.76098  -4.367976   966.1888
Iteration 3      69.56556  -4.469316   877.0900
Iteration 4      63.97678  -4.759496   704.2265
Iteration 5      71.45450  -5.533939   576.5033
Iteration 6      75.19838  -7.264575   511.4502
Iteration 7      81.29282  -9.173630   772.7803
Iteration 8      92.61787 -14.100181   592.2868
Iteration 9     104.85203 -12.859007   395.7271
Iteration 10     99.47174 -13.724145   275.8063
Iteration 11    100.49510 -14.362180   552.9542
Iteration 12    108.28597 -17.344073   311.3477
Iteration 13    111.61257 -17.510766   281.4529
Iteration 14    111.68330 -15.617246   315.2950
Iteration 15    112.97822 -18.594780   358.8411
Iteration 16    115.27897 -19.137953   286.8063
Iteration 17    121.49736 -19.647173   205.9642
```

An example Markov Chain

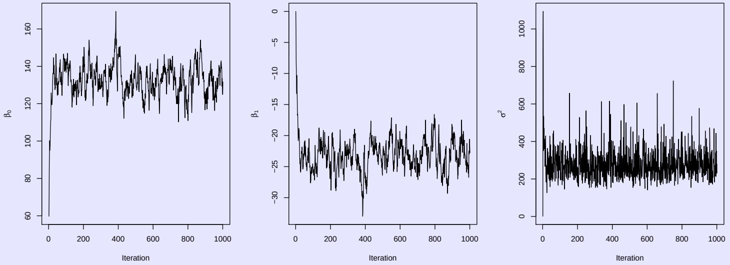
Gibbs sampling

Bivariate regression model



Gibbs sampling

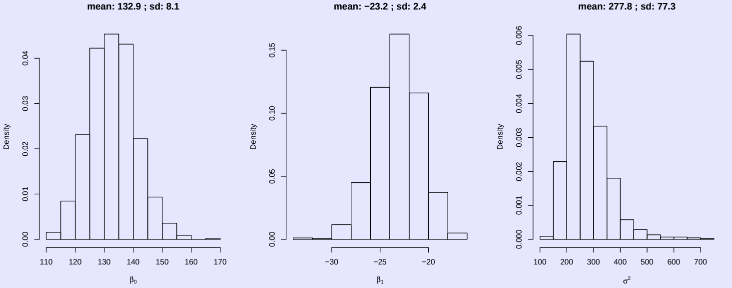
Bivariate regression model



After some iterations (burn-in), the chain converged to an invariant distribution.

Gibbs sampling

Bivariate regression model



After 100 iterations.

Metropolis-Hasting Algorithm

Metropolis-Hasting Algorithm?

- More general MCMC of the Gibbs sampling
- Differently from the Gibbs sampling, the full set of the conditional posterior for all parameters is not required.

Metropolis-Hasting Algorithm

Basic idea

Random draws from a posterior $f(\beta|y)$. The MH algorithm can be described by the following iterative steps for $t = 1, \ldots, T$:

1. Set $\beta = \beta^{(t-1)}$
2. Generate new candidate values β' from a proposal distribution $q(\beta'|\beta)$
3. Calculate $\alpha = \min\left(1, \frac{f(\beta'|y)q(\beta|\beta')}{f(\beta|y)q(\beta'|\beta)}\right) = \min\left(1, \frac{f(y|\beta')f(\beta')q(\beta|\beta')}{f(y|\beta)f(\beta)q(\beta'|\beta)}\right)$.
4. Update $\beta^{(t)} = \beta'$ with probability α (acceptance). Otherwise set $\beta^{(t)} = \beta$ (rejection of β').

Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

- You obtained as posterior (=product of prior and likelihood) the following density function:

$$f(\theta) = \frac{a^d}{\Gamma(d)}\theta^{-d-1}\exp\left(-\frac{a}{\theta}\right)$$

Application: Inverse Gamma Distribution

- You obtained as posterior (=product of prior and likelihood) the following density function:

$$f(\theta) = \frac{a^d}{\Gamma(d)}\theta^{-d-1}\exp\left(-\frac{a}{\theta}\right)$$

- Which form has the distribution?

Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

- You obtained as posterior (=product of prior and likelihood) the following density function:

$$f(\theta) \quad = \quad \frac{a^d}{\Gamma(d)} \theta^{-d-1} \exp\left(-\frac{a}{\theta}\right)$$

- Which form has the distribution?
- What is the expected value?

Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

- You obtained as posterior (=product of prior and likelihood) the following density function:

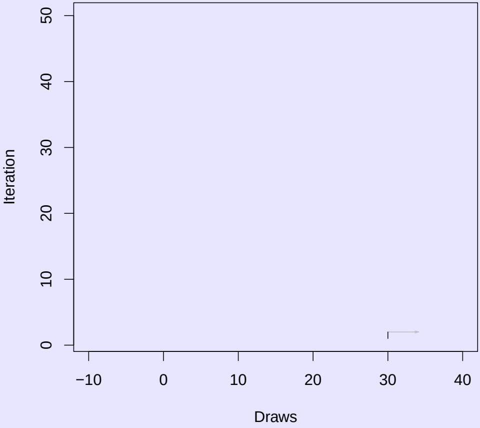
$$f(\theta) = \frac{a^d}{\Gamma(d)}\theta^{-d-1}\exp\left(-\frac{a}{\theta}\right)$$

- Which form has the distribution?
- What is the expected value?
- What is the variance?
