Markov Chain Monte Carlo

- Introduction to Bayesian Inference
  - Prior Distribution
  - Posterior Distribution
- Markov Chain Monte Carlo
  - Gibbs Sampling
  - Metropolis-Hasting
- Germination Rate Example

Introduction to Bayesian Inference

- Classical statistical inference
  - Parameters are fixed unknown quantities
  - Focus on the variability of the estimator

$$f(\widehat{\boldsymbol{\sigma}}; \boldsymbol{\sigma})$$

  - Hypothesis testing
  - Confidence intervals
  - Estimation

- Bayesian inference
  - Parameters are a realization of a random process

$$\boldsymbol{\sigma} \sim \pi(\boldsymbol{\sigma})$$

  - Random process used to model uncertainty
  - At some point the random process is treated as known
  - Prediction

$$\boldsymbol{\sigma}|\boldsymbol{y} \sim \pi(\boldsymbol{\sigma}|\boldsymbol{y})$$

Similar to the distinction we make when we talk about fixed effects and random effects

- Classical statistical inference/Fixed Effects
  - $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-})$
  - Estimable functions
  - $\mathrm{H}_O : \boldsymbol{K}'\boldsymbol{\beta} = 0$
  - Confidence Intervals
  - Never talk about the distribution of $\boldsymbol{\beta}$

- Bayesian Inference/Random Effects
  - $\boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$
  - $\boldsymbol{u}|\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{D} - \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\boldsymbol{D})$
  - Prediction intervals
  - Don't talk about testing

A simple example

$$\boldsymbol{y}|\boldsymbol{\sigma^2} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$$

Prior Distribution:

- Specify the distribution of the parameters
- Often selected to make the mathematics tractable

– Inverse Gamma

$$\frac{1}{\sigma^2} \sim \Gamma(\kappa, \theta)$$

$$f(\sigma^2; \kappa, \theta) = \frac{(\sigma^2)^{-(\kappa+1)} e^{-1/(\theta\sigma^2)}}{\theta^\kappa \Gamma(\kappa)}$$

$$\mathrm{E}(\sigma^2) = \frac{1}{\theta(\kappa - 1)}$$

$$\mathrm{var}(\sigma^2) = \frac{1}{\theta^2(\kappa-1)^2(\kappa-2)}$$

$$\kappa = \frac{\mathrm{E}^2(\sigma^2)}{\mathrm{var}(\sigma^2)} + 2$$

$$\theta = \frac{1}{\mathrm{E}(\sigma^2)} + \frac{\mathrm{E}(\sigma^2)}{\mathrm{var}(\sigma^2)}$$

Posterior Distribution

- Distribution of the parameter given both the prior information and the data.

$$
\begin{aligned}
\pi(\boldsymbol{\sigma}, \boldsymbol{y}) &= f(\boldsymbol{y}|\boldsymbol{\sigma})\pi(\sigma) \\
\pi(\boldsymbol{\sigma}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\sigma}, \boldsymbol{y})}{\int_{\boldsymbol{\sigma}} \pi(\boldsymbol{\sigma}, \boldsymbol{y}) d\sigma} \\
&= \frac{f(\boldsymbol{y}|\boldsymbol{\sigma})\pi(\sigma)}{\int_{\boldsymbol{\sigma}} f(\boldsymbol{y}|\boldsymbol{\sigma})\pi(\sigma) d\sigma}
\end{aligned}
$$

- In our simple example

$$
\begin{aligned}
f(\boldsymbol{y}|\sigma^2) &\propto (\sigma^2)^{-N/2} \exp\left(-N/2(s^2/\sigma^2)\right) \\
\pi(\sigma^2, \boldsymbol{y}) &\propto (\sigma^2)^{-(\kappa+N/2+1)} e^{-(Ns^2/2+1/\theta)(1/\sigma^2)} \\
\pi(\sigma^2|\boldsymbol{y}) &\propto (\sigma^2)^{-(\kappa+N/2+1)} e^{-(Ns^2/2+1/\theta)(1/\sigma^2)} \\
\pi(\sigma^2|\boldsymbol{y}) &= \frac{(\sigma^2)^{-(\kappa+N/2+1)} e^{-(Ns^2/2+1/\theta)(1/\sigma^2)}}{\int (\sigma^2)^{-(\kappa+N/2+1)} e^{-(Ns^2/2+1/\theta)(1/\sigma^2)} d\sigma^2}
\end{aligned}
$$

- It turns out that $\pi(\sigma^2|\boldsymbol{y})$ has an inverse gamma distribution with parameters $\kappa + N/2$ and $\theta/(N\theta s^2/2 + 1)$.

