STA2201H Methods of Applied Statistics II

Monica Alexander

Week 4: More Bayes and MCMC

Recap

- Posterior is proportional to the likelihood times the prior
- The prior encompasses knowledge to date (before data)
  - Prior setting involves choice
  - Pragmatic approach is to choose weakly informative prior
- Knowledge is then updated based on data collected
- Bayesian inference revolves around inference based on the posterior

More than one parameter

More than one parameter

What if the data model (likelihood function) includes more than 1 unknown parameter, e.g. do inference for μ if

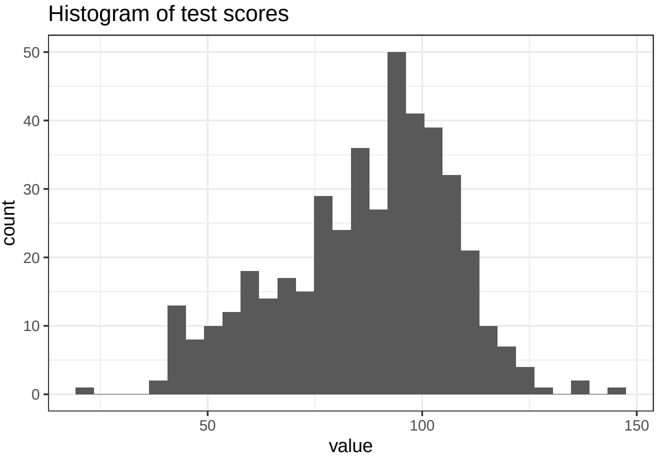
$$y_i \sim N(\mu, \sigma^2)$$

What do we want? $p(\mu, \sigma|\mathbf{y})$. If only mean is of interest, then want $p(\mu|\mathbf{y})$.

Example: kid's test scores

Gelman-Hill Chapter 3 Outcome of interest: cognitive tests scores for 3-4 year old kids. Denote the unknown mean test score by μ, and observed test score by y_i for kid i, with $i = 1, \ldots, n$.

Goal: estimate μ.



Example: kid's test scores

Let's assume Normal likelihood

$$y_i \sim N(\mu, \sigma^2)$$

- If we put a joint prior $p(\mu, \sigma)$ on the parameters, Bayes' rule tells us how to get the joint posterior distribution:

$$p(\mu, \sigma | \mathbf{y}) = \frac{p(\mathbf{y} | \mu, \sigma) p(\mu, \sigma)}{p(\mathbf{y})}$$

- And if inference about μ is our goal, we can get the marginal posterior distribution

$$p(\mu | \mathbf{y}) = \int_{\sigma} p(\mu, \sigma | \mathbf{y}) d\sigma$$

Example: kid's test scores

What priors to set for μ and σ^2? Let's assume that μ and σ^2 are independent a priori, $p(\mu, \sigma) = p(\mu)p(\sigma)$, and use

$$\mu \sim N(\mu_0, \sigma^2_{\mu_0})$$

and

$$1/\sigma^2 \sim Gamma(\nu_0/2, \nu_0/2 \cdot \sigma_0^2)$$

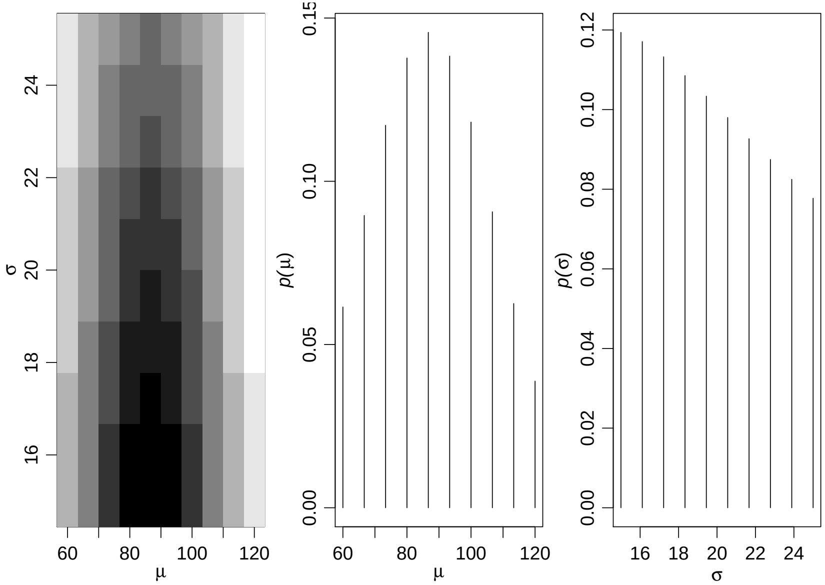
- The parameters of prior distributions are called **hyperparameters**
- For illustrative purposes, will set **hyperparameters** to be $\mu_0 = 86.8$, $\sigma_{\mu_0} = \sigma_0 = 20.4$ and $\nu_0 = 1$.