- …

Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

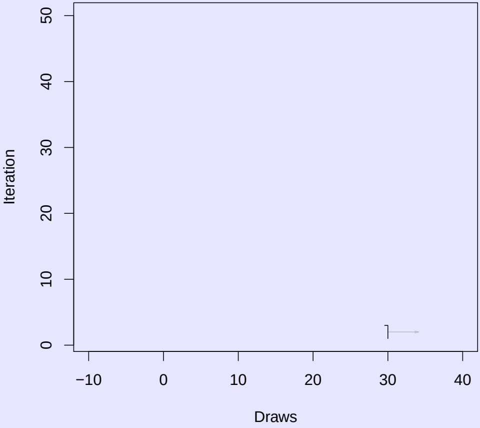
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

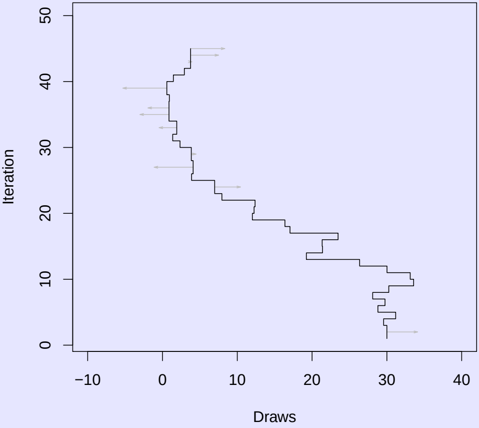
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

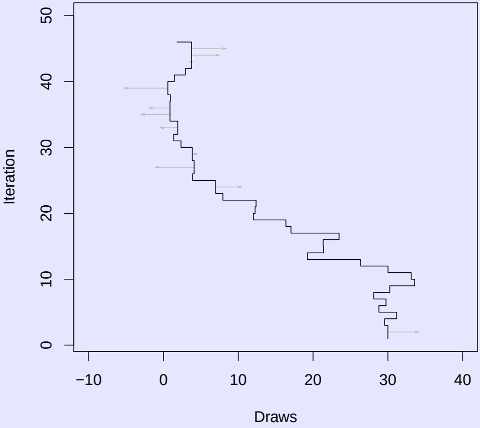
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

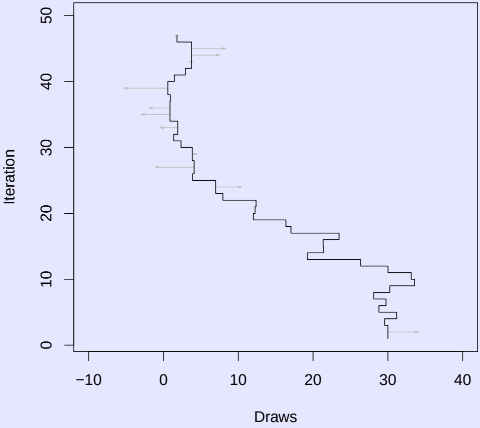
Application: Inverse Gamma Distribution

- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Application: Inverse Gamma Distribution

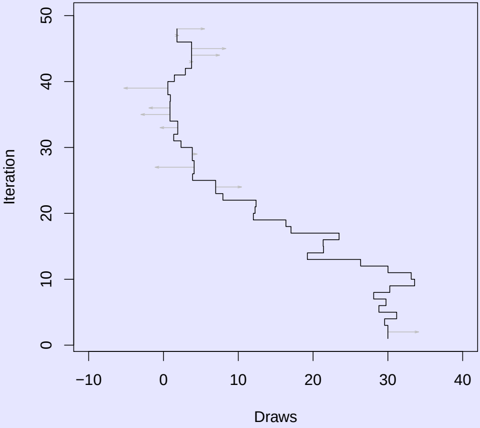
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

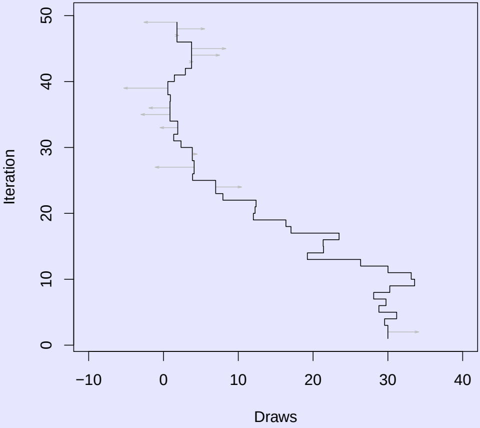
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

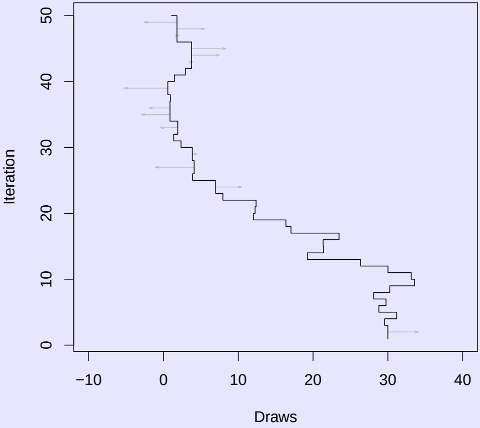
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

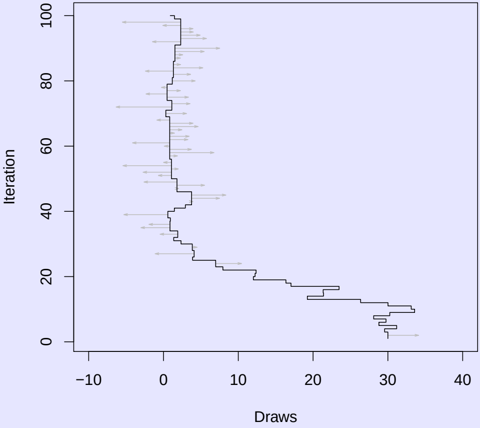
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

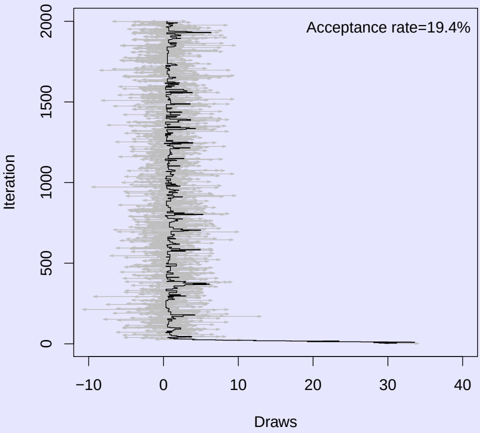
- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

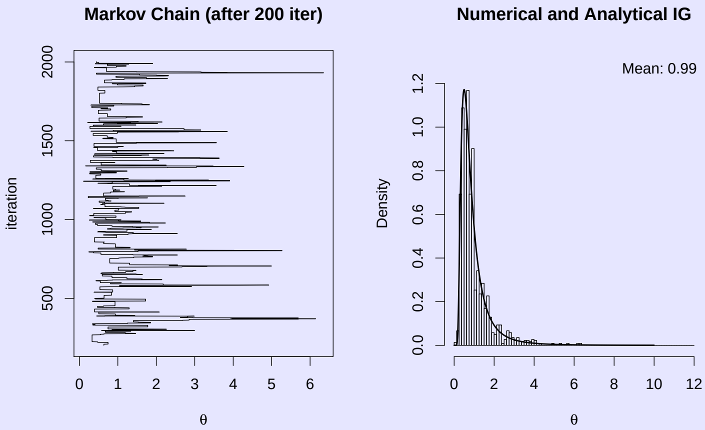
Application: Inverse Gamma Distribution

- Parameter values: a=2, d=3
- Initial value: 30
- Proposal dist.: normal distribution with sd=3



Metropolis-Hasting Algorithm

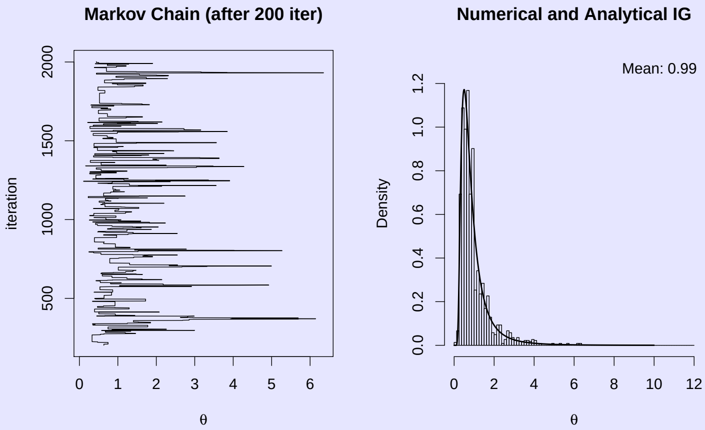
Application: Inverse Gamma Distribution