- The posterior mean of σ^2 is

$$\widehat{\sigma}^2 = s^2 \frac{N/2}{N/2 + (\kappa - 1)} + \frac{1}{\theta(\kappa - 1)} \left[\frac{(\kappa - 1)}{N/2 + (\kappa - 1)} \right]$$

- Posterior mean estimate is a weighted average of s^2 and the mean of prior distribution $1/(\theta(\kappa - 1))$.

- What happens when the distribution is more complicated?

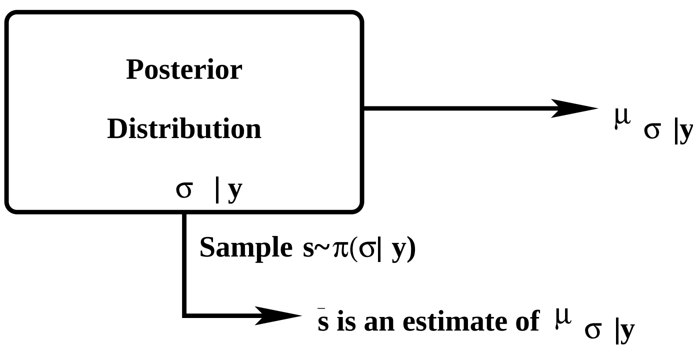
- Posterior mean

$$\mathrm{E}(\sigma^2 | \boldsymbol{y}) = \frac{\int_{\sigma^2} \sigma^2 \pi(\sigma^2, \boldsymbol{y}) d\sigma^2}{\int_{\sigma^2} \pi(\sigma^2, \boldsymbol{y}) d\sigma^2}$$

Markov Chain Monte Carlo

- Features of the posterior distribution such as the mean and variance are examples of

$$\mathrm{E}(g(\boldsymbol{\sigma})|\boldsymbol{y}) = \frac{\int_{\boldsymbol{\sigma}} g(\boldsymbol{\sigma})\pi(\boldsymbol{\sigma},\boldsymbol{y})d\boldsymbol{\sigma}}{\int_{\boldsymbol{\sigma}} \pi(\boldsymbol{\sigma},\boldsymbol{y})d\boldsymbol{\sigma}}$$



- If we could draw a sample $\boldsymbol{\sigma}_i$ from $\pi(\boldsymbol{\sigma}|\boldsymbol{y})$ then

$$\mathrm{E}(g(\boldsymbol{\sigma}_i)) = \mathrm{E}(g(\boldsymbol{\sigma}))$$

- If we draw M samples each of which is from $\pi(\boldsymbol{\sigma}|\boldsymbol{y})$ then

$$\mathrm{E}\left(\frac{\sum_i^M g(\boldsymbol{\sigma}_i)}{M}\right) = \mathrm{E}(g(\boldsymbol{\sigma}))$$

- The samples do **NOT** need to be independent samples.
  - Taking correlated samples does not create biased estimators
  - Need more correlated samples for the same level of precision

- How do we draw a sample from $\pi(\boldsymbol{\sigma}|\boldsymbol{y})$?

- How many samples do we need?

Drawing the next sample

- Markov Chain: Sample $i + 1$ ($\boldsymbol{\sigma}_{i+1}$) is based on the previous sample ($\boldsymbol{\sigma}_i$).

- Monte Carlo: Sample $i + 1$ a random sample from $q(\boldsymbol{\sigma}_{i+1}|\boldsymbol{\sigma}_{\boldsymbol{i}})$.

- $\boldsymbol{\sigma}_{i+1}|\boldsymbol{\sigma}_i \sim q(\boldsymbol{\sigma}_{i+1}|\boldsymbol{\sigma}_{\boldsymbol{i}})$

- Suppose that $f(\boldsymbol{\sigma}_i) = \pi(\boldsymbol{\sigma}|\boldsymbol{y})$, what is the distribution of $\boldsymbol{\sigma}_{i+1}$?

$$f(\boldsymbol{\sigma}_{i+1}|\boldsymbol{y}) = \int q(\boldsymbol{\sigma}_{i+1}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\boldsymbol{y})d\boldsymbol{\sigma}$$

