COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Improper Priors

What if we want to make the priors as less informative as possible?

In the Beta-Binomial model this was easily done by choosing Beta(1, 1) prior (which is the same as U(0, 1)). But what do we do in the normal model setup?

We might want to choose $\kappa_0 = \nu_0 = 0$ as they represent the prior "sample size". However, this is technically impossible. But we can notice that the posterior mean and variance converge to the sample mean and variance if we let $\kappa_0 \to 0$ & $\nu_0 \to 0$

More formally, we can use an *improper* prior

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Notice this is not a pdf since $\iint \pi(\theta, \sigma^2) d\theta \, d\sigma^2 = \infty$. However, it can be shown that both posterior distributions are *proper* with

$$\sigma^2 | x_1, \dots x_n \sim IG\left(\frac{n-1}{2}, \frac{1}{2}\sum (x_i - \bar{x})^2\right)$$

Exercise 0

In the Binomial model consider the prior

$$\pi(\theta) \propto \frac{1}{\theta}$$

Find the posterior distribution and show that it is proper iff

$$\sum_{i=1}^{n} x_i \geq 1$$

Bias

Recall from frequentist statistics that an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is called *unbiased* if $\mathrm{E}(\hat{\theta}) = \theta$.

However, this concept is not applicable in Bayesian analysis since the parameter is a random variable.

For example, in the normal model

$$E(\theta|x_1,..,x_n) = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}\mu_0 + \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}\bar{x} = w\bar{x} + (1-w)\mu_0$$

The only way to compare this to a "true value" θ is to think that we know the data were generate from a specific θ, then the "bias" of the Bayesian estimator is $w\theta + (1-w)\mu_0$, so in a sense Bayesian estimators are always "biased".

MSE

Bias is an overrated property of estimators. What is more desirable is that the *mean squared error* is as small as possible.

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

It can be shown that

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$$

Bayesian estimators have smaller variances than their frequentist counterparts, and often they will also have smaller MSE. Some argue that if you know even just a little bit about the population you are about to sample from, you should be able to find values of the hyperparameters of the prior such that this inequality holds.

Introduction to Gibbs

- For many multiparameter Bayesian models the joint posterior distribution is non-standard and difficult to sample from directly.
- However, it is often the case that it is easy to sample from the full conditional distribution of each parameter.
- In such cases, posterior approximation can be made with the Gibbs sampler, an iterative algorithm that constructs a dependent sequence (a Markov chain) of parameter values whose distribution converges to the target joint posterior distribution.

Gibbs sampler – General Description

- The Gibbs sampler is an MCMC method to generate a draw from a target multivariate distribution $f(x_1, \ldots, x_p)$.
- It is used when draws from the joint distribution are hard to compute directly.
- Draws from the conditional distributions are easy to obtain

$$f(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$$

- Start with some initial guess values $x_1^{(0)}, \ldots, x_p^{(0)}$
- Iterate many times to obtain a Markov chain that converges to the target distribution $f(x_1, \ldots, x_p)$.

Gibbs Sampler Iteration Steps

- Given the current values $x_1^{(t)}, \ldots, x_p^{(t)}$, obtain new values:

$$x_1^{(t+1)} \sim f(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)})$$

$$x_2^{(t+1)} \sim f(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$$

$$x_3^{(t+1)} \sim f(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_p^{(t)})$$

$$\ldots$$

$$x_p^{(t+1)} \sim f(x_p \mid x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$$

- Under mild regularity conditions it can be shown that $(x_1^{(t)}, x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)})$ *converges* to a draw from the target f.

