The normal distribution

Patrick Breheny

January 24

Introduction

- Typically, once a model has more than one parameter, it is not possible to find conjugate priors anymore – obviously, this rules out virtually all interesting analyses
- In particular, even the normal distribution with its two parameters, μ and σ^2, cannot be analyzed with conjugate methods
- Conjugate approaches do exist, however, for each parameter individually, if we were to act as if the other parameter was known
- This lecture explores those approaches, both because it lends insight into normal models and because it illustrates the basic idea of Gibbs sampling

Conjugate prior for the normal mean

- First, let's suppose that the variance σ^2 is known
- **Exercise**: For $Y_i \overset{\text{iid}}{\sim} \text{N}(\theta, \sigma^2)$ with σ^2 known, the conjugate prior for θ is also normal
- Note that this requires "completing the square":

$$x^2 + bx + c = \left(x + \frac{1}{2}b\right)^2 + k,$$

where the two expressions differ only with respect to the constant term

Posterior distribution

Thus, if the prior distribution on the mean is

$$\theta \sim N\left(\mu_0, \frac{\sigma^2}{n_0}\right),$$

the posterior distribution is

$$\theta|\mathbf{y} \sim N\left(\mu_n, \frac{\sigma^2}{n_0+n}\right),$$

where

$$\mu_n = \frac{n_0\mu_0 + n\bar{y}}{n_0+n}$$

Shrinkage

- Note that the posterior mean can also be written as

$$\mu_n = w\mu_0 + (1-w)\bar{y}$$

where

$$w = \frac{n_0}{n_0 + n}$$

- Thus, as we have seen with other distributions, the posterior mean is a weighted average of the prior mean and sample mean, with the relative weights determined by the sample size and prior variance (which is in turn determined here by n_0, the "effective prior sample size")
- This phenomenon, where the posterior is shrunk towards the prior, is often referred to as *shrinkage*; we will see examples of shrinkage throughout this course

Precision

- Recall that the *precision* is the inverse of the variance: $\tau = 1/\sigma^2$
- Again, it is important to distinguish between the precision with which we know the mean (let's call this ω) and the precision that reflects the fundamental variability of the outcome (let's call this τ), so that our model is:

$$Y_i \sim \mathrm{N}(\theta, \tau^{-1})$$
$$\theta \sim \mathrm{N}(\mu_0, \omega_0^{-1})$$

Precision (cont'd)

- Now, our posterior for θ is

$$\theta|y \sim \mathrm{N}(\mu_n, \omega_n^{-1})$$

where

$$\mu_n = \frac{\omega_0\mu_0 + n\tau\bar{y}}{\omega_0 + n\tau}$$
$$\omega_n = \omega_0 + n\tau;$$

in other words, the posterior precision for the mean is the sum of the prior precision and the information (recall that $n\tau$ is the Fisher information)

- As noted previously, BUGS and JAGS parameterize the normal distribution in terms of the precision, as it is typically easier to work with in Bayesian calculations

Predictive distribution

- The posterior predictive distribution for the normal mean case is particularly easy to think about, as it is equivalent to the sum of two independent normal quantities: $\epsilon \sim \mathrm{N}(0, \sigma^2)$ and $\theta|y \sim \mathrm{N}(\mu_n, \sigma_\theta^2)$
- Thus,

$$Y|y \sim \mathrm{N}(\mu_n, \sigma^2 + \sigma_\theta^2)$$

- This is the same approach used in frequentist "prediction intervals"; one of the rare cases where it is possible in frequentist statistics to take parameter uncertainty into account when carrying out prediction