Let's start with a discrete approximation

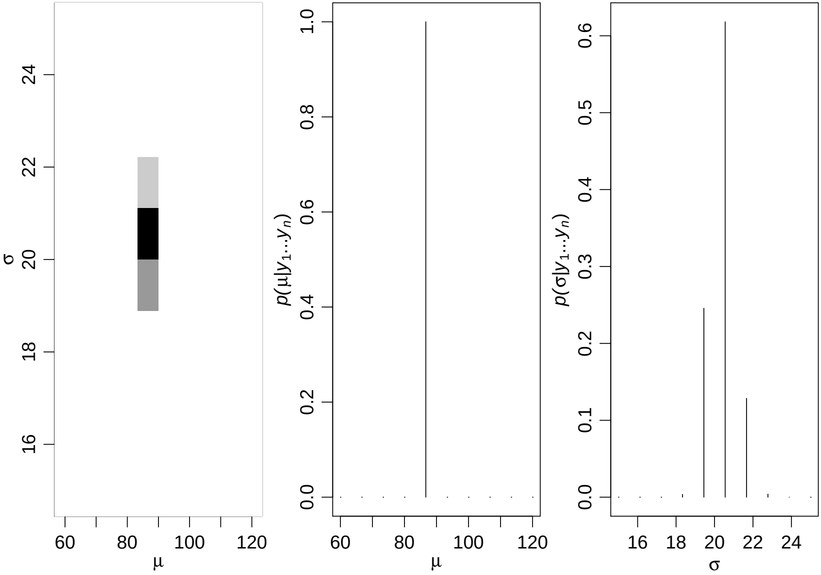
For illustrative purposes, start with a discrete approximation to these priors.

- E.g. use discrete grid of values for μ, and set $p(\mu) = f(\mu)/\sum f(\mu)$ where $f(.)$ is given by the Normal pdf for μ.
- Can use these to calculate discrete likelihood and thus a discrete approximation to the posterior

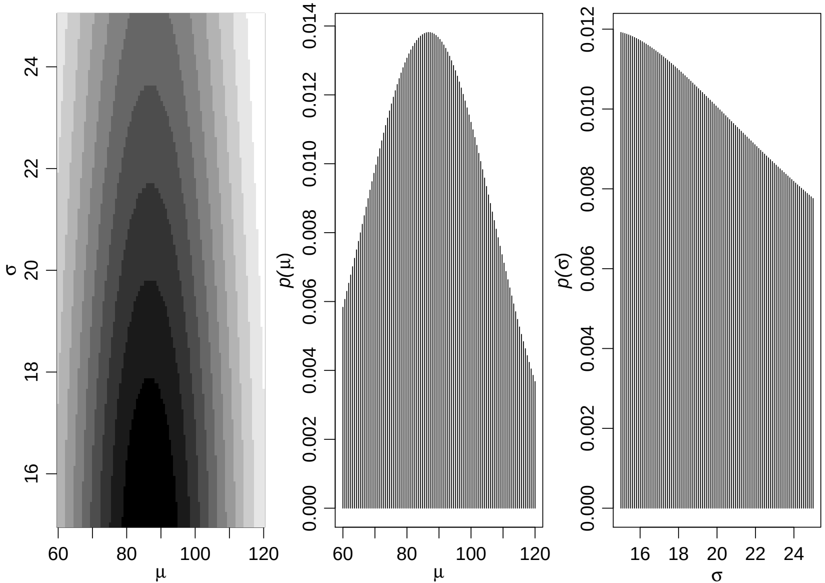
Joint and marginal prior distributions



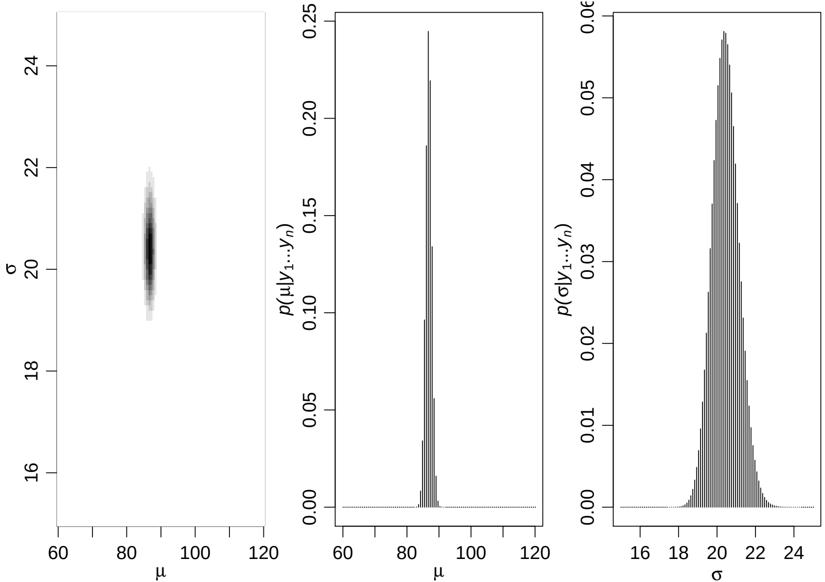
Joint and marginal posterior distributions