- Need to select $q(\boldsymbol{\sigma}_{i+1}|\boldsymbol{\sigma})$

  1. Simple to compute
  2. $f(\boldsymbol{\sigma}_{i+1}|\boldsymbol{y}) \to \pi(\boldsymbol{\sigma}|\boldsymbol{y})$
  3. $f(\boldsymbol{\sigma}_{i+1}|\boldsymbol{y}) \to \pi(\boldsymbol{\sigma}|\boldsymbol{y})$ quickly
  4. $\mathrm{var}(\bar{g}(\boldsymbol{\sigma}.))$ is small

Gibbs Sampling (A special case)

$$q(\boldsymbol{\sigma}_{i,j+1}|\boldsymbol{\sigma}_i) = \pi(\sigma_{j+1}|\boldsymbol{y}, \boldsymbol{\sigma}_{i,-j})$$

- Update each component using the conditional distribution of that component given the data and the current value of the remaining components.
- Requires that $\pi(\sigma_{j+1}|\boldsymbol{y}, \boldsymbol{\sigma}_{i,-j})$ has a "nice form."

Metropolis–Hasting

$$\boldsymbol{\sigma}_{i+1} = \begin{cases} \boldsymbol{\sigma}_i & \text{with Probability } 1 - \alpha(\boldsymbol{\sigma}_?, \boldsymbol{\sigma}_i) \\ \boldsymbol{\sigma}_? & \text{with Probability } \alpha(\boldsymbol{\sigma}_?, \boldsymbol{\sigma}_i) \end{cases}$$

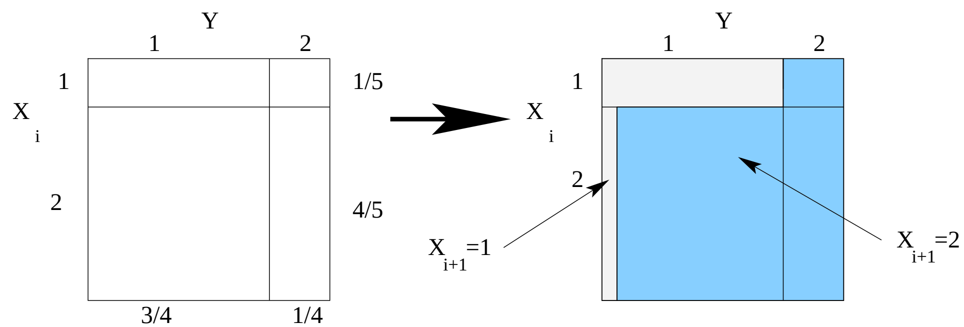
$$\boldsymbol{\sigma}_? \sim q(\boldsymbol{\sigma}_? | \boldsymbol{\sigma}_i)$$

$$\alpha(\boldsymbol{\sigma}_?, \boldsymbol{\sigma}_i) = \min\left(1, \frac{\pi(\boldsymbol{\sigma}_?, \boldsymbol{y})}{q(\boldsymbol{\sigma}_? | \boldsymbol{\sigma}_i)} \Big/ \frac{\pi(\boldsymbol{\sigma}_i, \boldsymbol{y})}{q(\boldsymbol{\sigma}_i | \boldsymbol{\sigma}_?)}\right)$$

- Draw a proposed sample from $q(\boldsymbol{\sigma}|\boldsymbol{\sigma}_i)$ call it $\boldsymbol{\sigma}_?$
- Calculate an acceptance probability $\alpha(\boldsymbol{\sigma}_?, \boldsymbol{\sigma}_i)$

- The chance will accept the proposed sample will be high if
  - Posterior for the new value is large relative to the proposed density for the new value $(\pi(\boldsymbol{\sigma}_?)/q(\boldsymbol{\sigma}_?|\boldsymbol{\sigma}_i))$
  - "Large" is measured relative the current value
- Effect is the more a proposed distribution under samples a value the more difficult it is to move away from that value
- If your proposed distribution is very light in one of the tails, then
  - You are unlikely to sample a value in that tail
  - Once you do sample a value in the tail, it will be difficult to leave that tail.

