Outline

Gibbs Sampling with two Parameters

Suppose we have a two dimensional parameter space

$$\theta = (\theta_1, \theta_2).$$

An important example is

$$Y_i \sim N(\mu, \sigma^2), \quad \theta = (\mu, \sigma).$$

We want to "compute" $\pi(\theta_1, \theta_2)$

π might be the prior or the posterior (usually the posterior).

We define a Markov chain with stationary distribution π by:

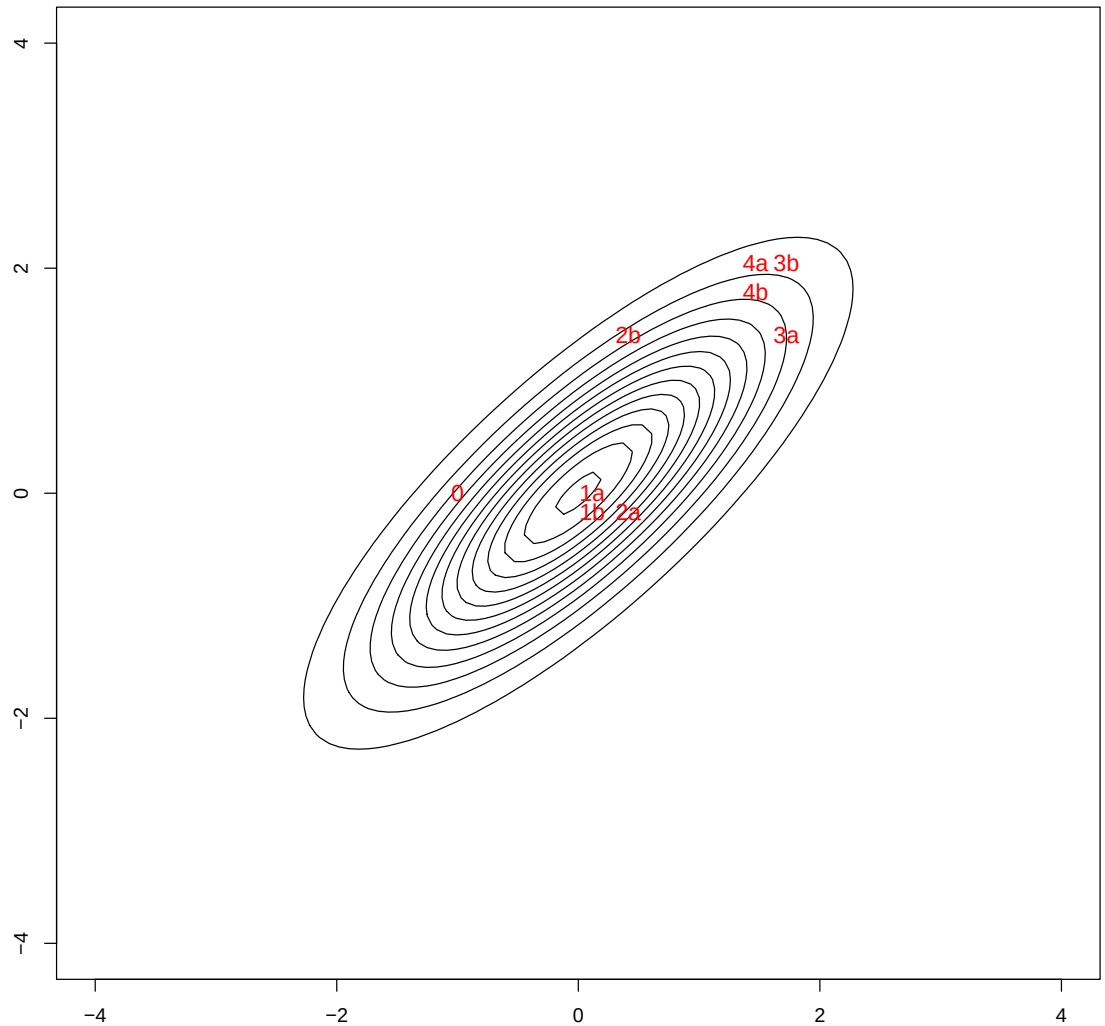
Given current values (θ_1^0, θ_2^0), draw the next pair, (θ_1^1, θ_2^1) using:

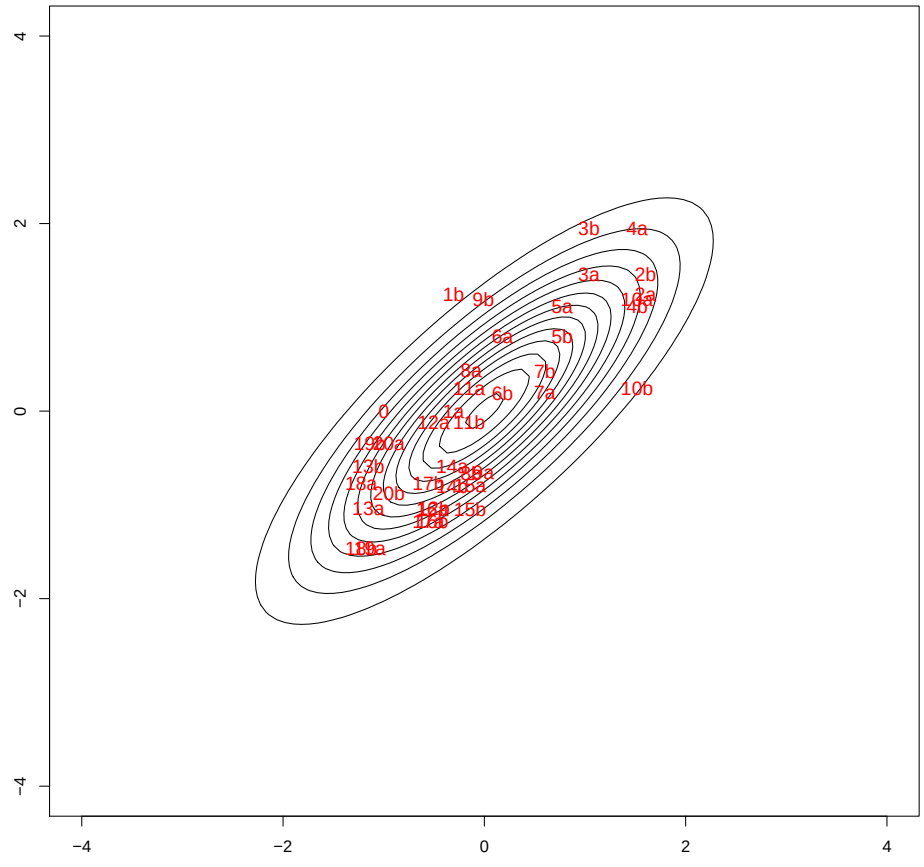
1. draw $\theta_1^1 \sim \theta_1 \mid \theta_2 = \theta_2^0$.
2. draw $\theta_2^1 \sim \theta_2 \mid \theta_2 = \theta_1^1$.

where $\theta_1 \mid \theta_2$ is the conditional for θ_1 given θ_2 under the joint distribution corresponding to π.

Same for $\theta_2 \mid \theta_1$.

i^{th} draw is labelled (ia, ib)
where ia is the first update $(\theta_1 \mid \theta_2)$
and ib is the second update $(\theta_2 \mid \theta_1)$.




4
2
0
−2
−4
−4
−2
0
2
4
3b
4a
3a
2b
1b
9b
5a
6a
5b
8a
7b
11a
6b
7a
10b
0
14
12a
11b
13b
13a
15b

Note:

1.
We clearly have a Markov Chain.

2.
If (θ_1^0, θ_2^0) is a draw from π then θ_2^0 is a draw from the marginal of θ_2 under π. Then (θ_1^1, θ_2^0) is a draw from the joint π, and so on ...

The stationary distribution of the Markov chain is π !!!!

Example: Bivariate Normal

The previous graphs were generating using:

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$$

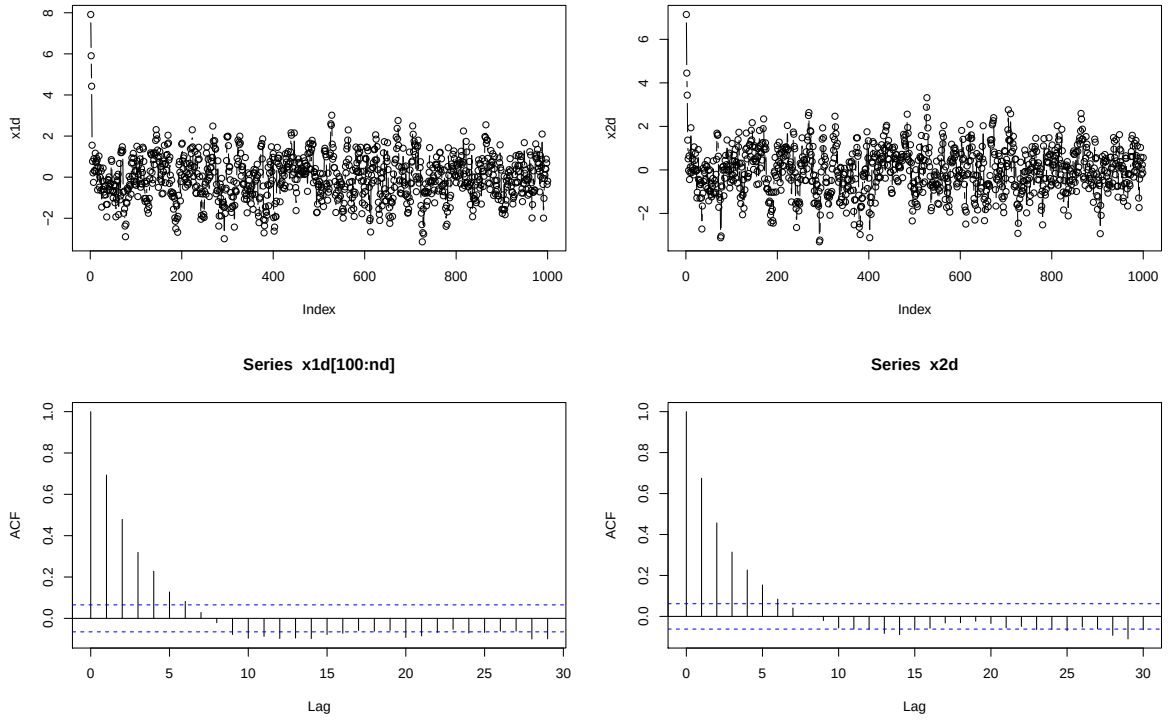
$$\theta_1 \mid \theta_2 \sim N(\rho\,\theta_2, (1-\rho^2)).$$