- Parameter values: a=2, d=3

Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution



- Parameter values: a=2, d=3
- Average of IG: $\frac{a}{d-1} = 1$

Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

- MH also works even if we know only the following posterior

$$
\begin{aligned}
f(\theta) &= \frac{a^d}{\Gamma(d)}\theta^{-d-1}\exp\left(-\frac{a}{\theta}\right) \\
&\propto \theta^{-d-1}\exp\left(-\frac{a}{\theta}\right)
\end{aligned}
$$

Metropolis-Hasting Algorithm

Application: Inverse Gamma Distribution

- MH also works even if we know only the following posterior

$$
\begin{aligned}
f(\theta) &= \frac{a^d}{\Gamma(d)}\theta^{-d-1}\exp\left(-\frac{a}{\theta}\right) \\
&\propto \theta^{-d-1}\exp\left(-\frac{a}{\theta}\right)
\end{aligned}
$$

- $\frac{f(y|\beta')f(\beta')q(\beta|\beta')}{f(y|\beta)f(\beta)q(\beta'|\beta)}$ can cancel out the constant term!

Slice Sampling

A drawback of MH Algorithm

If a proposal distribution's variance of proposal distribution is

- too small: Markov chains converge very slowly.
- too large: acceptance rate is too small (inefficient).

Any algorithm which requires no proposal distribution?

Slice Sampling

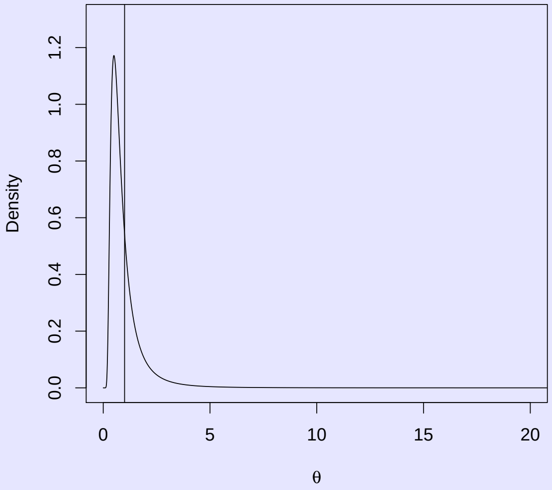
Procedure

- Goal: Generation of random numbers from a pdf $f(x)$
- You have no program for $f(x)$.