Finer grid: priors



Finer grid: posteriors



How did I get those previous graphs?

- Defined a grid of μ and σ values
  - e.g. first example for μ was `seq(60,120,length=10)`
- Calculated density at each grid point
  - e.g. using `dnorm`
- Standardized e.g. $p(\mu) = f(\mu)/\sum f(\mu)$
- Calculated prior grid $p(\mu) \cdot p(\sigma)$
- Calculated posterior grid $p(\mu, \sigma|y) = \frac{p(y|\mu,\sigma) \cdot p(\mu) \cdot p(\sigma)}{p(y)}$
- Calculated marginals of posterior grid by summing over relevant parameter e.g. $p(\mu|y) = \sum_\sigma p(\mu, \sigma|y)$

Now with continuous priors

$$p(\mu, \sigma|\mathbf{y}) = \frac{p(\mathbf{y}|\mu, \sigma)p(\mu, \sigma)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mu, \sigma)p(\mu, \sigma)}{\int_{\mu}\int_{\sigma} p(\mathbf{y}|\mu, \sigma)p(\mu, \sigma)d\sigma d\mu}$$

The bad news:

- Common choices of priors (e.g. what we have chosen) do not result in a closed-form expression for these posterior distributions

The good news:

- Not a problem if we can obtain a sample from the posterior distribution, which is very common in Bayesian inference.

Simulation-based inference and Monte Carlo

The general idea in simulation-based inference: we can make inference about a random variable μ, using a sample $\mu^{(1)}, \ldots, \mu^{(S)}$ from its probability distribution. This is called a **Monte Carlo (MC)** approximation.

```
my_sample <- rnorm(5000, mean = 0, sd = 1)
mean(my_sample)
```

```
## [1] 0.01147015
```

```
sd(my_sample)
```

```
## [1] 1.003294
```

Monte Carlo

- Why can we use a sample mean as an approximation to the mean of a random variable?
- Just about any aspect of the distribution of μ can be approximated arbitrarily exactly with a large enough Monte Carlo sample, e.g.
  - the α-percentile of $\mu^{(1)}, \ldots, \mu^{(S)} \rightarrow$ the α-percentile of the distribution, e.g. the median
  - We can approximate $Pr(\mu \geq x)$ for any constant x by the proportion of samples for which $\mu \geq x$, because

$$1/S \sum_{s=1}^{S} I(\mu^{(s)} \geq x) \rightarrow Pr(\mu \geq x)$$

Monte Carlo

- With a simulation, it also becomes very easy to analyze the distributions of any function of 1 or more random variables, e.g.
  - use $1/\mu^{(s)}$ to study $1/\mu$
- Samples from marginal distributions may be obtained from samples from joint distributions, e.g.
  - the distribution of μ_1, where $(\mu_1, \mu_2) \sim N_2(\mathbf{0}, \Sigma)$ can be studied using samples of $\mu_1^{(s)}$ where $(\mu_1^{(s)}, \mu_2^{(s)}) \sim N_2(\mathbf{0}, \Sigma)$

A note on simulation-based inference in general

- We are going to be talking about and using samples to make inferences a lot, because it's usually the most tractable way of making inferences from the posterior distribution
- But simulation techniques are super useful more generally
- Simulate datasets, run models, make sure you get back what you put in
- Simulate to understand sampling behavior

Summary

Bayes rule for more than one parameter

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

- $p(\mathbf{y}) = \int_{\theta'} p(\mathbf{y}|\theta')p(\theta')d\theta'$
- The marginal posterior for just one parameter is given by $p(\theta_1|\mathbf{y}) = \int_{\theta'_2} \cdots \int_{\theta'_p} p(\theta_1, \theta'_2, \ldots \theta'_p|\mathbf{y})d\theta'_2 \ldots d\theta'_p$.
- Problem: often don't have a closed form solution for posterior
- Solution: We can make inference about any random variable θ, using a sample from its probability distribution. This is called a Monte Carlo (MC) approximation.
- If we are able to obtain a sample from posterior $p(\theta|\mathbf{y})$ then
  - for each parameter we have a sample of its marginal e.g. $\theta_1^{(1)}, \ldots, \theta_1^{(S)} \sim p(\theta_1|y)$.
  - we can report any summary we'd like, e.g. posterior mean (sample mean), posterior median or other percentiles (sample percentiles).