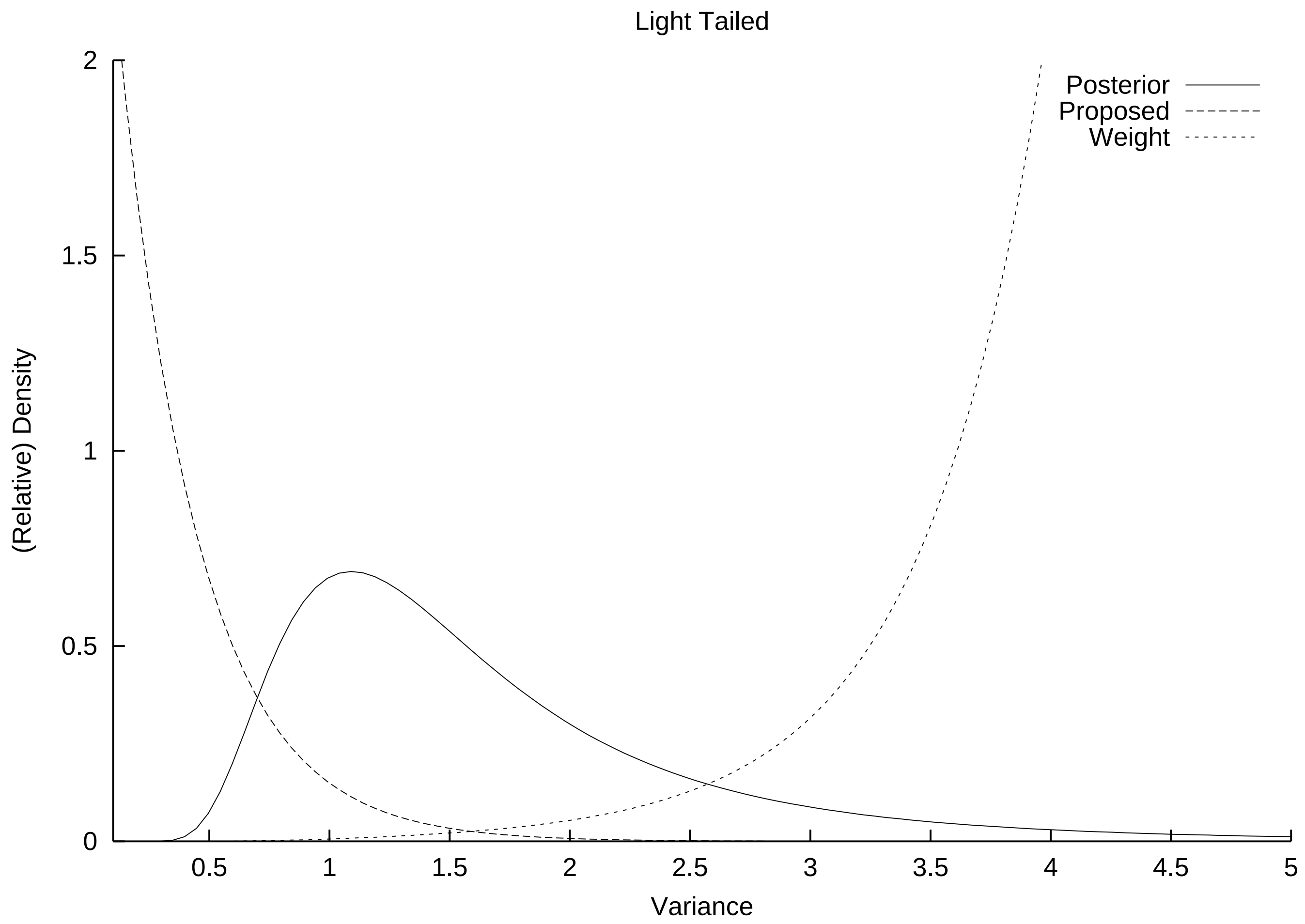
Example

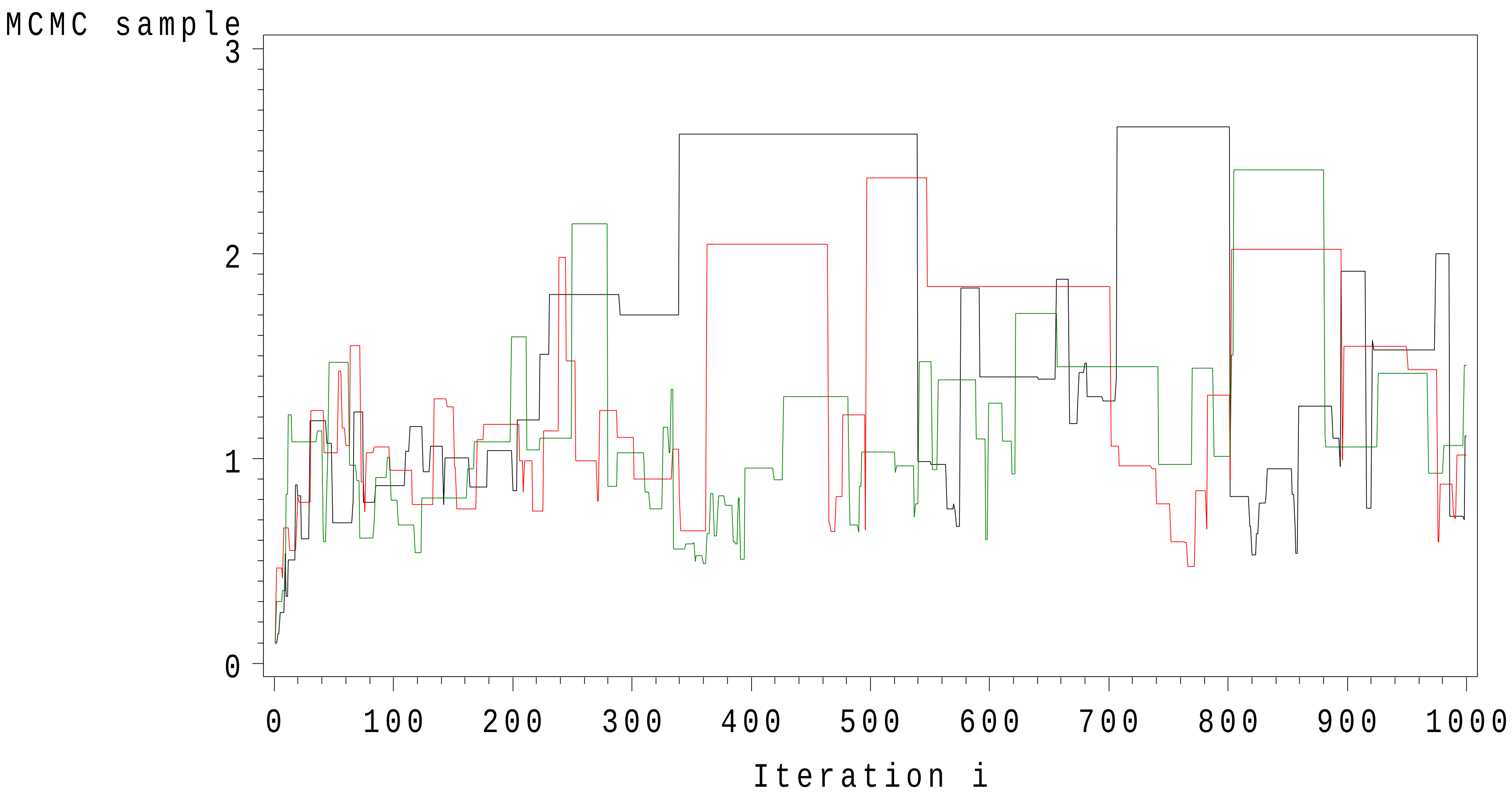


X_i	$\pi(X)$	$q(X)$	Weight	$\alpha(Y,X)$ $Y=1$	$Y=2$
1	3	.75	4	$1 = 4/4$	$1 = min(1, 48/4)$
2	12	.25	48	$1/12 = 4/48$	$1 = 48/48$
	15				