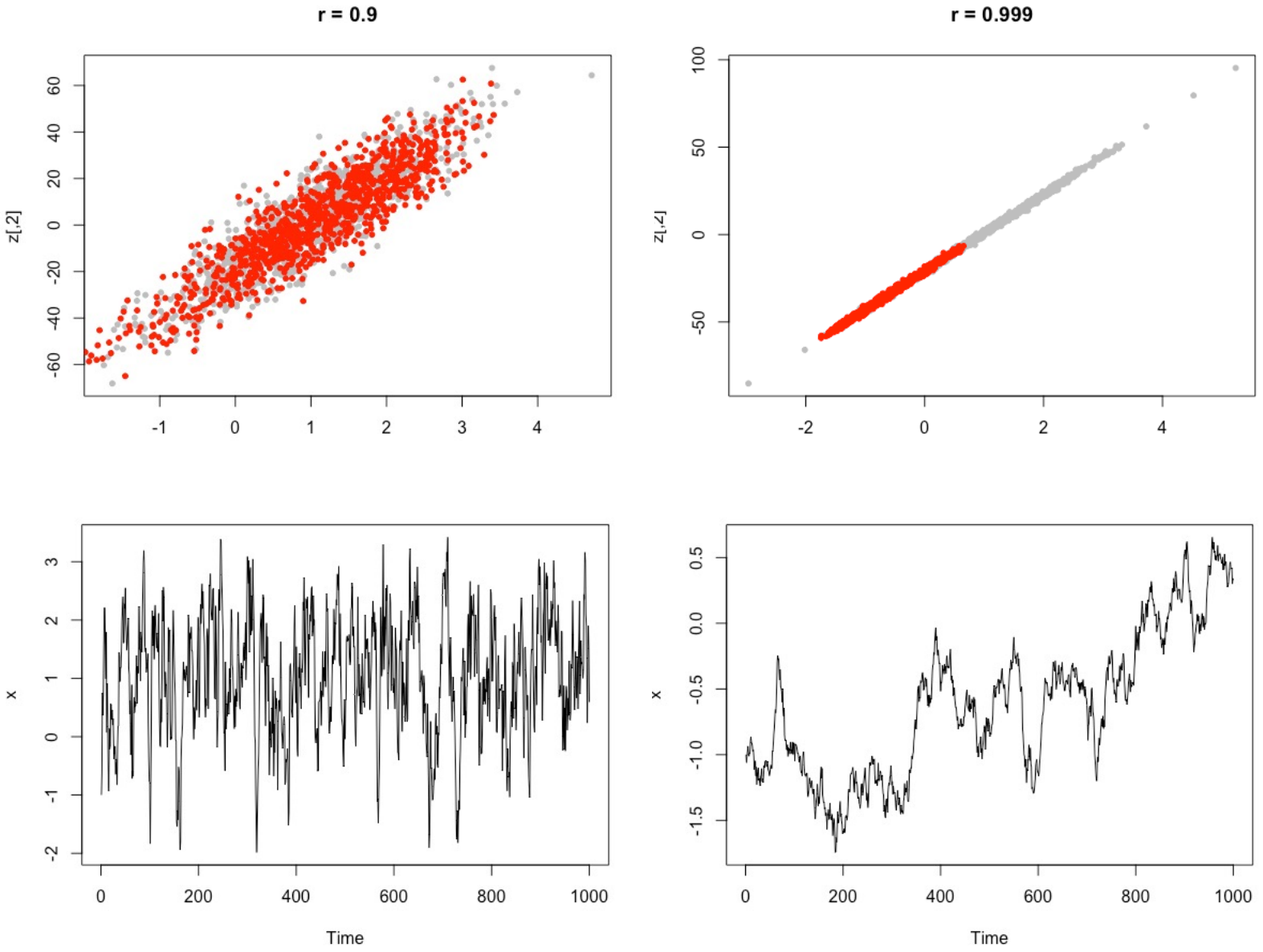
Gibbs Sampler Connection with Bayesian Models

- In Bayesian analysis the target density $f(x_1, \ldots, x_p)$ is actually the posterior of $\boldsymbol{\theta}$
- That is, the target is $f(\boldsymbol{\theta} \mid x_1, \ldots, x_n)$.
- Then in the limit we obtain a draw from $f(\boldsymbol{\theta} \mid x_1, \ldots, x_n)$.
- Gibbs sampler is still applicable to non-Baysian models. For example, in missing data it can be used to sample X_{mis} one component at a time, which results in a draw from $(Y_{\text{mis}} \mid Y_{\text{obs}})$ and $\boldsymbol{\theta}$ is then estimated from complete data $(Y_{\text{obs}}, Y_{\text{mis}} \mid \boldsymbol{\theta})$.
- In complicated models we can even "expand" the parameter space to use the so called *data augmentation*.

Gibbs Sampler Toy Example

- We will apply the Gibbs sampler to simulate from a bivariate normal distribution.
- Formulas for the conditional distributions are readily available.
- Note there is no need for such a method as exact methods exist for multivariate normal (e.g. MASS package).
- We will use the exact simulation to judge the performance of our Gibbs sampler under different values for the parameters.
- See R file

Gibbs sampler example



Exercise 1

Consider the autoexponential model

$$f(x_1, x_2, x_3) \propto e^{-(x_1+x_2+x_3+\theta_{12}x_1x_2+\theta_{13}x_1x_3+\theta_{23}x_2x_3)}$$

where $\theta_{12}, \theta_{13}, \theta_{23} > 0$ are given constants

Derive the 3 conditional distributions

$$x_i | x_j, x_k, i \neq j \neq k = 1,2,3$$

Answer: All of them are exponential distributions.