Marginalization and sampling: key points

- Joint distribution of parameters

$$p(\theta_1, \theta_2 \mid y) \propto p(y \mid \theta_1, \theta_2) p(\theta_1, \theta_2)$$

- Marginalization

$$p(\theta_1 \mid y) = \int p(\theta_1, \theta_2 \mid y) d\theta_2$$

  $p(\theta_1 \mid y)$ is a marginal distribution

- Monte Carlo approximation

$$p(\theta_1 \mid y) \approx \frac{1}{S} \sum_{s=1}^{S} p(\theta_1, \theta_2^{(s)} \mid y),$$

  where $\theta_2^{(s)}$ are draws from $p(\theta_2 \mid y)$

MCMC

Reading

- Hoff (eBook through library) Chapters 6,9
- BDA Chapters 10-13
- Will only give very hand-wavy overview of HMC. More details:
  - Neal, MCMC using Hamiltonian dynamics: https://arxiv.org/pdf/1206.1901.pdf
  - Betancourt, A Conceptual Introduction to Hamiltonian Monte Carlo: https://arxiv.org/abs/1701.02434
- Stanislav Ulam's autobiography, 'Adventures of a Mathematician', is a great read
- Arianna Rosenbluth's obituary

Where are we at

- Bayesian inference revolves around inference based on the posterior
- Posterior usually hard to write down in closed form
- But as long as we can get a set of samples from posterior, we can do inference

Where are we going

- How do we get samples from posterior?
- How do we run models in R?
- How do we check models?

Gibbs Sampling

Gibbs Sampling

Instead of trying to sample directly from $p(\mu, \sigma|\mathbf{y})$, sample parameters sequentially from their **full conditional** distributions (conditioning on data as well as all other model parameters).

Given starting values $\mu^{(1)}$ and $\sigma^{(1)}$, draw samples $s = 1, 2, \ldots$ some large number as follows:

1. sample $\mu^{(s+1)}$ from $p(\mu|\mathbf{y}, \sigma^{(s)})$
2. sample $\sigma^{(s+1)}$ from $p(\sigma|\mathbf{y}, \mu^{(s+1)})$

Gibbs Sampling for μ and σ when $y_i \sim N(\mu, \sigma^2)$

In full conditional distributions, data and all other model parameters are assumed known. So let's figure out what they are by considering the following settings:

- Obtain $p(\mu|\mathbf{y}, \sigma^{(s)})$, in other words, obtain the posterior for μ when assuming that σ is known.
- Obtain $p(\sigma|\mathbf{y}, \mu^{(s)})$, in other words, obtain the posterior for σ when assuming that μ is known.

Estimating mean test score

Assuming σ is known

It turns out that with a normal prior on μ

$$\mu \sim N(\mu_0, \sigma^2_{\mu_0})$$

and a likelihood function

$$p\left(\mathbf{y}|\mu, \sigma^2\right) = \prod_{i=1}^{n} p\left(y_i|\mu\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2\sigma^2}\left(y_i - \mu\right)^2\right)$$

The posterior is normal too:

$$\mu|\boldsymbol{y}, \sigma^2 \sim N\left(\frac{\mu_0/\sigma^2_{\mu 0} + n \cdot \bar{y}/\sigma^2}{1/\sigma^2_{\mu 0} + n/\sigma^2}, \frac{1}{1/\sigma^2_{\mu 0} + n/\sigma^2}\right)$$

How did the normal posterior come about?

$$p\left(\mu|\boldsymbol{y},\sigma^2\right) \propto p(\mu)p\left(\boldsymbol{y}|\mu,\sigma^2\right) \text{ (Bayes' rule)}$$

$$\propto \exp(\frac{-1}{2\sigma_{\mu0}^2}\left(\mu-\mu_0\right)^2\Bigg) \cdot \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right) \propto \exp\left(-\frac{1}{2}f(\mu)\right)$$

Short detour

What's the interpretation here?

$$\mu|\boldsymbol{y},\sigma^2 \sim N\left(\frac{\mu_0/\sigma^2_{\mu 0} + n \cdot \bar{y}/\sigma^2}{1/\sigma^2_{\mu 0} + n/\sigma^2}, \frac{1}{1/\sigma^2_{\mu 0} + n/\sigma^2}\right)$$

- What happens when n is big?