1. Choose a starting value x_0 for which $f(x_0) > 0$.
2. Sample a y value uniformly between 0 and $f(x_0)$.
3. Draw a horizontal line across the curve at this y position.
4. Sample a point (x, y) from the line segments within the curve.
5. Repeat from step 2 using the new x value.

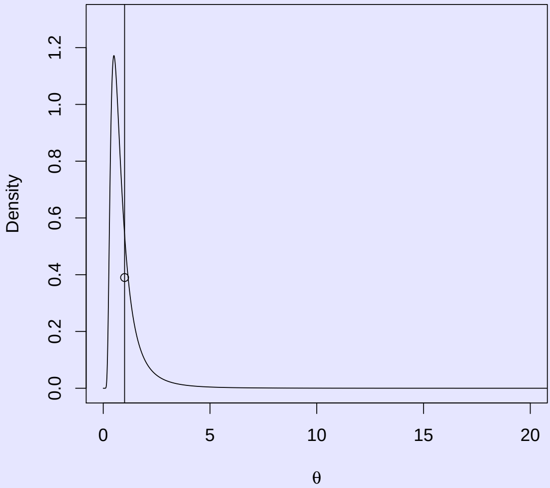
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



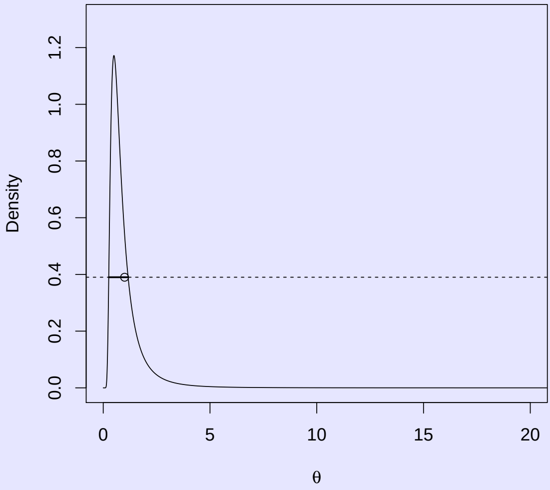
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



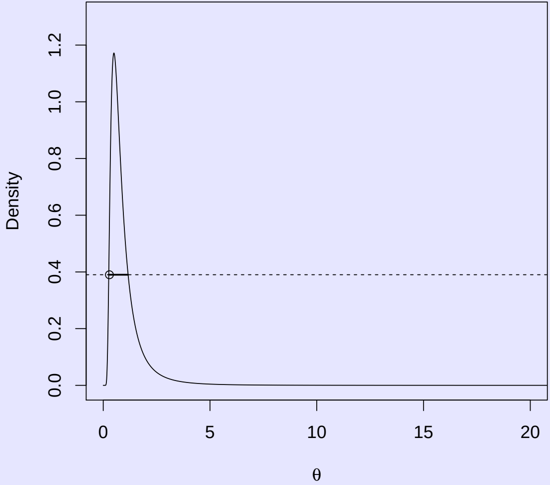
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



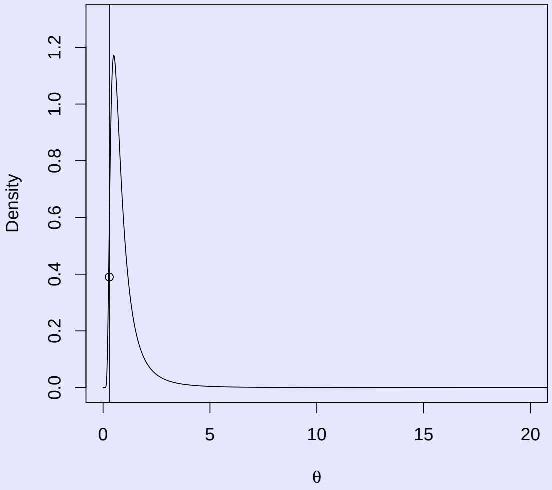
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



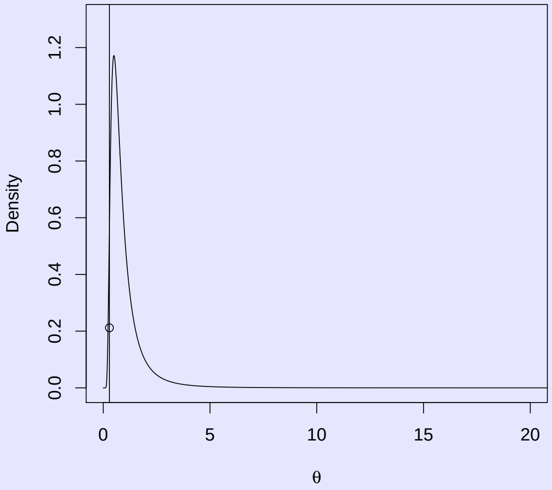
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



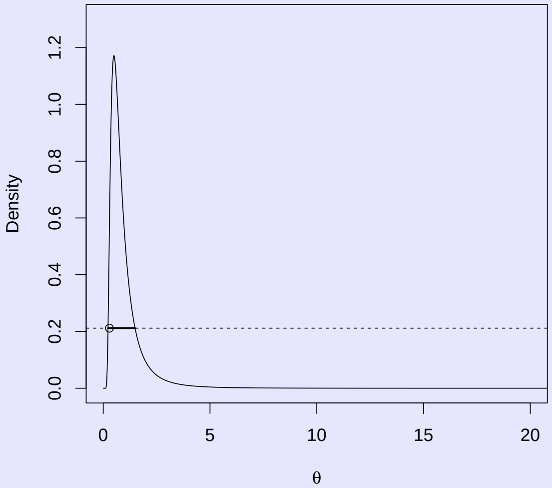
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



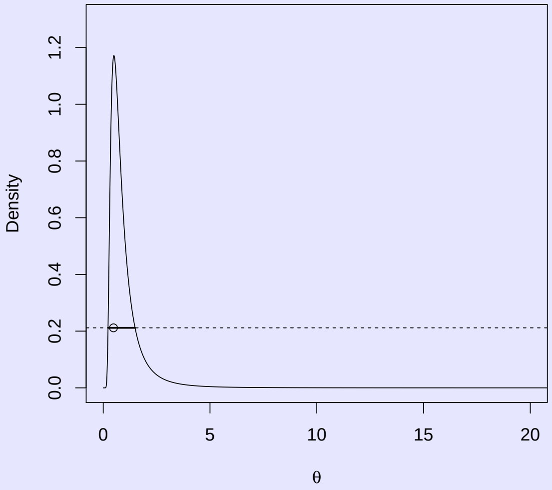
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



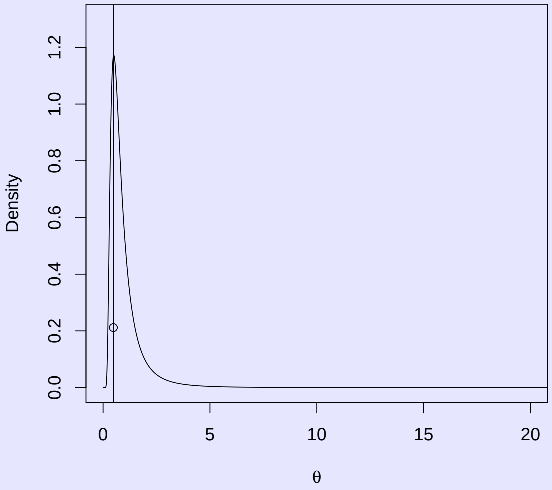
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



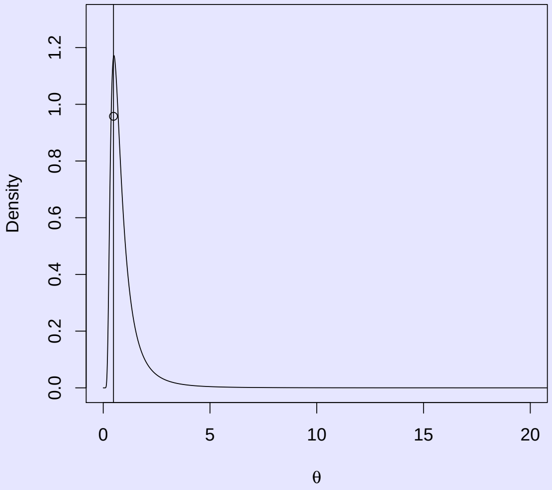
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



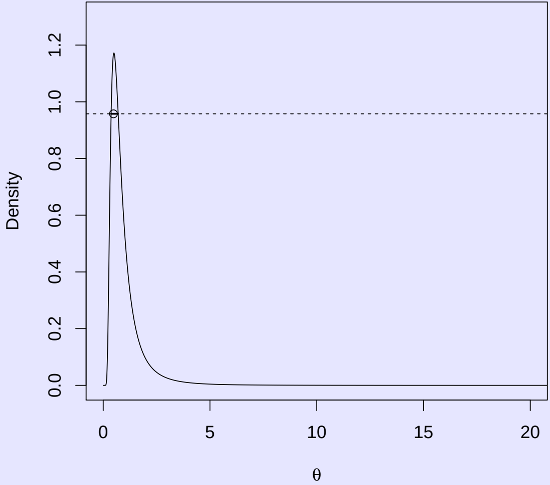
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



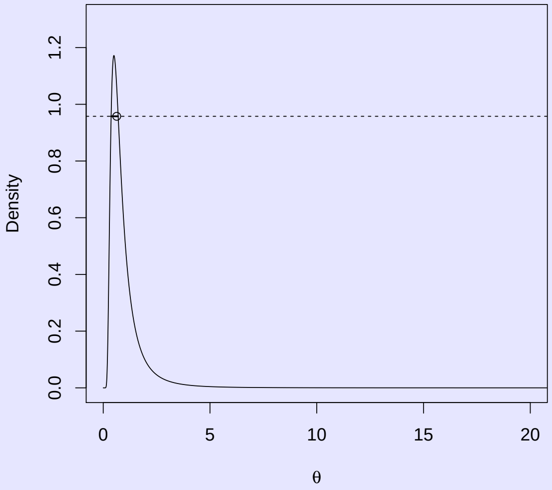
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



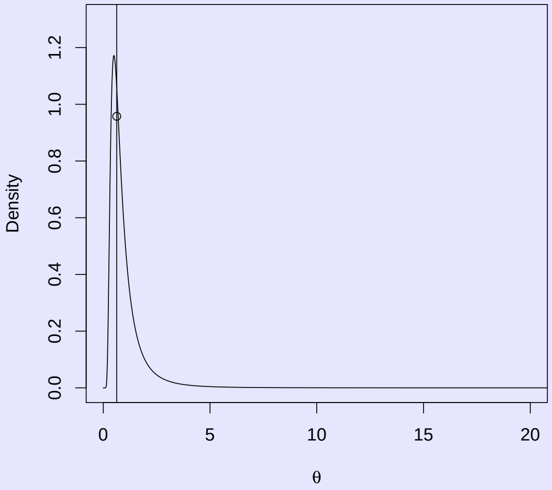
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