$$\theta_2 \mid \theta_1 \sim N(\rho\,\theta_1, (1-\rho^2)).$$

Here is R code to do the Gibbs sampler:

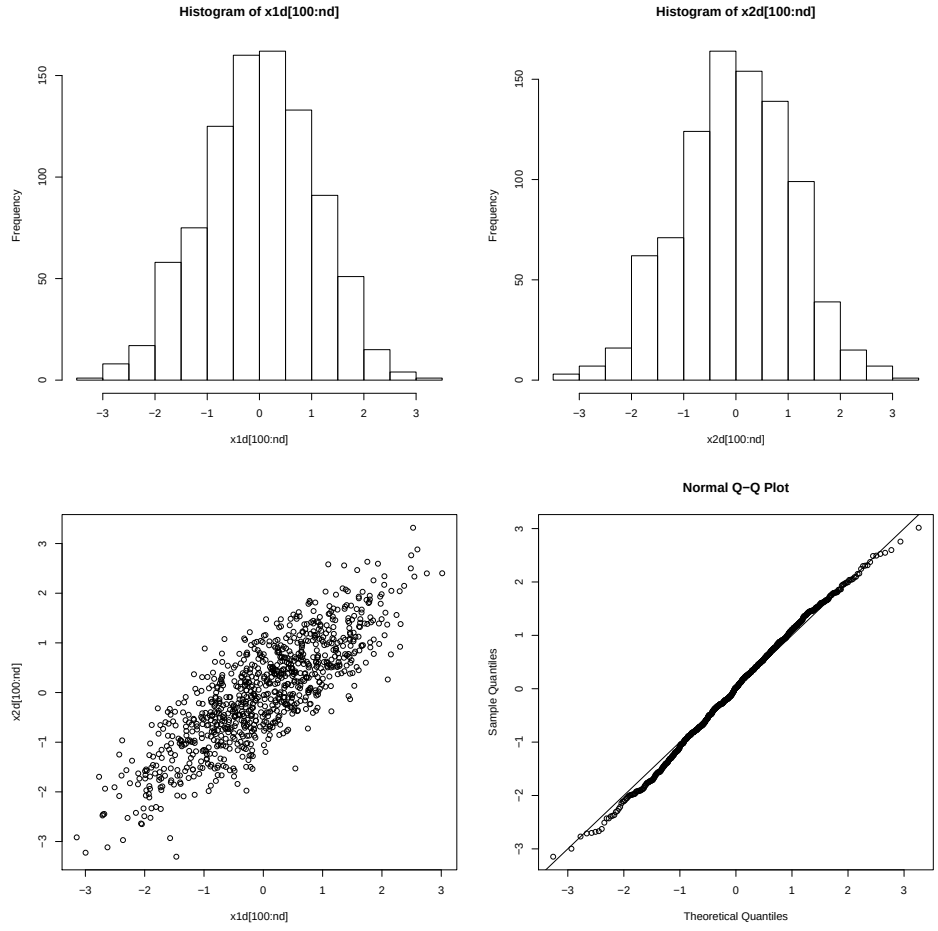
```
nd = 500
x1d=rep(0,nd)
x2d=rep(0,nd)
x1=10
x2=10
for(i in 1:nd) {
x1 = rnorm(1,rho*x2,sqrt(1-rho^2)) # x1 | x2
x2 = rnorm(1,rho*x1,sqrt(1-rho^2)) # x2 | x1
x1d[i] = x1; x2d[i]=x2
}
```

(1,1): marginal draws of x1, (1,2): marginal draws of x2
(2,1): acf of x1 draws, (2,2): acf of x2 draws



Note: for the ACF's I dropped the first 100 draws.
We often need to drop the initial draws, during which the Markov Chain "burns in" and "forgets" the starting values which may, or may not, be good.

(1,1): marginal draws of x1, (1,2): marginal draws of x2
(2,1): x1 vs. x2, (2,2): normal qqplot of x1



Again, first 100 draws dropped.