Reference priors

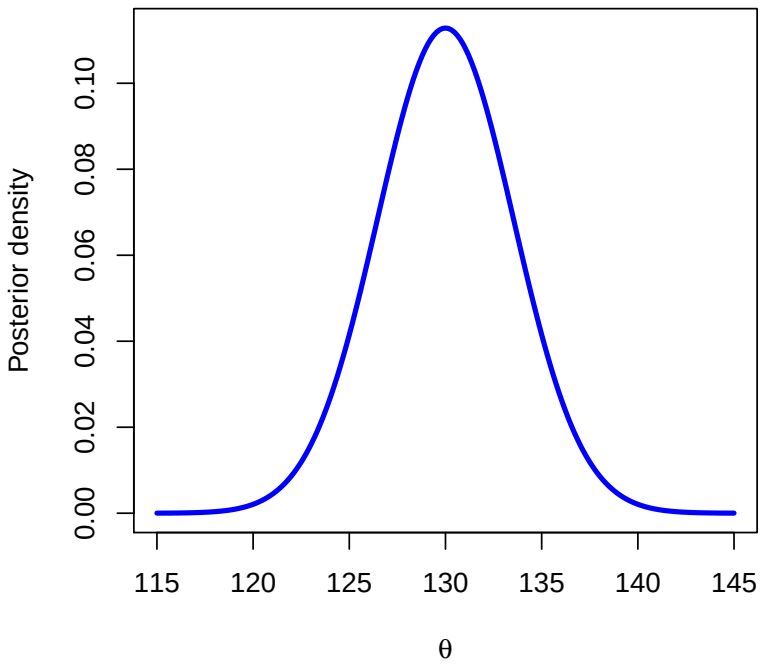
- For the normal distribution, $I(\theta) = \frac{1}{\sigma^2}$
- This is constant with respect to theta; thus, the Jeffreys approach would suggest a uniform (or "flat") distribution over the entire real line
- Obviously, this is improper
- BUGS does provide a `dflat` distribution; JAGS does not, although the same basic effect may be realized by either taking `theta` $\sim$ `dunif(-100, 100)` (or some other very wide range) or `theta` $\sim$ `dnorm(0, 0.0001)` (or some other very small precision)

Example: THM in tap water

- Water companies and the EPA regularly monitor the concentrations of various chemicals that are present in tap water
- This example deals with measurements of trihalomethanes (THMs), which are thought to be carcinogenic in humans at high concentrations
- Suppose that an assay has a known measurement error of $\sigma = 5\,\mu$g/L, and that, for a certain water supply, the measurements are 128 μg/L and 132 μg/L

Uninformative prior

Suppose we used an uninformative prior:



$$\hat{\theta} = \bar{\theta} = 130$$

$$HDI_{95} = (123.1, 136.9)$$

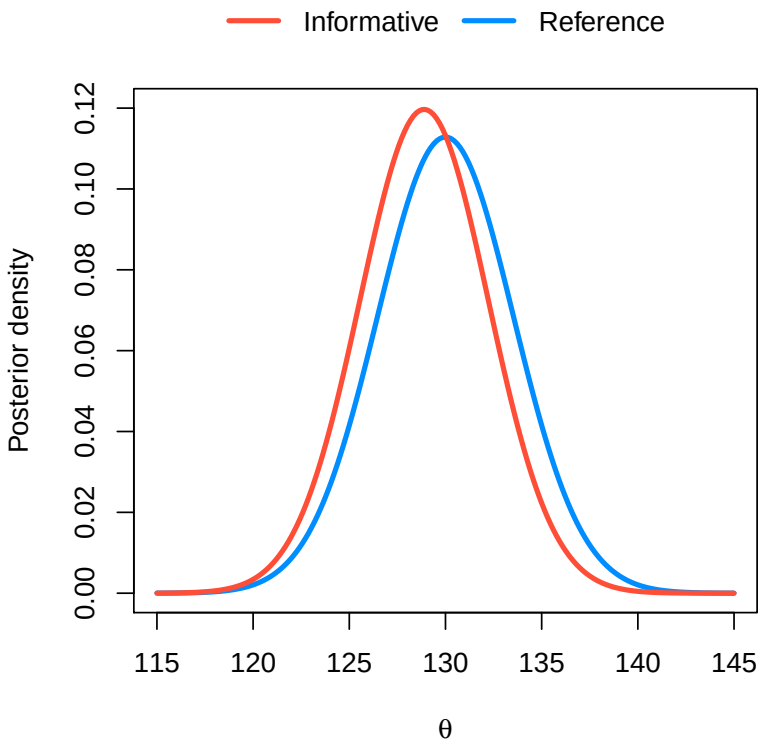
Note that in this case, the posterior mean is identical to the posterior mode, the HPD interval is identical to the central interval, and all of these results are identical to that of a standard frequentist analysis