Normal Model Continued

Recall the normal model from last time:

$$X_1, \ldots, X_n | \theta, \sigma^2 \sim N(\theta, \sigma^2)$$

$$\theta | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$$

$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

This prior distribution relates the prior variance of θ to the sampling variance of our data in such a way that μ_0 can be thought of as κ_0 prior samples from the population. What if we want to specify our uncertainty about θ as being independent of σ^2, so that

$$\pi(\theta, \sigma^2) = \pi(\theta) \times \pi(\sigma^2)$$

Semiconjugate Prior

Suppose that

$$X_1, \ldots, X_n | \theta, \sigma^2 \sim \mathrm{N}(\theta, \sigma^2)$$
$$\theta | \sigma^2 \sim N(\mu_0, \tau_0^2)$$
$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

Then we have that:

$$\theta | x_1, \ldots, x_n, \sigma^2 \sim N(\mu_n, \tau_n^2)$$

where

$$\tau_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}, \mu_n = \tau_n^2 \left(\frac{n}{\sigma^2}\bar{x} + \frac{1}{\tau_0^2}\mu_0\right)$$

However, now the posterior of $\sigma^2 | x_1, \ldots x_n$ *is not* $IG\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right)$

Approach 1: Grid approximation

The posterior distribution is:

$$f(\theta, \sigma^2 | x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n | \theta, \sigma^2)\pi(\theta, \sigma^2)}{f(x_1, \dots, x_n)}$$
$$\propto f(x_1, \dots, x_n | \theta, \sigma^2)\pi(\theta, \sigma^2)$$

In a discrete approximation, we would approximate the "normalizing constant" $f(x_1, \dots, x_n)$ by using a grid over which we compute it numerically. Then we only need to calculate

$$f(x_1, \dots, x_n | \theta, \sigma^2)\pi(\theta, \sigma^2)$$
$$= \left[\prod_{i=1}^{n} dnorm(x_i, \theta, \sigma^2)\right] \times dnorm(\theta, \mu_0, \tau_0)$$
$$\times dinvgamma\left(\sigma^2, \frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

Example 1 (from last time)

We will use the values $\mu_0 = 165$ and $\tau_0^2 = 4, \nu_0 = 1, \sigma_0^2 = 0.01$

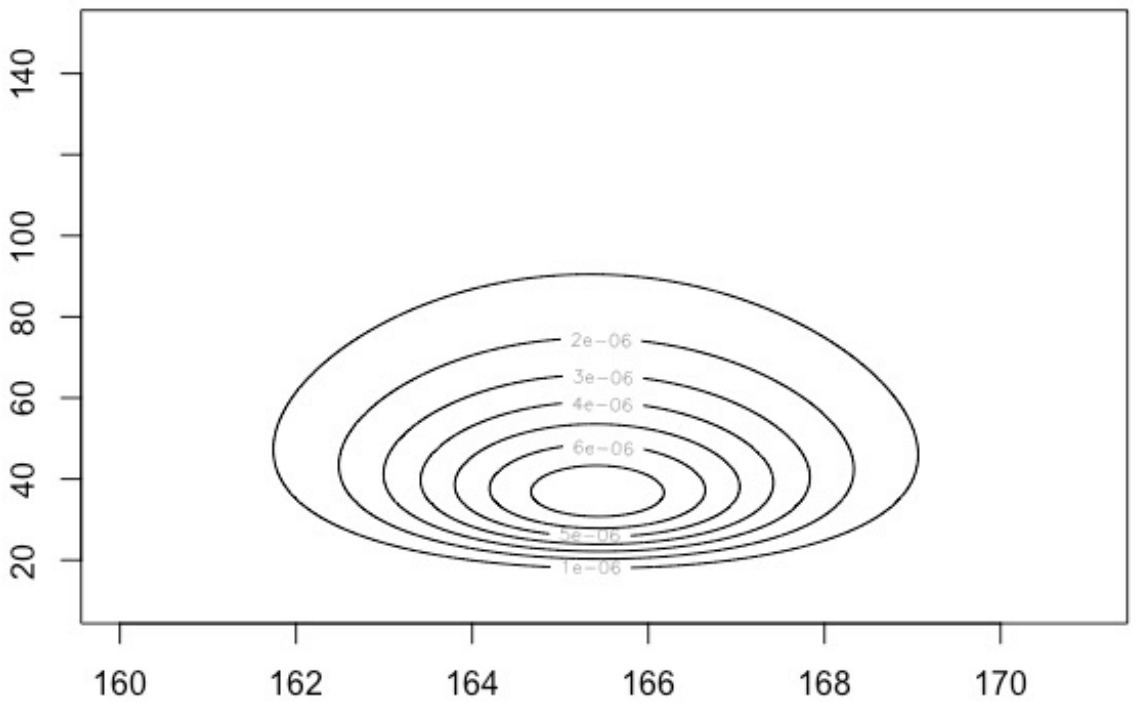
The sample mean and variance are:

$$\bar{x} = 165.52, s^2 = 52.453$$

The grid should be over (θ, σ^2) and we can use 100 × 100 or larger, evenly spaced apart.