Normal Mean and Variance

Observe

$$Y_i \sim N(\mu, \sigma^2), \quad iid$$

prior:

$$\mu \sim N(\bar{\mu}, \tau^2), \quad \sigma^2 \sim \frac{\nu\lambda}{\chi^2_\nu}.$$

with

$$p(\mu, \sigma) = p(\mu)\, p(\sigma), \text{ they are independent!!.}$$

Gibbs Sampler:

Pick starting values for μ and σ and then draw:

- $\mu \mid \sigma, y$.
- $\sigma \mid \mu, y$.

We certainly know how to do the first draw.

Given μ we observe

$$\epsilon_i = Y_i - \mu \sim N(0, \sigma^2).$$

so we know how to do the second draw!

To implement the Gibbs sampler we can write a function for each of the draws

```
# mu|sigma ----------
drmu = function(y,sigma,mbar,tau) {
#draw mu | sigma, y~N(mu,sigma^2)
#y: data
#mu ~ N(mbar,tau^2)
n = length(y)
a= n/sigma^2
b=1/tau^2
ybar=mean(y)
mpost = (a*ybar+b*mbar)/(a+b)
spost = sqrt(1/(a+b))
return(rnorm(1,mpost,spost))
}
```

```
# sigma|mu ----------
drsigma = function(y,mu,nu,lambda) {
#draw sigma | mu, y~N(mu,sigma^2)
#y: data
#sigma^2 ~ nu*lambda/chi^2_nu
n=length(y)
S = sum((y-mu)^2)
return(sqrt((nu*lambda+S)/rchisq(1,nu+n)))
}
```

And then our Gibbs sampler is simply:

```
#Gibbs to draw from the posterior

nd = 1000 #number of gibbs iterations
muv = rep(0,nd) #storage for mu draws
sigv = rep(0,nd) #storage for sigma draws

mud=0 #current mu draw (this is the starting value)
sigd=1 #current sigma draw (this is the starting value)

for(i in 1:nd) {
#mu|sigma
mud = drmu(y,sigd,mbar,tau)
#sigma | mu
sigd = drsigma(y,mud,nu,lambda)
muv[i]=mud; sigv[i]=sigd
}
```

Normal (μ, σ), Simulated Data

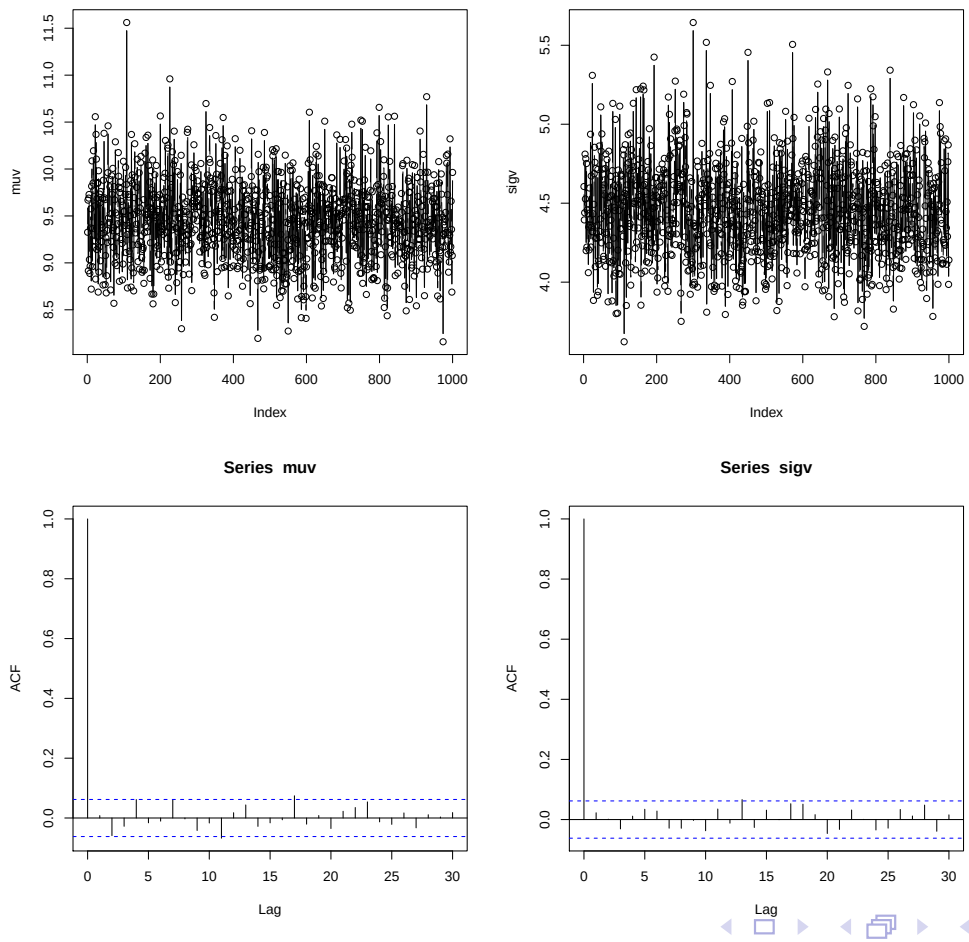
Let's try this with some simulated data.
Simulate the data and specify the prior:

```
#do simulated example
mu = 10 #true mu
sigma=5 #true sigma
n=100   # number of observations
set.seed(99)

#simulate data
y = mu + sigma*rnorm(n)

#specify prior
mbar=0
tau=10
nu=5
lambda = 3^2
```

```
#plot mcmc
par(mfrow=c(2,2))
plot(muv,type='b')
plot(sigv,type='b')
acf(muv)
acf(sigv)
dev.copy2pdf(file='mu-sig-ts.pdf',height=10,width=10)
```