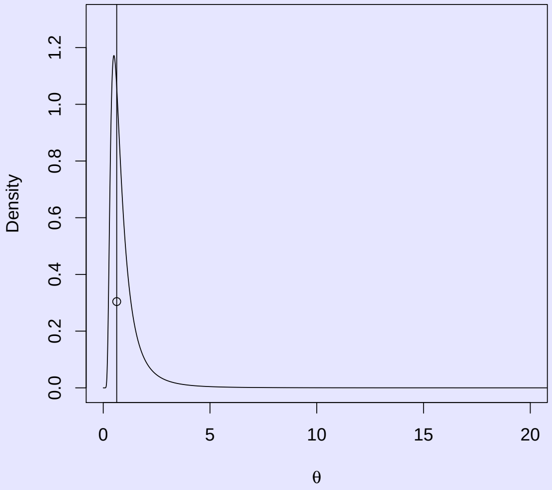
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



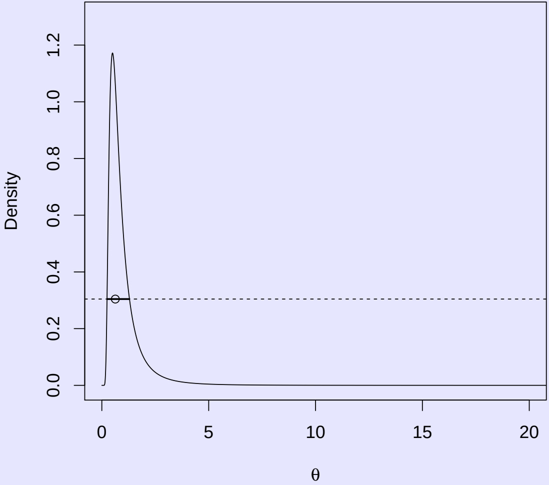
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



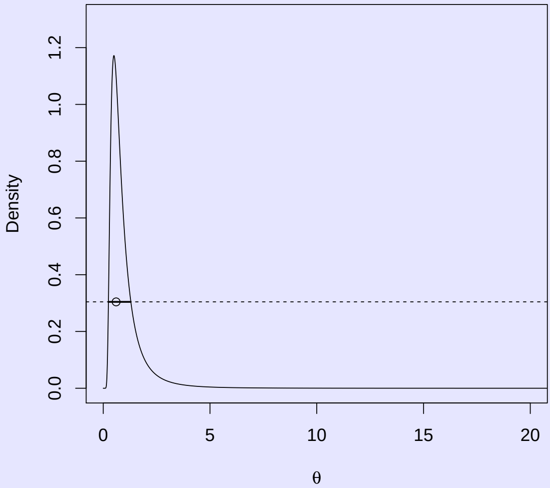
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



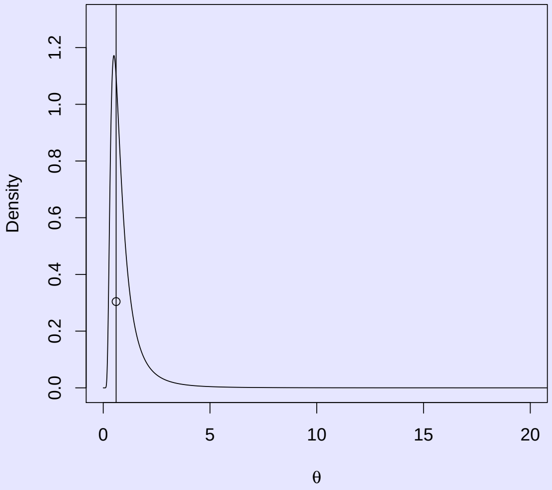
Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$

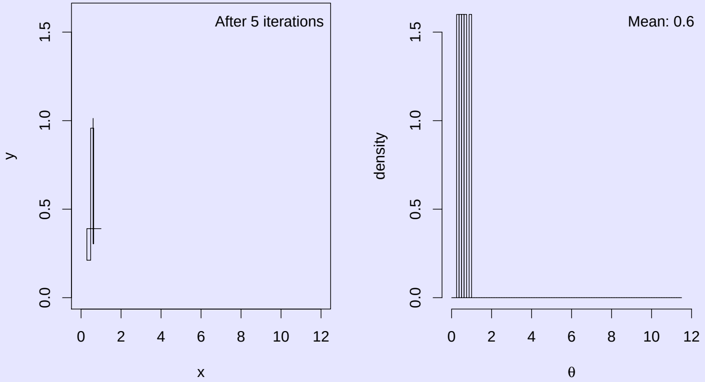


Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$

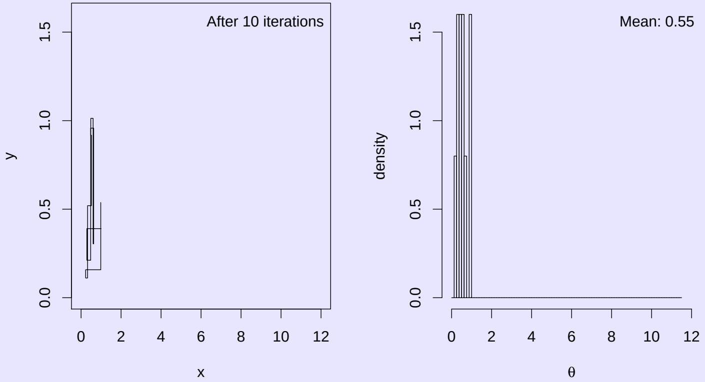


Example: IG(a=2,d=3) with $x_0 = 1$

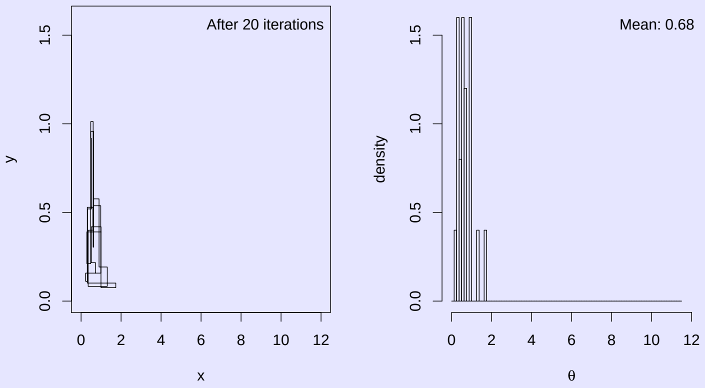


Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$

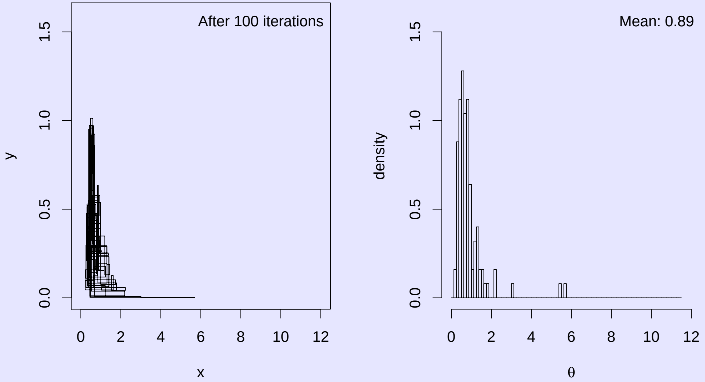


Example: IG(a=2,d=3) with $x_0 = 1$



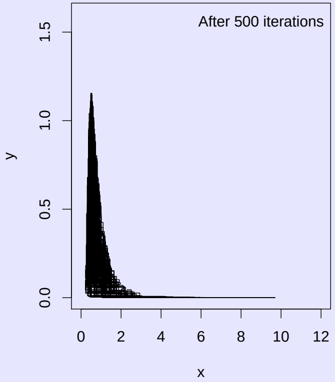
Slice Sampling

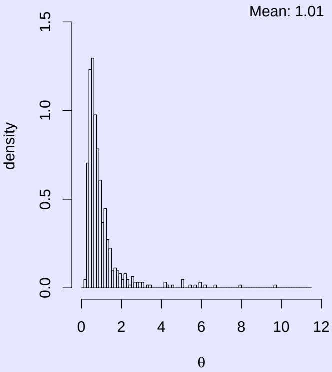
Example: IG(a=2,d=3) with $x_0 = 1$



Slice Sampling

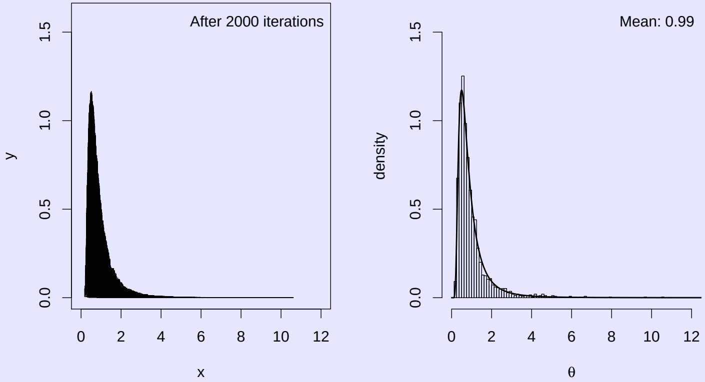
Example: IG(a=2,d=3) with $x_0 = 1$





Slice Sampling

Example: IG(a=2,d=3) with $x_0 = 1$



Markov-Chain-Monte-Carlo (MCMC)

Gibbs, MH and Slice sampling

Markov-Chain-Monte-Carlo (MCMC)

Gibbs, MH and Slice sampling

- Tools for random number generation based on a certain distribution. . .

Markov-Chain-Monte-Carlo (MCMC)

Gibbs, MH and Slice sampling

- Tools for random number generation based on a certain distribution. . .
- with many iterations.

Markov-Chain-Monte-Carlo (MCMC)

Gibbs, MH and Slice sampling