Informative prior

- Suppose, however, that historical data on THM levels from other water supplies have a mean of 120 μg/L and a standard deviation of 10 μg/L
- This suggests a prior on θ, the THM concentration of the water supply under investigation, of $\theta \sim \mathrm{N}(120, 10^2)$
- With the parameterization we have been using, $n_0 = 5^2/10^2 = 1/4$; *i.e.*, the prior counts for about 1/4th of an observation

Informative prior (cont'd)

With the informative prior:



$$\hat{\theta} = \bar{\theta} = 128.9$$
$$HDI_{95} = (122.4, 135.4)$$

Note that the informative prior pulls the posterior to the left as well as makes it narrower

Conjugate prior #1

- Let us now reverse the situation, and suppose that we know the mean μ of the normal distribution, but that the variance θ is unknown
- Rather than work with the variance, however, we will find it easier to work with the precision τ
- **Exercise**: For $Y_i \overset{\text{iid}}{\sim} \text{N}(\mu, \tau^{-1})$ with μ known, the conjugate prior for τ is Gamma
- **Exercise**: For $\tau \sim \text{Gamma}(\alpha, \beta)$,

$$\tau|\mathbf{y} \sim \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\text{RSS}\right),$$

where $\text{RSS} = \sum_i (y_i - \mu)^2$

Conjugate prior #2

- An equivalent way of expressing the conjugate prior is as a *scaled* χ^2 distribution: if $cX \sim \chi^2(\nu)$ for a positive constant c, then $X \sim \text{Gamma}(\nu/2, c/2)$
- I will use the expression $\text{Scaled-}\chi^2(\nu, c)$ to denote this distribution (which, of course, is a Gamma distribution, just with an alternate parameterization)
- What does this mean for specifying a prior?
- If we let $\tau \sim \text{Scaled-}\chi^2(n_0, \text{RSS}_0)$, then

$$\tau|\mathbf{y} \sim \text{Scaled-}\chi^2(n_0 + n, \text{RSS}_0 + \text{RSS})$$

Conjugate priors for σ^2?

- Are there conjugate priors for σ^2? Sort of
- Yes, in the sense that if $X \sim \text{Gamma}$, then $1/X$ is said to follow an "inverse-gamma" distribution; similarly there is an "scaled inverse χ^2" distribution
- These distributions do have closed forms for their distribution functions, mean, variance, mode, etc., but are not familiar to most people and cannot be directly specified in BUGS/JAGS