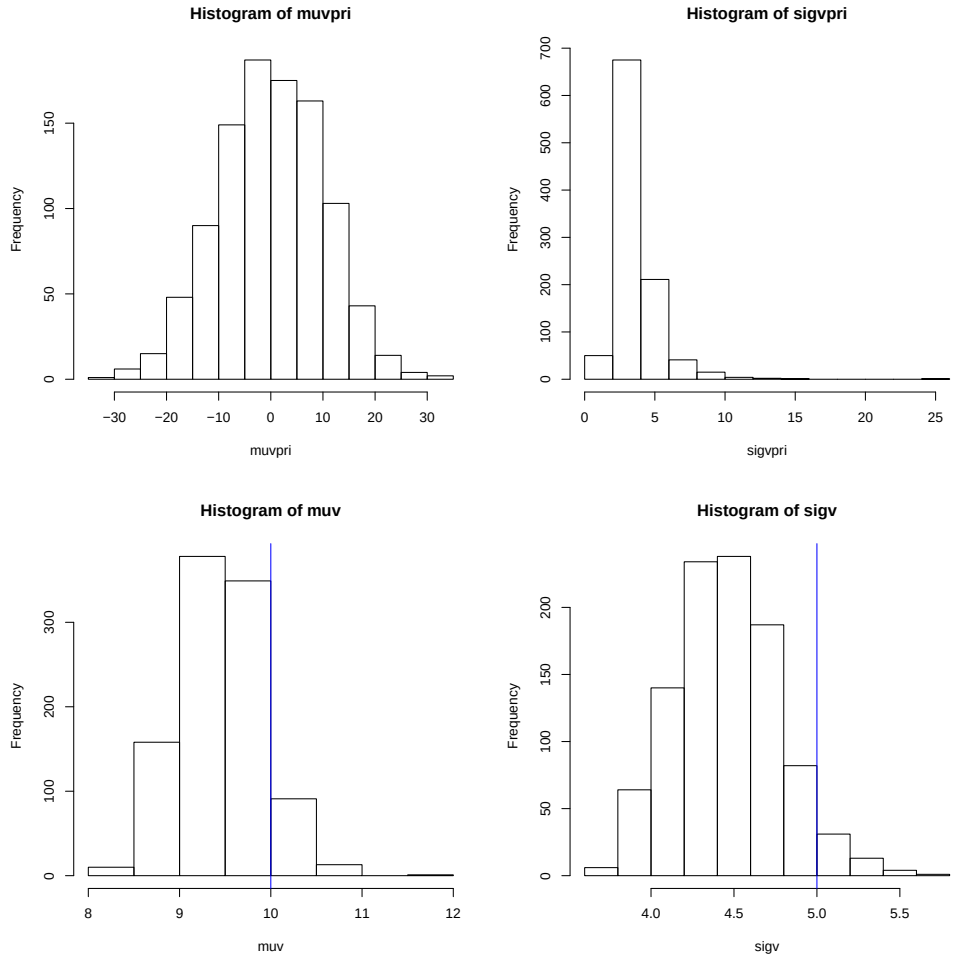
Wow!

no burn-in!!