μ known, σ unknown

Steps for Bayesian inference about σ are same as before but easiest set-up is based on $1/\sigma^2$, which is called the precision:

$$p\left(1/\sigma^2|\boldsymbol{y},\mu\right) = \frac{p\left(1/\sigma^2\right)p(\boldsymbol{y}|\mu,\sigma)}{p(\boldsymbol{y})} \propto p\left(1/\sigma^2\right)p(\boldsymbol{y}|\mu,\sigma)$$

The Gamma distribution for the precision is a conjugate prior :

- If $1/\sigma^2 \sim \text{Gamma}\left(\nu_0/2, \nu_0/2 \cdot \sigma_0^2\right)$, then

$$1/\sigma^2|\boldsymbol{y},\mu \sim \text{Gamma}\left(\nu_n/2, \nu_n/2 \cdot \sigma_n^2\right)$$

with

$$\begin{aligned} \nu_n &= \nu_0 + n \\ \sigma_n^2 &= 1/\nu_n\left(\nu_0\sigma_0^2 + ns_n^2\{\mu\}\right) \\ s_n^2\{\mu\} &= 1/n\sum\left(y_i-\mu\right)^2 \end{aligned}$$

μ and σ unknown

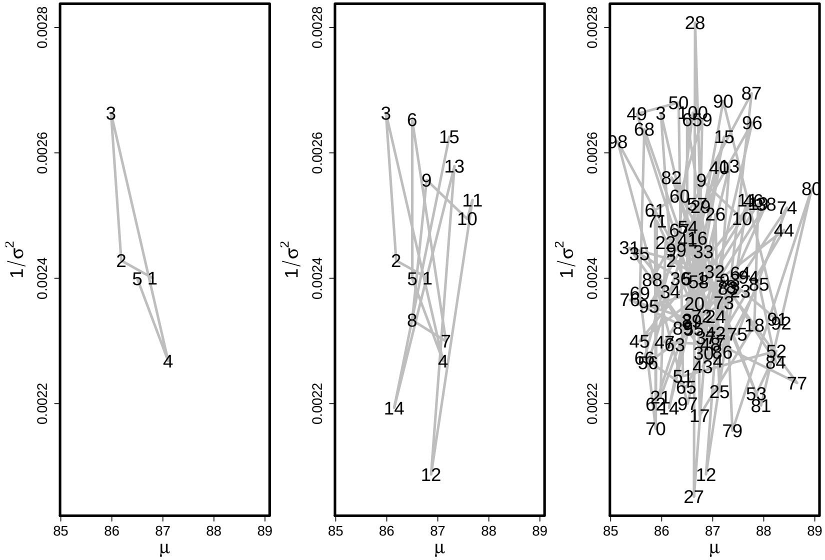
Use same priors as before such that we know what these full conditionals are:

$$
\begin{aligned}
\mu &\sim N\left(\mu_0, \sigma^2_{\mu 0}\right) \\
\mu | \boldsymbol{y}, \sigma^2 &\sim N\left(\frac{\mu_0/\sigma^2_{\mu 0} + n \cdot \bar{y}/\sigma^2}{1/\sigma^2_{\mu 0} + n/\sigma^2}, \frac{1}{1/\sigma^2_{\mu 0} + n/\sigma^2}\right) \\
1/\sigma^2 &\sim \text{Gamma}\left(\nu_0/2, \nu_0/2 \cdot \sigma^2_0\right) \\
1/\sigma^2 | \boldsymbol{y}, \mu &\sim \text{Gamma}\left(\nu_n/2, \nu_n/2 \cdot \sigma^2_n\right)
\end{aligned}
$$