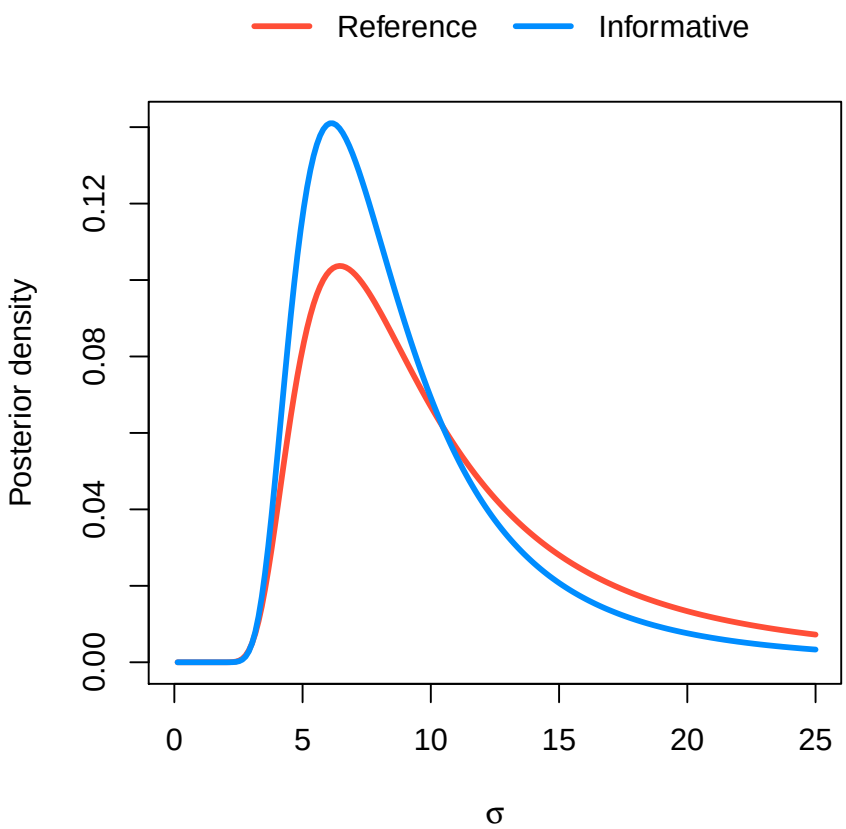
Reference priors

- With a Gaussian distribution, the Fisher information for the standard deviation is $I(\sigma) = 2/\sigma^2$; the Jeffreys prior is therefore $p(\sigma) \propto \sigma^{-1}$
- The Jeffreys prior for the precision is also $p(\tau) \propto \tau^{-1}$
- These are improper, but can be approximated in BUGS/JAGS with `dgamma(0.0001, 0.0001)`
- Alternatively, the Jeffreys prior on $\log(\sigma)$ is $p(\log \sigma) \propto 1$
- This is obviously improper, but can be approximated with `log.sigma ~ dunif(-10,10)`

THM example

- Let's continue with our THM example, but now suppose that we are assessing the variability of the instrument by taking a few measurements on a sample with known concentration of 100 μg/L
- Suppose that we obtained two measurements: 105 μg/L and 110 μg/L (*i.e.*, measurement errors of 5 and 10 μg/L)
- We'll analyze this data with both a reference prior and an informative prior
- For the latter, suppose we have a vague notion that the standard deviation is about 5, so we use a scaled χ^2 distribution with $\text{RSS}_0 = 25$, $n_0 = 1$

Results



	Reference	Informative
$\hat{\sigma}$	6.5	6.1
$\bar{\sigma}$	14.0	9.8
$\mathrm{SD}(\sigma)$	24.4	7.3
CI_{95}	(4.1,49.7)	(4.0,26.4)
HDI_{95}	(2.8,34.7)	(3.1,20.9)

Multiple unknown parameters

- Typically, of course, neither the mean nor the variance of the distribution are known; what then?
- We have previously considered some models with multiple parameters; for example the two-sample Poisson childbirth data
- In that situation, the priors for λ_1 and λ_2 (the childbirth rates for the two education levels) were independent, as were the data from the two groups
- Thus, we could obtain the posterior distributions for λ_1 and λ_2 separately, and use conjugate results for each parameter

Semi-conjugacy

- This is not the case for the normal distribution: the posterior distributions for μ and σ^2 are dependent on each other and there is no standard parametric distribution which is conjugate for their joint likelihood
- However, as we have seen, μ and σ^2 (or τ) have conjugate distributions if we condition on knowing the other parameter(s) in the model
- This phenomenon is known as *semi-conjugacy*; although it does not lead to closed-form solutions for the entire posterior, it will help us to sample from it

A sampler

Our sampling approach will be as follows:

- Set some initial value for τ: τ_0
- For $b = 1, 2, \ldots, B$,
  - Draw $\mu_b \sim \mathrm{N}(\mu_n, \omega_n^{-1}) | \tau = \tau_{b-1}$
  - Draw $\tau_b \sim \text{Scaled-}\chi^2(n + n_0, \mathrm{RSS} + \mathrm{RSS}_0) | \mu = \mu_b$

Alternatively, we could have drawn τ first, then μ; the order is not important

Gibbs sampling

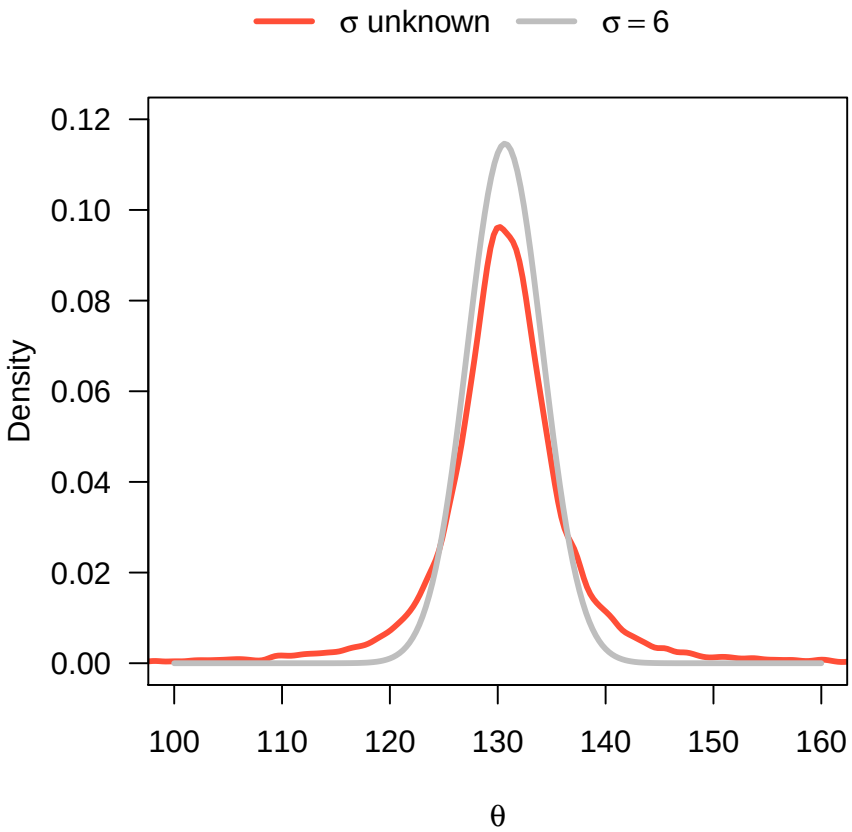
- The algorithm on the previous slide is an example of a *Gibbs sampler*
- We'll discuss Gibbs sampling in more detail later on, but the example illustrates the basic idea: we sample each parameter individually, conditioning on the *most recent* values of the other parameters
- This idea is not without its flaws (we will see one of them in a moment), but its huge virtue is that it allows us to sample from extremely complicated and high-dimensional posterior distributions by breaking the problem down into sampling one parameter at a time

Gibbs sampler: R code

```
gibbs <- function(x, mu, omega, n0, RSS0, B=10000) {
  n <- length(x)
  ybar <- mean(x)
  tau.init <- 1/var(x)
  theta <- matrix(NA, nrow=B, ncol=2)
  for (b in 1:B) {
    inf <- if (b==1) n*tau.init else n*theta[b-1,2]
    Mu <- (omega*mu + inf*ybar)/(omega+inf)
    Omega <- omega+inf
    theta[b,1] <- rnorm(1, Mu, sd=1/sqrt(Omega))
    RSS <- sum((x-theta[b,1])^2)
    theta[b,2] <- rgamma(1, (n0+n)/2, (RSS+RSS0)/2)
  }
  theta
}
```

Results: Mean

Below is the *marginal* distribution of θ, the expected value of Y:



	σ unknown	$\sigma = 6$
$\bar{\theta}$	130.7	130.7
$\mathrm{SD}(\theta)$	10.7	3.5
CI_{95}	(114.9,145.7)	(123.8,137.5)

Averaging over unknown parameters

- Note that we obtain the marginal distribution for θ by integrating σ out of the joint distribution; with Monte Carlo integration, this integral is approximated by the simpler process of *averaging over the unknown parameters*
- Not surprisingly, this causes the posterior distribution to become more spread out, but note that its very shape changes: $\theta|\mathbf{y}$ no longer follows a normal distribution; its tails are much thicker

Connection with Student's t

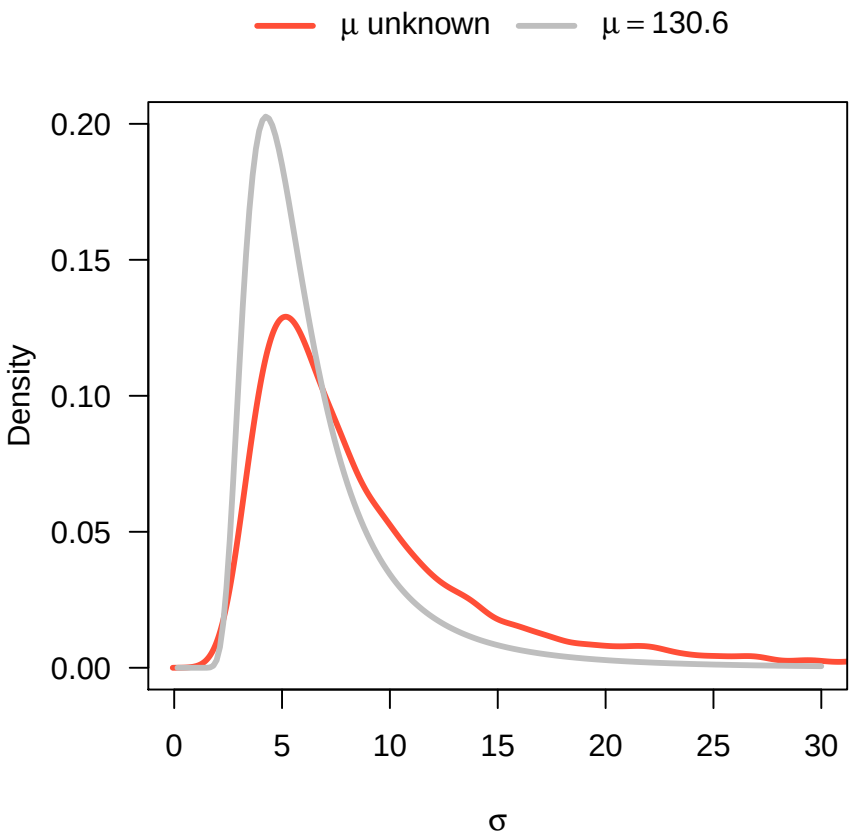
- In fact, it can be shown that if

$$X|\tau \sim \mathrm{N}(\mu, \tau^{-1})$$
$$\tau \sim \mathrm{Gamma}(\alpha, \beta),$$

  then X follows a t distribution
- Thus, in this particular case (provided we use reference priors), we again obtain standard frequentist results
- However, it is worth noting that this phenomenon arises naturally in Bayesian statistics, and will continue to arise in ever more complicated models; this is typically not the case in frequentist statistics

Results: σ

A similar phenomenon occurs with the posterior distribution of σ:



	μ unknown	$\mu = 130.7$
$\bar{\sigma}$	10.8	6.8
$\mathrm{SD}(\sigma)$	15.6	4.8
HDI_{95}	(2.2, 27.1)	(2.2, 14.8)