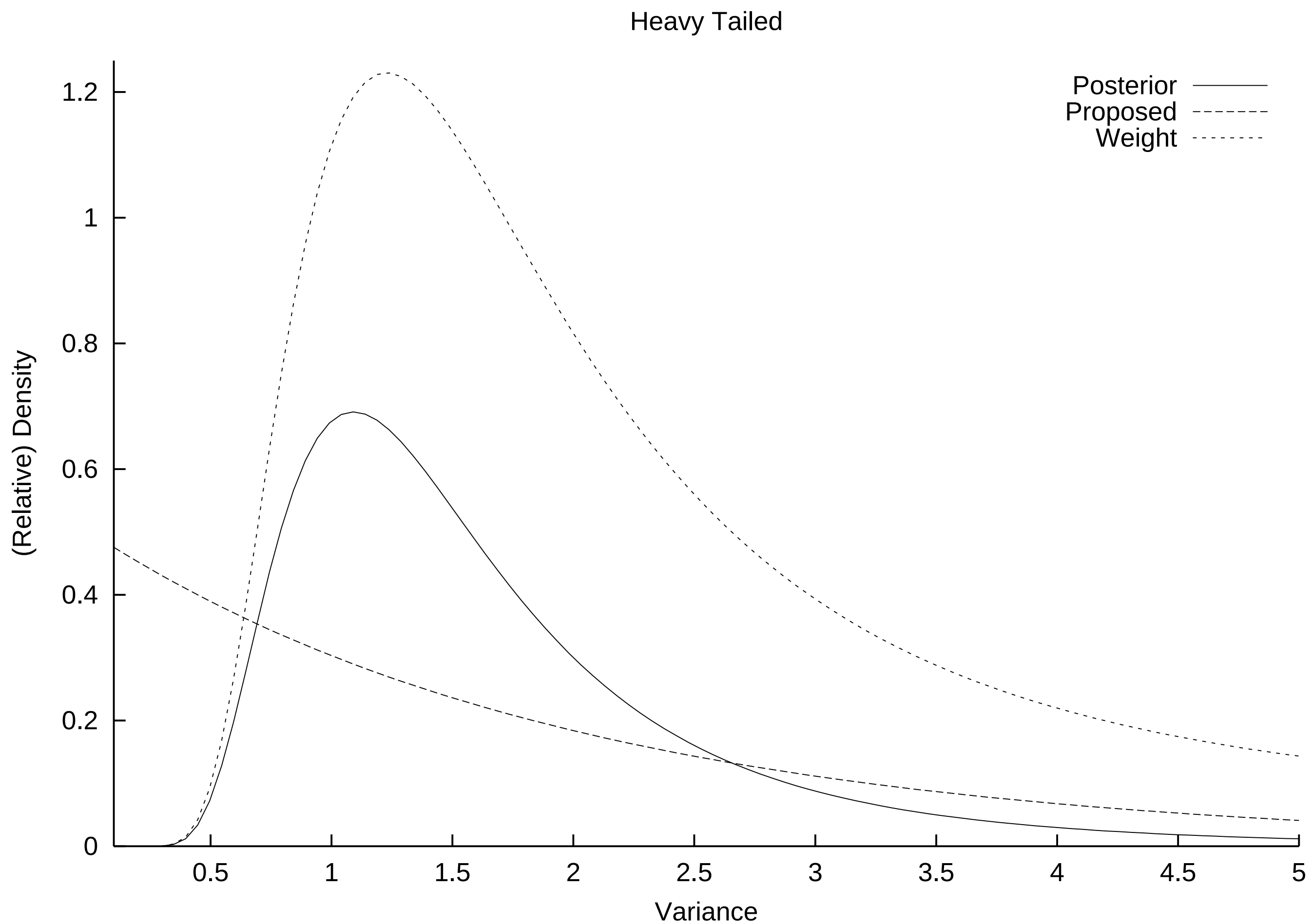
	$\Pr(Y\|X)$		
X_i	$Y=1$	$Y=2$	$\Pr(X_{i+1})$
1	.75	.25	$.75 * 3/15 + (1/16) * (12/15) = .2$
2	$.75 * \frac{1}{12}$	$.25 + .75 * \frac{11}{12}$	$.25 * \frac{3}{15} + \frac{15}{16} * \frac{12}{15}$
	$1/16$	$5/16$	$.8$