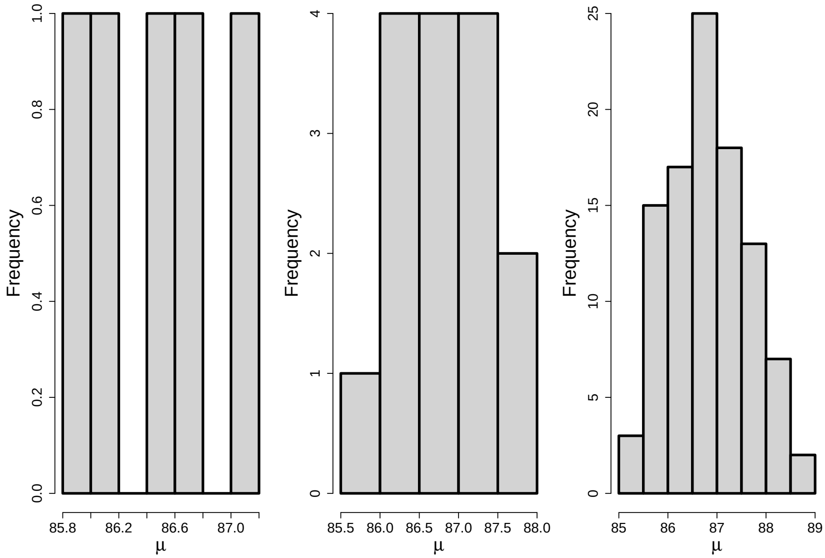
with

$\nu_n = \nu_0 + n; \sigma^2_n = 1/\nu_n\left(\nu_0\sigma^2_0 + ns^2_n\{\mu\}\right)$
$n \cdot s^2_n\{\mu\} = \sum\left(y_i - \mu\right)^2 = (n-1)s^2\{y\} + n(\bar{y} - \mu)^2$(Hoff p.94)
and $\mu_0 = \hat{\mu}, \sigma_{\mu 0} = s\{y\}, \nu_0 = 1$ and $\sigma_0 = s\{y\}$

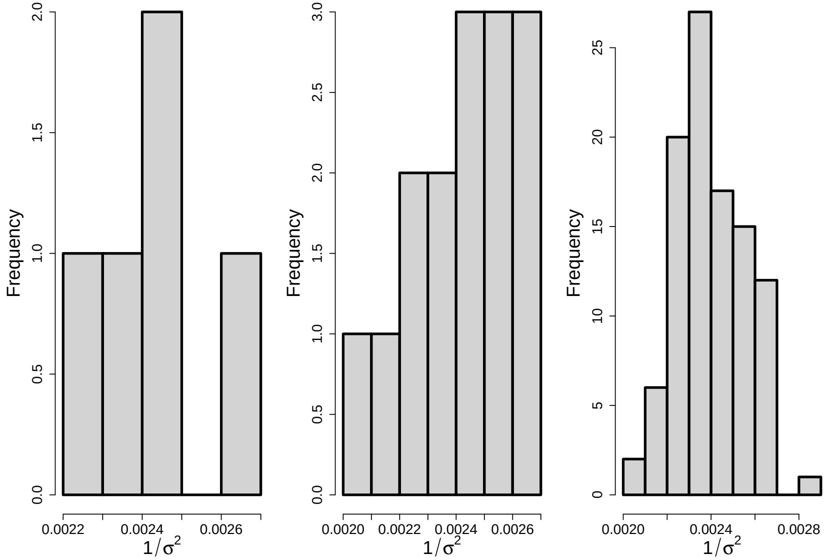
Gibbs Sampling (first 5, 10, 100 iterations)



Histograms of μ (5, 15, 100 samples)



Histograms of $1/\sigma^2$ (5, 15, 100 samples)



Gibbs Sampling

- When doing inference for μ when σ is known, and inference for σ when μ is known, conjugate priors can be used and the posteriors are available in closed form.
- When doing inference for both simultaneously, we can still use Bayes' rule but finding closed-form expressions for the (joint) posteriors is hard/impossible, so can use sampling.
- Gibbs sampling is one of such methods, where (subsets of) parameters are sampled sequentially from their full conditional distributions (given by conditioning on the data as well as all other parameters).

MCMC

MCMC

The Gibbs sampling algorithm is a Markov Chain Monte Carlo (MCMC) algorithm.

The **Markov Chain** part

- Let $\boldsymbol{\phi}^{(s)} = \left(\mu^{(s)}, \sigma^{(s)}\right)$ be the s-th draw of parameters.
- $\phi^{(s)}$ depends on $\phi^{(s-1)}, \phi^{(s-2)}, \ldots, \phi^{(1)}$ only through $\phi^{(s-1)}$
- This is called the Markov property, and so the sequence is called a Markov chain
- Each sample $\phi^{(s)}$is drawn such that eventually, for some large s, $\phi^{(s)}$ is a draw from the target distribution, which in this example is the posterior distribution $(\mu, \sigma)|\boldsymbol{y}$.

The **Monte Carlo** part

- We approximate quantities of interest, e.g. $E(\mu|\boldsymbol{y})$, using resulting samples.

MCMC: beyond Gibbs

- In Gibbs Sampling we used conjugate priors to write down the full conditionals
- Sometimes conjugate priors are not available or are unsuitable.
- For more complex models, we often cannot find closed form solutions for the posterior

Any algorithm for obtaining samples from a target distribution (e.g. $p(\theta|y)$) whereby the s-th sample $\theta^{(s)}$ depends on $\theta^{(s-1)}$ is called an MCMC algorithm.