Dependency in Gibbs samplers

- In this particular example, the Gibbs sampler worked beautifully; as alluded to earlier, however, this is not always the case
- It is tempting to think that with Gibbs sampling, we are obtaining independent draws from the posterior, but that is not the case: when we draw τ_b, it depends on θ_b, which in turn depends on τ_{b-1}, so consecutive draws of τ (and θ) are dependent

Problematic example

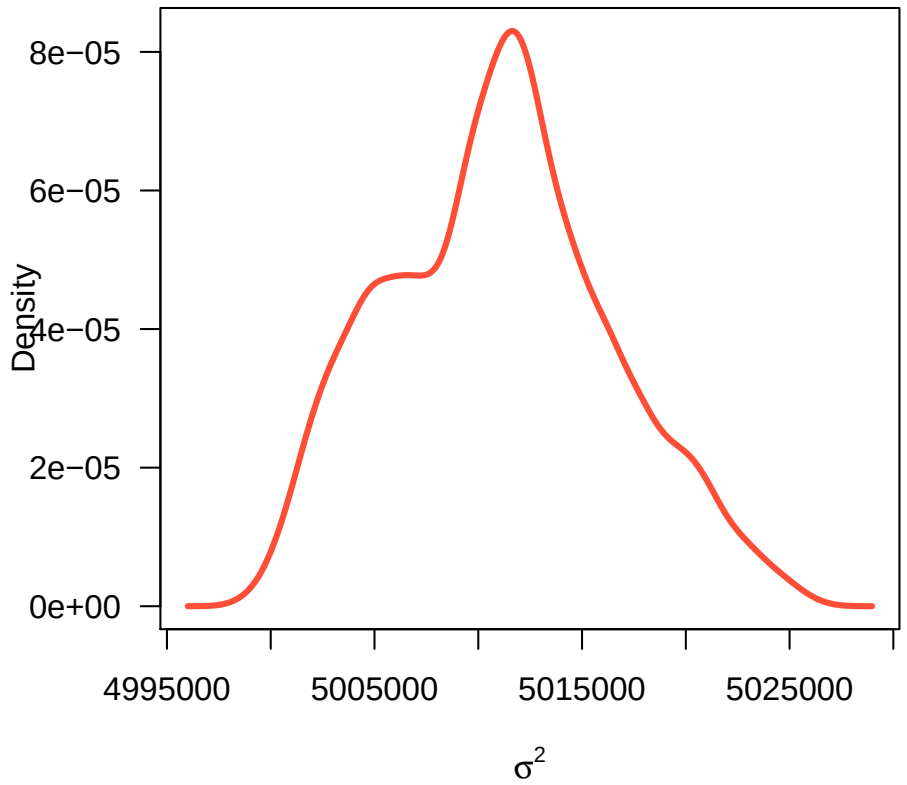
To see an example of this, suppose we fit the following model in JAGS:

```
model <- function() {
  ## Likelihood
  for (i in 1:n) {
    x[i] ~ dnorm(mu, pow(sigma.sq, -1))
  }

  ## Prior
  mu ~ dnorm(0, 0.0001)
  sigma.sq ~ dunif(0, 1e7)
}
```

Problematic results

We obtain:



The posterior distribution for σ^2 is nowhere even close to 5,000,000 in reality; what's going on?

An explanation

- What has happened is the following: if you do not supply an initial value to BUGS/JAGS, it will generate them from the prior: in this case, σ^2 is drawn from a $\text{Unif}(0, 10^7)$ distribution, which produced 5,005,129
- With a variance this large and only three observations, the mean, θ, could be almost anywhere
- This causes the Gibbs sampler to just bounce around in the extreme tails of the posterior and never find the central mass of the posterior

Final remarks

From now on, we'll be doing a lot of Gibbs sampling, but keep this example in mind as a cautionary tale:

- We may need to give thought to initial values
- "Uninformative" priors can be problematic, especially when using ones which are not invariant
- We need to be careful when carrying out Gibbs sampling to check that this sort of thing has not happened to us (we will be discussing diagnostic methods for such checking in future lectures)