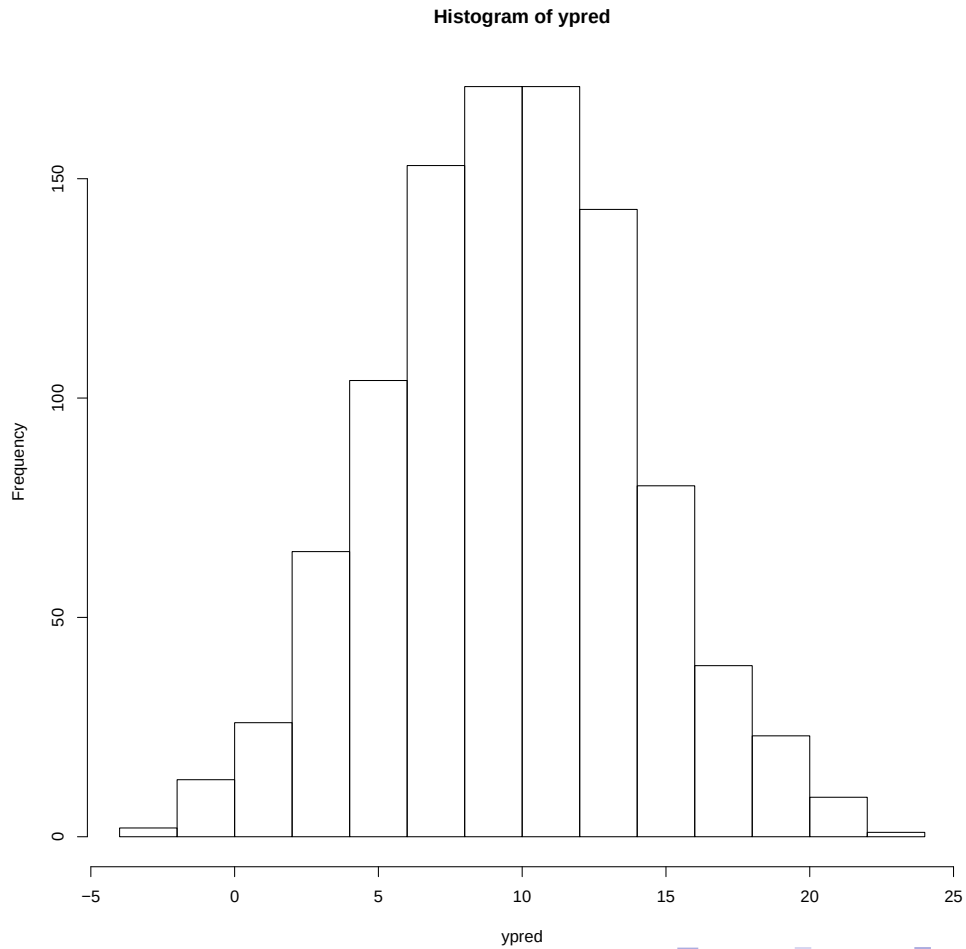
```
#plot inference: prior and posterior
par(mfrow=c(2,2))
#prior draws
muvpri = mbar + tau*rnorm(nd)
sigvpri = sqrt((nu*lambda)/rchisq(nd,nu))
hist(muvpri)
hist(sigvpri)
#posterior draws
hist(muv)
abline(v=mu,col='blue')
hist(sigv)
abline(v=sigma,col='blue')
```

(1,1): prior on μ, (1,2): prior on σ.
(1,1): posterior on μ, (1,2): posterior on σ.



```
#prediction
ypred = rep(0,nd)
for(i in 1:nd) {
ypred[i] = rnorm(1,muv[i],sigv[i])
}
par(mfrow=c(1,1))
hist(ypred)
```

For each draw of
$\theta = (\mu, \sigma)$
from the posterior,
we draw
$Y \mid \theta \sim N(\mu, \sigma^2)$.



The General Gibbs Sampler

We can extend the idea of the Gibbs sampler to any number of parameters!
Let

$$\theta = (\theta_1, \theta_2, \ldots, \theta_k).$$

Here each θ_j can be a subset or "block" of parameters of any size.
We then iterate by sequentially drawing from all the conditionals:

$$\theta_j \mid \theta_1, \theta_2, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k.$$

That is:

- update each θ_j by drawing from its conditional given the rest.
- condition using the most recent draws of "the rest".

If we start with

$$\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_k^0).$$

then we get to

$$\theta^1 = (\theta_1^1, \theta_2^1, \ldots, \theta_k^1).$$

by drawing

$$\theta_j^1 \sim \theta_j \mid \theta_1 = \theta_1^1, \theta_2 = \theta_2^1, \ldots, \theta_{j-1} = \theta_{j-1}^1, \theta_{j+1} = \theta_{j+1}^0, \ldots, \theta_k = \theta_k^0$$

$$j = 1, 2, \ldots, k.$$

Hierarchical Normal Means

Observe:

$$Y_{ij} \sim N(\theta_j, \sigma_j^2), \;\; j = 1, 2, \ldots, m, i = 1, 2, \ldots, n_j.$$

We have m groups of observations.

Within each group, we observe iid normal data with a mean θ_j and standard deviation σ_j that depend on the group.

Let's ignore the σ_j for a while and focus on our choice of prior (or *model*) for the θ_j.

Suppose the groups have something to do with each other in that they are the same kind of thing.