See R code for posterior distributions.

Example 1 (from Lecture 1)



Exercise 1

Obtain the marginal posteriors of θ and σ^2 by adding over the grid the values of one parameter for each value of the other (see page 91 for details).

Approach 2: Gibbs sampler

We need to find the two conditional distributions. The conditional distribution of σ^2 given θ and $\{x_1,...,x_n\}$ is

$$f(\sigma^2|\theta, x_1, \dots, x_n) \propto f(x_1, \dots, x_n|\theta, \sigma^2)\pi(\theta, \sigma^2)$$
$$\propto f(x_1, \dots, x_n|\theta, \sigma^2)\pi(\sigma^2)$$
$$\propto \left[\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2}}\right] \times \left[(\sigma^2)^{\frac{\nu_0}{2}-1} e^{-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2}{2}}\right]$$
$$\propto (\sigma^2)^{\frac{\nu_0+n}{2}-1} e^{-\frac{1}{\sigma^2}\left(\frac{\nu_0\sigma_0^2}{2}+\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2}\right)}$$

This means that the distribution is IG:

$$\sigma^2|\theta, x_1, \dots, x_n \sim IG\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2(\theta)}{2}\right)$$

where

$$\nu_n = \nu_0 + n, \sigma_n^2(\theta) = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + n\, s_n^2(\theta)], s_n^2(\theta) = \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{n}$$

Gibbs Sampler

Suppose we have a starting value $\sigma^{2(1)}$

Then generate

$$\theta^{(1)} \sim f\left(\theta \middle| \sigma^{2(1)}, x_1, \ldots, x_n\right)$$

followed by

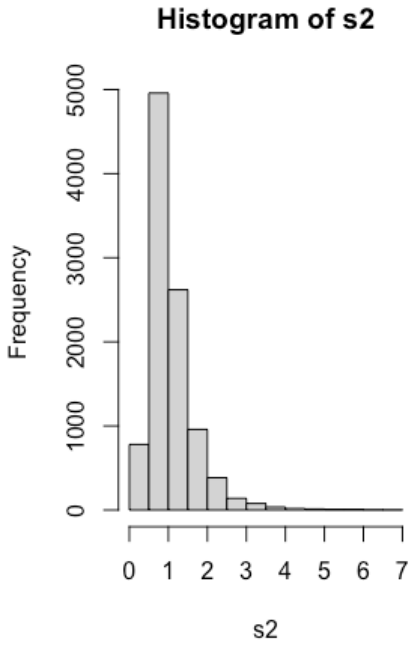
$$\sigma^{2(2)} \sim f\left(\sigma^2 \middle| \theta^{(1)}, x_1, \ldots, x_n\right)$$

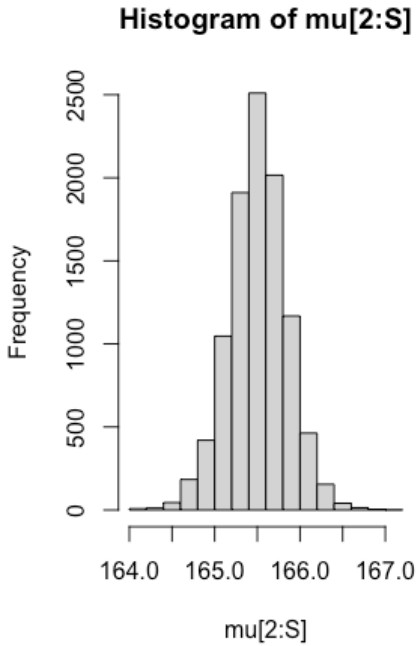
Note that the conditional of θ was Normal, given at the beginning of the discussion.

Keep iterating *S* times between the two until "convergence".

At the end we will have an approximate sample $(\theta^{(1)}, \sigma^{2(1)})$, .., $(\theta^{(S)}, \sigma^{2(S)})$, which we can use to estimate the joint and marginal posteriors of the parameters.

Example 1 continued





Next time

MCMC diagnostics