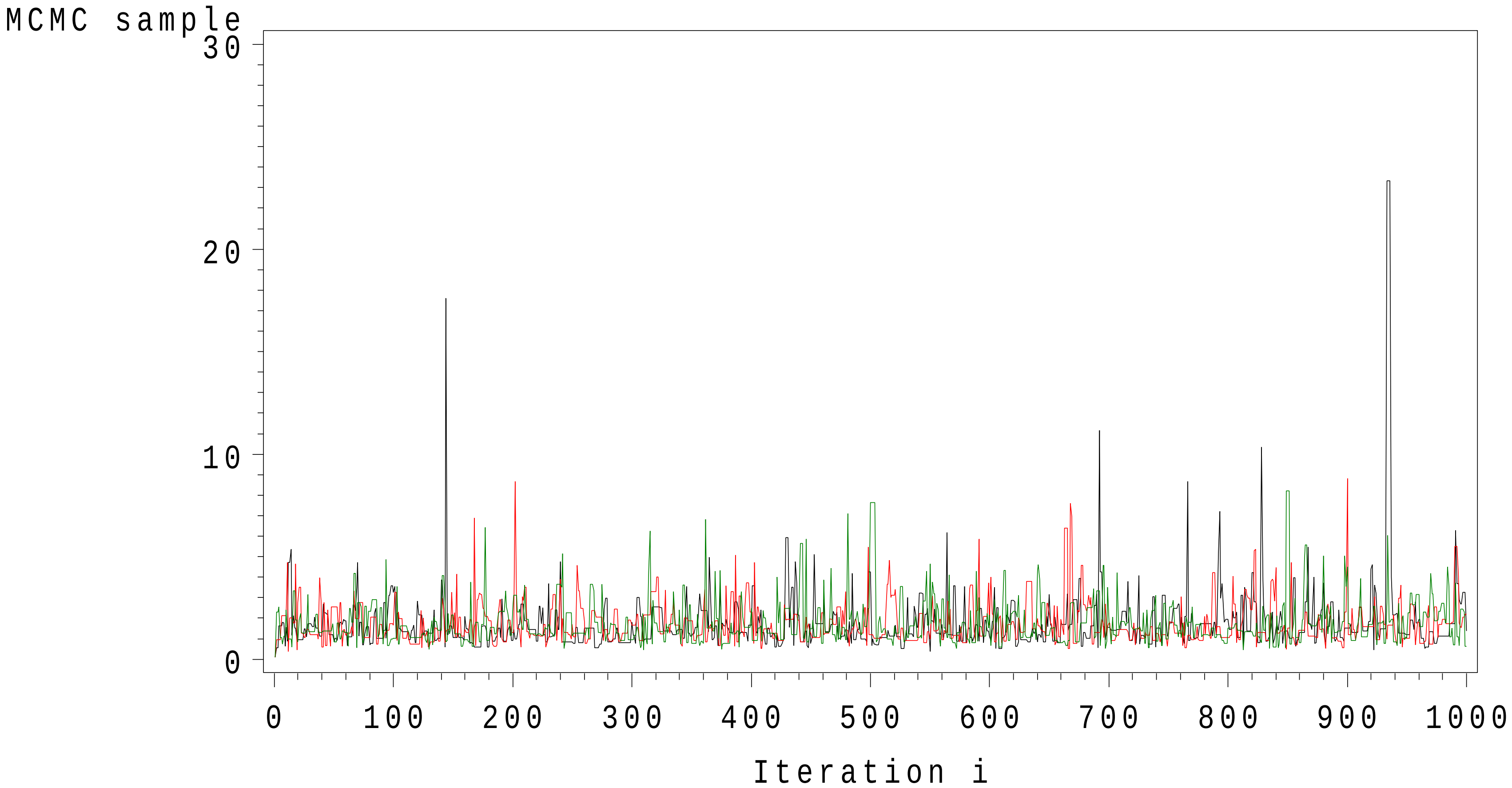
- Using the proposed distribution, $\exp(1/3)$
- Because it is very light in the upper tail, the proposed distribution will very rarely generate a large value.
- Once a large value is generated it will be difficult to return to a small value.


Light Tailed
Posterior
Proposed
Weight
(Relative) Density
Variance
0
0.5
1
1.5
2
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