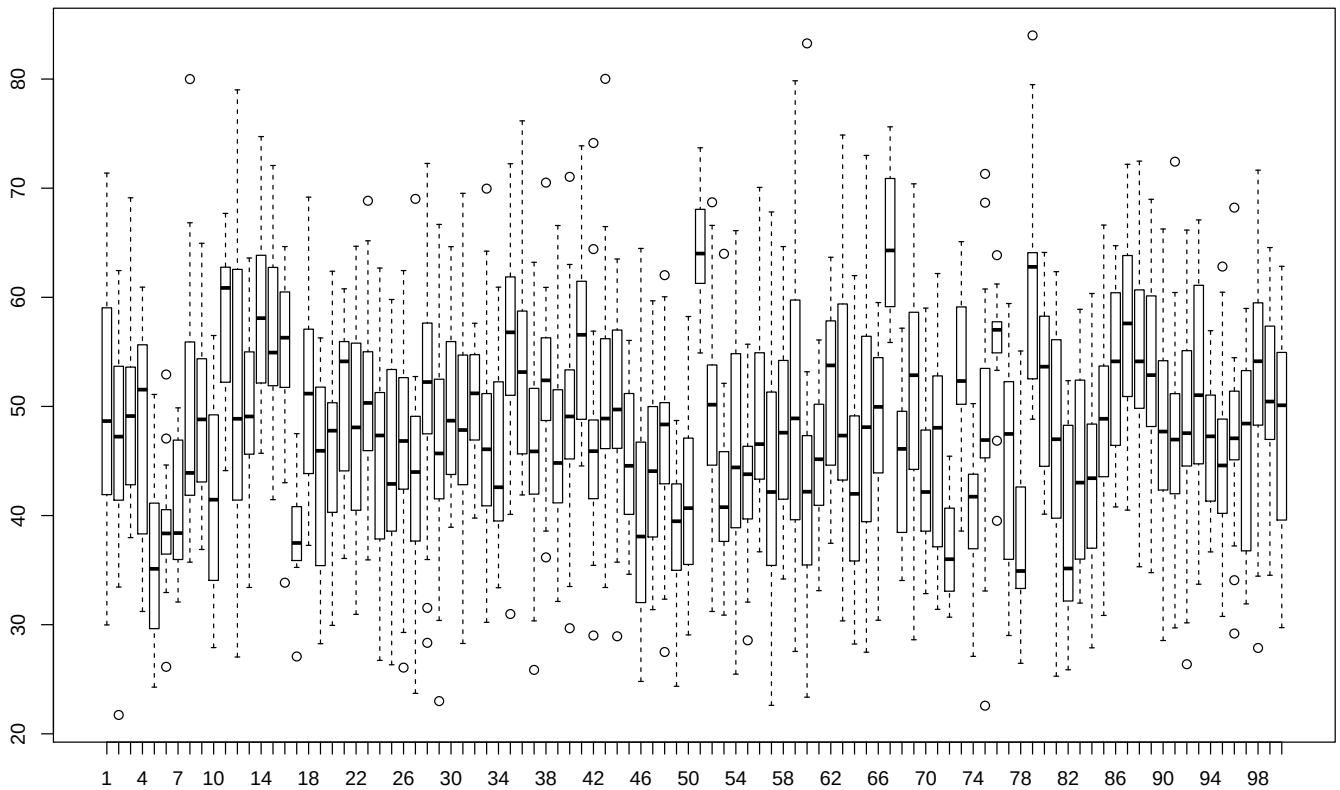
It may help us to think about our prior for the θ_j jointly.

Hoff's example is:

- Each group corresponds to a school.
- Within each group, each $Y_{i,j}$ is a students score on a math test.

$Y_{i,j}$: math score for student i, at school j.

The data, each boxplots displays math test scores for a particular school.



Rather than think about each school separately, we think about an overall level for the schools and then how much this overall level varies across schools:

$$\theta_j \sim N(\mu_\theta, \sigma_\theta^2).$$

μ_θ: overall mean level across schools.
σ_θ: variation in individual school mean level.

For example, if σ_θ were 0, that would be like combining the groups together.

If we stop here, we are just using the same prior for each θ_j.

So, for example,

$$E(\theta_j \mid y_j, \sigma_j) = \frac{a_j\, \bar{y}_j + b\, \mu_\theta}{a_j + b},$$

$$a_j = \frac{n_j}{\sigma_j^2}, \quad b = \frac{1}{\sigma_\theta^2}.$$

We shrink each θ_j towards the common "grand mean" μ_θ and the amount of shrinkage is crucially controlled by σ_θ.

But, our inference for θ_j only depends on the observations in group j.

Where it get interesting is when we say to ourselves:

Well, I like the idea of thinking about the θ's together, but I am really not too sure about what the overall mean μ_θ should be and how much the θ_j vary about μ_θ, that is, σ_θ.

Maybe if I have a lot of observations in most of the groups telling me that their θ's are big that should suggest bigger θ's in the rest of the groups.

$$\theta_j \sim N(\mu_\theta, \sigma_\theta^2).$$

We can put priors on μ_θ and σ_θ:

$$\mu_\theta \sim N(\bar{\mu}, \sigma_\mu^2), \;\; \sigma_\theta^2 \sim \frac{\nu\,\lambda}{\chi_\nu^2}, \;\; \text{independent.}$$

Now, all of the data will help us learn about $(\mu_\theta, \sigma_\theta)$.

Learning about μ determines where we shrink to.

Learning about σ_θ determines how much we shrink.

By thinking about the groups together, we have *adapative shrinkage*.

$$\theta_j \sim N(\mu_\theta, \sigma_\theta^2).$$

$$\mu_\theta \sim N(\bar{\mu}, \sigma_\mu^2), \quad \sigma_\theta^2 \sim \frac{\nu\,\lambda}{\chi_\nu^2}, \quad \text{independent.}$$

Note that if we observed the θ_j, this would just be our standard normal inference problem for a normal mean and standard deviation $(\mu_\theta, \sigma_\theta)$.