MCMC: intuition

- Suppose we already have $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(s)}$ and we have another θ. Should we add θ to our samples?
- Want to compare $p(\theta|y)$ to $p(\theta^{(s)}|y)$
- But we don't need to compute the posteriors

$$\frac{p(\theta|y)}{p\left(\theta^{(s)}|y\right)} = \frac{p(y|\theta)p(\theta)p(y)}{p(y)p\left(y|\theta^{(s)}\right)p\left(\theta^{(s)}\right)} = \frac{p(y|\theta)p(\theta)}{p\left(y|\theta^{(s)}\right)p\left(\theta^{(s)}\right)}$$

- If $p(\theta|y) > p\left(\theta^{(s)}|y\right)$ we set $\theta = \theta^{(s+1)}$.
- If $r = p(\theta|y)/p\left(\theta^{(s)}|y\right) < 1$ we expect to have θ appear r times as often as $\theta^{(s)}$ so we accept θ with probability r.

The Metropolis algorithm

1. Propose a new θ^* which is a draw from a symmetric distribution around $\theta^{(s)}$, call it $J(\theta^*|\theta^{(s)})$ (symmetry means $J_t\left(\theta_a|\theta_b\right) = J_t\left(\theta_b|\theta_a\right)$)
2. Calculate the ratio

$$r = \frac{p\left(\theta^*|\boldsymbol{y}\right)}{p\left(\theta^{(s)}|\boldsymbol{y}\right)} = \frac{p\left(\theta^*\right)p\left(\boldsymbol{y}|\theta^*\right)}{p\left(\theta^{(s)}\right)p\left(\boldsymbol{y}|\theta^{(s)}\right)}$$

3.
   - If $r > 1$, set $\theta^{(s+1)} = \theta^*$.
   - If $r < 1$, set $\theta^{(s+1)} = \theta^*$ with probability r and $\theta^{(s+1)} = \theta^s$ with probability $1 - r$.

The Metropolis-Hasting algorithm

Extension to non-symmetric proposal distributions. To correct for the asymmetry, the ratio r is replaced by a ratio of ratios:

$$r = \frac{p\left(\theta^{*}|y\right)/J_{s+1}\left(\theta^{*}|\theta^{s}\right)}{p\left(\theta^{s}|y\right)/J_{s+1}\left(\theta^{s}|\theta^{*}\right)}$$

- Metropolis a special case of MH
- Gibbs a special case of MH with $r = ?$

Does it work? (in theory)

Supplementary details: Hoff Chapter 10

- For most problems, we can construct an MH algorithm that generates an *irreducible*, *aperiodic* and *recurrent* Markov chain, that converges to a stationary distribution which is the posterior distribution of interest.
- What that means is: Even though samples are obtained sequentially (and depend on the previous sample), eventually, the *s*-th draw can be considered to be a random draw from the target distribution, i.e.:

$$\lim_{s \to \infty} \Pr\left(\theta^{(s)} \in A\right) = \int_A p(\theta|\boldsymbol{y}) d\theta$$

Irreducible, aperiodic and recurrent

What is an irreducible, aperiodic and recurrent Markov chain?

- Irreducible: able to move from one value of θ with $p(\theta|\boldsymbol{y}) > 0$ to another with non-zero probability mass
- Aperiodic: no periodic values for θ with $p(\theta|\boldsymbol{y}) > 0$
- Recurrent: if we start at θ with $p(\theta|\boldsymbol{y}) > 0$, we are guaranteed to return to it.

Once you are sampling from the stationary distribution, you are always sampling from the stationary distribution.

Proposal distributions

In most problems it is not hard to construct MH algorithms that generate Markov chains that are irreducible, aperiodic and recurrent.

What should we choose for our proposal distribution $J(\theta^*|\theta^{(s)})$?

A normal proposal distribution centered at the current value works. This is called **Random Walk Metropolis**

- take our current position and add some noise

$$J(\theta^*|\theta^{(s)}) \sim N(\theta^{(s)}, \sigma^2)$$

i.e.

$$\theta^* = \theta^{(s)} + \epsilon^*$$

Issues with random walk Metropolis:

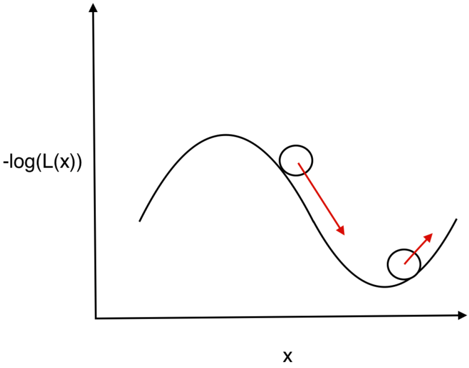
- slow
- highly correlated samples
- breaks down in high dimensions

Proposal distributions

Common alternative is the class of proposal distributions that lead to **Hamiltonian Monte Carlo**. Not going into detail here (see Neal and Betancourt in readings for more info), but