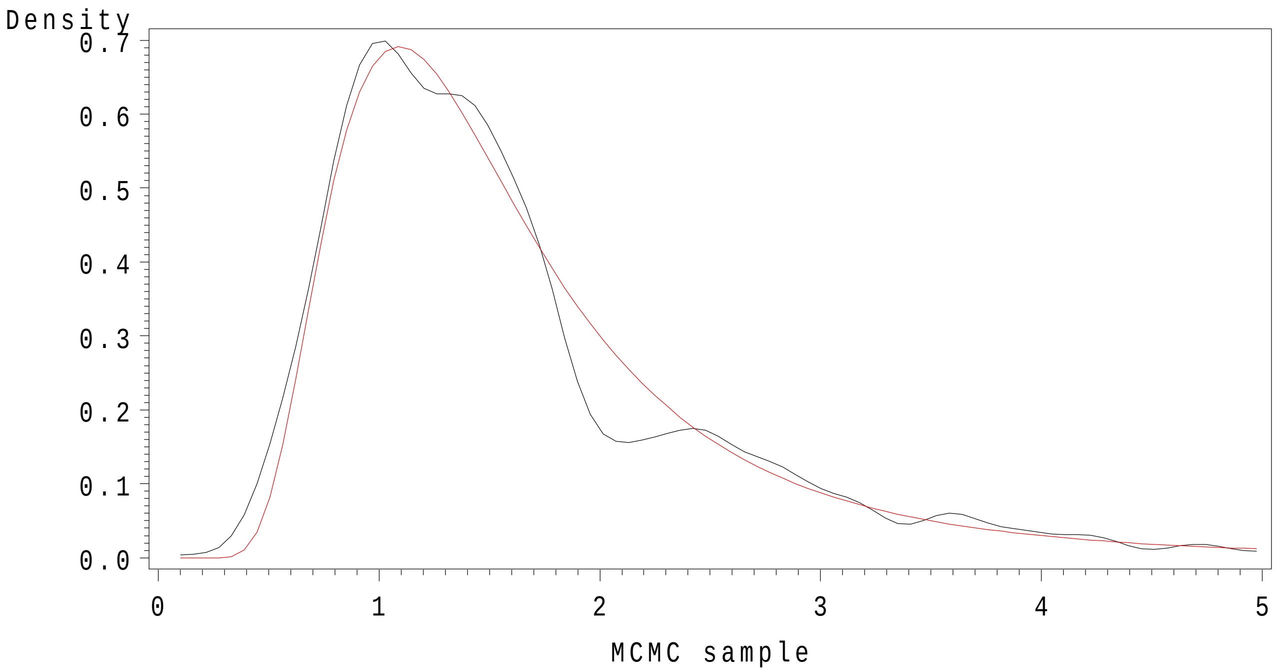

MCMC sample
3
2
1
0
0
100
200
300
400
500
600
700
800
900
1000
Iteration i

- Using the proposed distribution, $\exp(2)$
- Because it has a very heavy upper tail, the proposed distribution will generate to many large value.
- However, it will tend to reject large value in favor of more moderate values.


Heavy Tailed
Posterior
Proposed
Weight
(Relative) Density
Variance
0
0.2
0.4
0.6
0.8
1
1.2
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5


MCMC sample
30
20
10
0
0
100
200
300
400
500
600
700
800
900
1000
Iteration i

Here are the theoretical and estimated density functions from a MCMC chain of length 1,000 using the proposed distribution $\exp(2)$. The estimated density function was obtained using `Proc KDE` in `SAS`.



Germination Rate Bayesian Analysis

Need priors for $\boldsymbol{\beta}$ and σ_D^2

$$\pi(\boldsymbol{\beta}) \sim \mathrm{N}(0, 10^6)$$
$$\pi(\boldsymbol{\sigma^2}_D) \sim \Gamma^{-1}(10^{-6}, 10^6).$$

In WinBUGS the model is

```
model{
for(i in 1:N){
   r[i]~dbin(p[i],n[i]);
   eta[i]<-beta[trt[i]]+b[i];
   p[i]<-exp(eta[i])/(1.+exp(eta[i]));
   b[i]~dnorm(0,b.tol);
}
```

```
for(j in 1:4){
    beta[j]~dnorm(0,1.E-6);
}
b.tol~dgamma(1.E-6,1.E-6);
plate.var<-1./b.tol;
meseed<-.5*(beta[1]+beta[2]-beta[3]-beta[4]);
meroot<-.5*(beta[1]-beta[2]+beta[3]-beta[4]);
int<-(beta[1]-beta[2]-beta[3]+beta[4]);
}
```

The following results were obtained after discarding the first 999 samples from a chain of length of 10,000.

Plate Variance:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
plate.var	0.04218	0.08115	0.004749	1.38E-6	0.005883	0.2808	1000	10001

Fixed Effects

Main Effect for seed type

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
meseed	0.4751	0.2207	0.004237	0.06438	0.4682	0.9281	1000	10001

Main effect for root extract

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
meroot	-1.073	0.217	0.002897	-1.505	-1.067	-0.6584	1000	10001

Interaction

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
int	-0.9415	0.4368	0.006994	-1.807	-0.9366	-0.08989	1000	10001

Treatment

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta[1]	-0.7098	0.1759	0.00244	-1.067	-0.7091	-0.3641	1000	10001
beta[2]	0.8337	0.177	0.004013	0.503	0.8265	1.212	1000	10001
beta[3]	-0.7142	0.2655	0.003338	-1.25	-0.7077	-0.21	1000	10001
beta[4]	-0.1122	0.2427	0.003828	-0.612	-0.1066	0.3493	1000	10001

Comparison

Parameter	PQL estimate	PQL SE	Post. Mean	Post. S.D.
σ_D^2	.059	.072	.042	.081
O aeg 75*Bean	$-.695$	.190	$-.710$	.176
O aeg 75*Cucumber	$.848$	.190	$.834$	.177
O aeg 73*Bean	$-.697$	.277	$-.714$	.266
O aeg 73*Cucumber	$-.132$	.252	$-.112$	.243