- Tools for random number generation based on a certain distribution. . .
- with many iterations.
- Random number generation in each iteration depends on the last iteration.

Markov-Chain-Monte-Carlo (MCMC)

Gibbs, MH and Slice sampling

- Tools for random number generation based on a certain distribution. . .
- with many iterations.
- Random number generation in each iteration depends on the last iteration.
- $\rightarrow$ Markov-Chain-Monte-Carlo

Markov-Chain-Monte-Carlo (MCMC)

Gibbs, MH and Slice sampling

- Tools for random number generation based on a certain distribution. . .
- with many iterations.
- Random number generation in each iteration depends on the last iteration.
- $\rightarrow$ Markov-Chain-Monte-Carlo

Markov-Chain-Monte-Carlo (MCMC)

Gibbs, MH and Slice sampling

- Tools for random number generation based on a certain distribution...
- with many iterations.
- Random number generation in each iteration depends on the last iteration.
- $\rightarrow$ Markov-Chain-Monte-Carlo

Markov-Chain-Monte-Carlo (MCMC)

Markov chain

- A stochastic process satisfying the Markov property.

Markov-Chain-Monte-Carlo (MCMC)

Markov chain

- A stochastic process satisfying the Markov property.
- Markov property: the probability distribution at $t+1$ depends only on the state of the system at t.

Markov-Chain-Monte-Carlo (MCMC)

Markov chain

- A stochastic process satisfying the Markov property.