- Physical analogy of a particle moving around some parameter space i.e. treat the Markov chain like a physical particle.
- The 'energy' of particle (potential and kinetic) is dependent on its position and momentum. We know position, we generate momentum.
- Treat the negative log density like a physical potential.
- Proposal distributions makes use of gradient information of the log of the posterior in order to move more quickly toward regions of high probability.
- Give the particle a random shove instead of a random step.
- When it fails, it fails badly (cf RWMH, which fails silently)

HMC



HMC

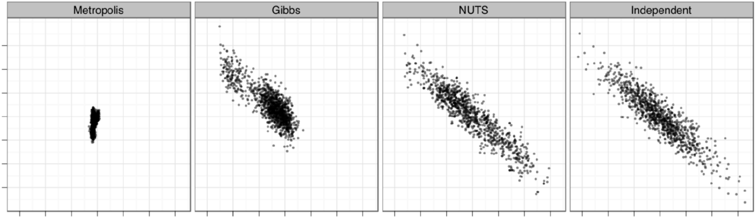


Figure from Hoffman and Gelman

Where are we at

- For most problems, we can construct an MCMC algorithm that generates an irreducible, aperiodic and recurrent Markov chain, that converges to a stationary distribution which is the posterior distribution of interest.
- Then, eventually, an MCMC sample is a draw from its target distribution, hence MCMC algorithms can be used to generate samples from posterior distributions. . .

Can we just use all the samples?

Does it work (in practice?)

MCMC samples are NOT independent draws from a target distribution:

- The first draw is set by the user and thus not a random draw from the target distribution.
- Draw $s + 1$ is correlated with draw s

We can use samples from an MCMC algorithm to do inference but ONLY IF we have "waited long enough" for those samples to be representative of the distribution of interest.

Need to (try and) check:

- That the MCMC chain has converged away from the initial values into the target (stationary) distribution.
- That the chain has generated a representative sample

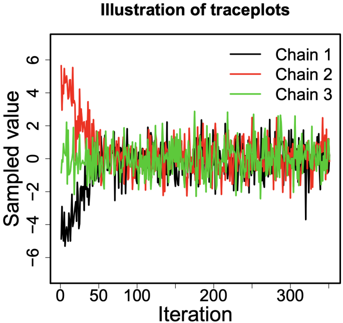
MCMC diagnostic tools

- Trace plots
- Effective sample size
- $\hat{R}$

Trace plots

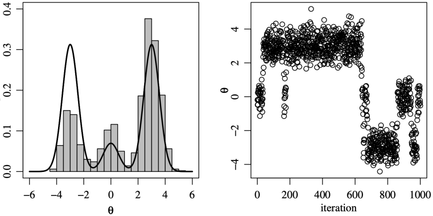
Trace plots: can be used to detect **burn-in** and **mixing**.

- Initial MCMC samples may not be representative of the posterior distribution.
- The initial phase of an MCMC chain is called the burn-in phase, during which the chain converges towards the target distribution.
- Samples from the burn-in period should be discarded.
- Chains should be ‘mixing’ well



Effective sample size

S_{eff} = the number of independent MC samples that would give the same precision for the mean as the MCMC samples, estimated by $\bar{\theta} = 1/S \sum_s \theta^{(s)}$, as obtained with the MCMC sample of size S.



Hoff Figure 6.5, $S_{eff} = 18.4$ ($S = 1,000$).

Effective sample size

- The precision (standard error) of the mean with MC samples is just
$$Var(\bar{\theta})_{MC} = \frac{Var(\theta)}{S}$$
- But for MCMC, the samples S are not independent, and they are, in general, positively correlated
- So in general $Var(\bar{\theta})_{MCMC} > Var(\bar{\theta})_{MC}$
- The effective sample size is calculated such that

$$\mathsf{Var}_{\mathrm{MCMC}}[\bar{\theta}] = \frac{\mathsf{Var}[\theta]}{S_{\mathrm{eff}}}$$

so that S_{eff} can be interpreted as the number of independent Monte Carlo samples necessary to give the same precision as the MCMC samples.

Potential scale reduction factor $\hat{R}$

The Gelman-Rubin convergence diagnostic statistic $\hat{R}$ (Gelman & Rubin 1992), or potential scale reduction factor/shrink factor, is based on a comparison between the average variance of samples within each chain to the variance of the pooled samples across chains

- If chains have mixed well, then $\hat{R}$ is close to 1
- Rules of thumb: Aim for $\hat{R} < 1.05$, and S_{eff} greater than 100.

More when we start fitting in R

Lab

- Webscraping
- No lab to hand in because of A1
- But please make sure you can install Stan and run example with Rmd without problems.