We could also think about the σ_j hierarchically and perhaps we will do this later.

For now let's just use

$$\sigma_j^2 \sim \frac{\nu_1 \lambda_1}{\chi^2_{\nu_1}}, \quad \textit{iid}.$$

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_m)$.

Let $y_j = (y_{1j}, y_{2j}, \ldots, y_{n_j j})$ and $y = (y_1, y_2, \ldots, y_m)$.

Then our full joint can be written,

$$p(\mu_\theta, \sigma_\theta, \theta, \sigma, y) =$$

$$p(\mu_\theta)\, p(\sigma_\theta)\, p(\theta \mid \mu_\theta, \sigma_\theta)\, p(\sigma)\, p(y \mid \theta, \sigma).$$

$$p(\mu_\theta, \sigma_\theta, \theta, \sigma, y) =$$

$$p(\mu_\theta)\, p(\sigma_\theta)\, p(\theta \mid \mu_\theta, \sigma_\theta)\, p(\sigma)\, p(y \mid \theta, \sigma).$$

where,

$p(\mu_\theta) : N(\bar{\mu}, \sigma_\mu^2)$
$p(\sigma_\theta) : \sigma_\theta^2 \sim \frac{\nu\,\lambda}{\chi_\nu^2}$
$p(\theta \mid \mu_\theta, \sigma_\theta) = \Pi_{j=1}^m\, p(\theta_j \mid \mu_\theta, \sigma_\theta^2),\ \theta_i \mid \mu_\theta, \sigma_\theta^2 \sim N(\mu_\theta, \sigma_\theta^2).$
$p(\sigma) = \Pi_{j=1}^m\, p(\sigma_j),\ \sigma_j^2 \sim \frac{\nu_1\,\lambda_1}{\chi_{\nu_1}^2}$
$p(y \mid \theta, \sigma) = \Pi_{j=1}^m\, p(y_j \mid \theta_j, \sigma_j)$
$p(y_j \mid \theta_j, \sigma_j) = \Pi_{i=1}^{n_j}\, p(y_{ij} \mid \theta_j, \sigma_j),\ y_{ij} \sim N(\theta_j, \sigma_j^2)$

Have to pick

$$(\bar{\mu}, \sigma_\mu, \nu, \lambda, \nu_1, \lambda_1).$$

Gibbs Sampler:

$$\mu_\theta \mid \sigma_\theta, \theta, \sigma, y$$

$$\sigma_\theta \mid \mu_\theta, \theta, \sigma, y$$

$$\theta \mid \mu_\theta, \sigma_\theta, \sigma, y$$

$$\sigma \mid \mu_\theta, \sigma_\theta, \theta, y$$

where,

$$\mu_\theta \mid \sigma_\theta, \theta, \sigma, y = \mu_\theta \mid \sigma_\theta, \theta \text{ (normal mean)}$$
$$\sigma_\theta \mid \mu_\theta, \theta, \sigma, y = \sigma_\theta \mid \mu_\theta, \theta \text{ (normal standard deviation)}$$
$$\theta \mid \mu_\theta, \sigma_\theta, \sigma, y = \theta \mid \mu_\theta, \sigma_\theta, \sigma, y$$
(m normal means, 1 for each θ_j)
$$\sigma \mid \mu_\theta, \sigma_\theta, \theta, y = \sigma \mid \theta, y$$
(m normal standard deviations, 1 for each σ_j)

Hierarchical Means Gibbs Sampler

```
drHierM = function(y,theta,sigma,mtheta,stheta,mm,sm,nu,lam,nu1,lam1) {
#y: list, y[[j]], observations for group j
#theta,sigma: y_ij ~ N(theta[j],sigma[j]^2)
#mtheta,stheta: theta_j ~ N(mtheta,stheta^2)
#mm,sm: mtheta ~ N(mm,sm^2)
#nu,lam: stheta^2 ~ nu*lam/chi^2_nu
#nu1,lam1: sigma_j ~ nu1*lam1/chi^2_nu1
   m=length(y)
   #draw theta----------
   for(j in 1:m) {
      theta[j] = drmu(y[[j]],sigma[j],mtheta,stheta)
   }
   #draw sigma----------
   for(j in 1:m) {
      sigma[j] = drsigma(y[[j]],theta[j],nu1,lam1)
   }
   #draw mtheta----------
   mtheta = drmu(theta,stheta,mm,sm)
   #draw stheta----------
   stheta = drsigma(theta,mtheta,nu,lam)

   return(list(theta=theta,sigma=sigma,mtheta=mtheta,stheta=stheta))
}
```

(1,1): post mean of θ_j vs. $\bar{y}_j$.
(1,2): post mean of σ_j vs. sd(y_j).
(2,1): n_j vs. absolute value of difference, θ_j - $\bar{y}_j$.