- Markov property: the probability distribution at $t + 1$ depends only on the state of the system at t.
- In a finite state space with all probabilities positive, there is a unique invariant distribution.

Markov-Chain-Monte-Carlo (MCMC)

Markov chain

- A stochastic process satisfying the Markov property.
- Markov property: the probability distribution at $t + 1$ depends only on the state of the system at t.
- In a finite state space with all probabilities positive, there is a unique invariant distribution.
  - Condition: The chain has to be irreducible and aperiodic.
  - For large enough n, the initial state plays almost no role.

Markov-Chain-Monte-Carlo (MCMC)

Advantages: simple and straightforward interpretation!

- Quite easy since we generate a „dataset" of draws from MC.
- Simple calculation of any quantity of interest:
  - the expected value or the median of that posterior, just calculate those of the generated draws.
  - 95% credible intervals, just find the 2.5th and 97.5th percentiles of the draws.
- You need just descriptive statistics to describe MC.

Wrap up

Bayesian regression models

- In conjugacy analysis posterior can be derived analytically.
- The larger n/the smaller the dispersion of prior, the more similar results with the maximum likelihood.
- Alternative to conjugacy analysis:
  - Deriving conditional posterior
  - Run Gibbs sampling (one of MCMC)
  - Obtain joint and marginal posterior
- Gibbs sampling is one possible MCMC algorithm.

Rejection Sampling

Rejection Sampling or Acceptance Rejection Sampling

- Goal: Generation of random numbers from a pdf $f(x)$
- You have no program for $f(x)$, but for another distribution $g(x)$
  1. You find a constant c so that $f(x) \leq cg(x)$ for all x.
  2. You generate a random number x^* from $g(x)$.
  3. You generate a random number u from $U[0, 1]$.
  4. If $u \leq f(x^*)/cg(x^*)$ you accept x^* as random number from $f(x)$. Otherwise you reject it.
  5. Go to Step 2.

Markov Chains

Markov process

- A stochastic process X_t taking values...
  - in the finite set $S = \{1, 2, \ldots, s\}$
  - t: time or iteration

Markov Chains

Markov process

- A stochastic process X_t taking values...
  - in the finite set $S = \{1, 2, \ldots, s\}$
  - t: time or iteration
- p_{ij} is transition probabilities
  - $p_{ij} = \Pr(X_{t+1} = j | X_t = i)$, $i, j, \in S$.

Markov Chains

Markov process

- A stochastic process X_t taking values…
  - in the finite set $S = \{1, 2, \dots, s\}$
  - t: time or iteration
- p_{ij} is transition probabilities
  - $p_{ij} = \Pr(X_{t+1} = j | X_t = i)$, $i, j, \in S$.
- Markov property: the probability distribution at $t + 1$ depends only on the state of the system at t.

Markov Chains

Markov process

- A stochastic process X_t taking values...
  - in the finite set $S = \{1, 2, \dots, s\}$
  - t: time or iteration
- p_{ij} is transition probabilities
  - $p_{ij} = \Pr(X_{t+1} = j | X_t = i)$, $i, j, \in S$.
- Markov property: the probability distribution at $t + 1$ depends only on the state of the system at t.
- Note that p_{ij} are constant over t.

Markov Chains

Transition matrix

- Transition probabilities: p_{ij}
  - $p_{ij} \geq 0$
  - $\sum_{j=1}^{s} p_{ij} = 1$
- $s \times s$ transition matrix $P = \{p_{ij}\}$

Markov Chains

Transition matrix

- Transition probabilities: p_{ij}
  - $p_{ij} \geq 0$
  - $\sum_{j=1}^{s} p_{ij} = 1$
- $s \times s$ transition matrix $P = \{p_{ij}\}$
- Distribution at $t + 2$?

Markov Chains

Transition matrix

- Transition probabilities: p_{ij}
  - $p_{ij} \geq 0$
  - $\sum_{j=1}^{s} p_{ij} = 1$
- $s \times s$ transition matrix $P = \{p_{ij}\}$
- Distribution at $t + 2$?
  - $p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}$

Markov Chains

Transition matrix

- Transition probabilities: p_{ij}
  - $p_{ij} \geq 0$
  - $\sum_{j=1}^{s} p_{ij} = 1$
- $s \times s$ transition matrix $P = \{p_{ij}\}$
- Distribution at $t + 2$?
  - $p_{ij}^{(2)} = \sum_k p_{ik} p_{kj}$
  - This is given by $PP \equiv P^2$

Markov Chains

Invariant distribution

- The probability distribution $\pi = (\pi_1, \dots \pi_s)'$ is invariant for P
- if $\pi' = \pi' P$,
- or $\pi_j = \sum_i \pi_i p_{ij}$ for $j = 1, \dots, s$.

Invariant distribution

- The probability distribution $\pi = (\pi_1, \ldots \pi_s)'$ is invariant for P
- if $\pi' = \pi' P$,
- or $\pi_j = \sum_i \pi_i p_{ij}$ for $j = 1, \ldots, s$.
- Note that π' is an eigenvector of P.

Markov Chains

Theorem

- Suppose S is finite and $p_{ij} > 0 \; \forall i, j$.
- Then there exists a unique probability distribution π_j, $j \in S$
- such that $\sum_i \pi_i p_{ij} = \pi_j \forall j \in S$.
- Further, $|p_{ij}^{(n)} - \pi_j| \leq r^n$,
- where $0 < r < 1$, for all i, j and $n \geq 1$.

Markov Chains

Theorem

- Suppose S is finite and $p_{ij} > 0 \; \forall i, j$.
- Then there exists a unique probability distribution π_j, $j \in S$
- such that $\sum_i \pi_i p_{ij} = \pi_j \forall j \in S$.
- Further, $|p_{ij}^{(n)} - \pi_j| \leq r^n$,
- where $0 < r < 1$, for all i, j and $n \geq 1$.

Substantive meaning of the theorem

- In a finite state space with all probabilities positive, there is a unique invariant distribution.
- For large enough n, the initial state plays almost no role.

Markov Chains

More generalized Theorem

- Let P be irreducible and aperiodic over a finite state space.
- Then there exists a unique probability distribution π_j, $j \in S$
- such that $\sum_i \pi_i p_{ij} = \pi_j \forall j \in S$
- and $|p_{ij}^{(n)} - \pi_j| \leq r^{n/\nu}$,
- where $0 < r < 1$, for all i, j and for some positive integer ν.

Markov Chains

More generalized Theorem

- Let P be irreducible and aperiodic over a finite state space.
- Then there exists a unique probability distribution π_j, $j \in S$
- such that $\sum_i \pi_i p_{ij} = \pi_j \forall j \in S$
- and $|p_{ij}^{(n)} - \pi_j| \leq r^{n/\nu}$,
- where $0 < r < 1$, for all i, j and for some positive integer ν.

- Irreducible: starting from state i, the process can reach any other state with positive probability.
- Aperiodic: if $p_{ii}^{(n)} > 0 \; \forall i$ and for sufficiently large n.

Markov Chains in Continuous Spaces

Markov-Process taking values in $\mathbb{R}$

- $f_{(X_1,\ldots,X_n|X_0=x_0)}(x_1,\ldots,x_n) = p(x_0,x_1)p(x_1,x_2)\cdots p(x_{n-1},x_n)$
- $p(x,y)$: Transitional kernel
- Invariant density: $\pi(y) = \int_{\mathbb{R}} \pi(x)p(x,y)dx$.