Week 12: Bayesian Computations (continued)

MATH-517 Statistical Computation and Visualization

Linda Mhalla

2024-12-06

Bayes' Rule

Let X be a random variable and θ a parameter, considered also a random variable:

$$f_{X,\theta}(x,\theta) = \underbrace{f_{X|\theta}(x \mid \theta)}_{\text{likelihood}} \underbrace{f_\theta(\theta)}_{\text{prior}} = \underbrace{f_{\theta|X}(\theta \mid x)}_{\text{posterior}} f_X(x).$$

- likelihood = frequentist model
- likelihood & prior = Bayesian model

Rewritten:

$$f_{\theta|X=x_0}(\theta \mid x_0) \propto f_{X|\theta}(x_0 \mid \theta) f_\theta(\theta),$$

in words: posterior $\propto$ likelihod $\times$ prior

- posterior has all the answers, but is often intractable $\Rightarrow$ MCMC

Metropolis–Hastings

Metropolis–Hastings (M–H) algorithm:

- **Input**: a proposal density $q(y \mid x)$, the target f (up to a constant)
- **for** $t = 1, 2, ...,$ update $X^{(t-1)}$ to $X^{(t)}$ by
  - generate $U^{(t)} \sim q(\cdot \mid X^{(t-1)})$
  - define

$$\alpha(X^{(t-1)}, U^{(t)}) = \min\left\{1, \frac{f(U^{(t)})q(X^{(t-1)} \mid U^{(t)})}{f(X^{(t-1)})q(U^{(t)} \mid X^{(t-1)})}\right\}$$

  - set $X^{(t)} := U^{(t)}$ with probability $\alpha(X^{(t-1)}, U^{(t)})$
  - otherwise set $X^{(t)} := X^{(t-1)}$

Metropolis–Hastings

- Under some conditions (see last week's lecture), the chain is ergodic (geometrically, i.e., exponentially fast convergence to stationarity or uniformly/strongly)
- MH with random walk is geometrically ergodic if and only if its target distribution has exponentially light tails; see this paper and this paper
- Independent MH is geometrically ergodic if and only if its proposal density is bounded below by a constant multiple of the target density; see this paper
- Metropolis–Hastings: extremely versatile approach to MCMC, but a good proposal (yielding good mixing rate/exploration of space) can be hard to find
- For the common random walk M–H, this is a scaling issue
  - too small and the chain will move too slowly; too large and the proposals will usually be rejected

Adaptive M–H: few words

- trial and error
  - if the acceptance rate seems too high, then we increase the proposal scaling
  - if the acceptance rate seems too low, then we decrease the scaling
- or let the computer decide on the fly
  - suppose we have a family $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ of possible Markov chains, each with stationary distribution $f(\cdot)$. Let the computer choose among them! At iteration n, use Markov chain P_{Γ_n}, where $\Gamma_n \in \mathcal{Y}$ chosen according to some adaptive rules (depending on chain's history, etc.)
  - $\Rightarrow$ Markov property and stationarity are destroyed. Will it still converge? Use "finite adaptation", i.e., stop adapting after a while

See Roberts and Rosenthat (2009) for examples of adaptive MCMC

Example (to follow): optimal proposal depends on the covariance matrix of the target, then take the empirical covariance at each step n

Adaptive M–H: few words

It is known from Roberts et al. (1997) and Roberts and Rosenthal (2001) that the proposal $\mathcal{N}(x, (2.38)^2\Sigma/d)$ is optimal in a particular large-dimensional context

- Haario et al (2001) propose a simple and effective adaptive random walk Metropolis
- run the MH with random walk with a Gaussian proposal for a fixed number of iterations for $s < s_0$
- estimate of covariance at state s

$$\Sigma^{(s)} = \frac{1}{s}\left(\sum_{i=1}^{s} X^{(i)} X^{(i)^T} - s\bar{X}^{(s)}\bar{X}^{(s)^T}\right)$$

- proposal for $s > s_0$ with $\delta = 2.38/\sqrt{d}$

$$U^{(s+1)} \sim \mathcal{N}\left(X^{(s)}, \delta^2\left(\Sigma^{(s)} + \epsilon I_d\right)\right)$$

Gibbs Sampler

Idea: take advantage of the hierarchical structure, i.e., decompose the multidimensional distribution into *full conditionals* and draw from those in a cyclic manner

- not as universal as M–H, since calculation of the conditional distributions not always possible

Full conditional: $f_i(x_i \mid x_{-i}) = f_i(x_i \mid x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$

The Gibbs sampler algorithm based on the target distribution f is

1. use the full conditional densities $f_1, \dots, f_d$ from f
2. start with the random variable $\mathbf{X} = (X_1, \dots, X_d)^\top$
3. simulate from the conditional densities

$$X_i \mid x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_d \sim f_i\left(x_i \mid x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_d\right)$$

for $i = 1, 2, \dots, d$

Systematic Gibbs Sampler

The systematic Gibbs sampler proceeds as follows from initial $x^{(0)} = (x_1^{(0)}, \dots, x_d^{(0)})^\top$:

- **for** $t = 1, 2, \dots$
  - generate $x_1^{(t)}$ from $X_1 \mid X_2 = x_2^{(t-1)}, X_3 = x_3^{(t-1)}, \dots, X_d = x_d^{(t-1)}$
  - generate $x_2^{(t)}$ from $X_2 \mid X_1 = x_1^{(t)}, X_3 = x_3^{(t-1)}, \dots, X_d = x_d^{(t-1)}$
  - ...
  - generate $x_d^{(t)}$ from $X_d \mid X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \dots, X_{d-1} = x_{d-1}^{(t-1)}$

$\Rightarrow$ full conditionals $f_1, \dots, f_d$ are the only densities used for simulation

The transition kernel is

$$K\left(x^{(t-1)}, x^{(t)}\right) = f_{X_1|X_{-1}}\left(x_1^{(t)} \mid x_2^{(t-1)}, \dots, x_d^{(t-1)}\right) \times f_{X_2|X_{-2}}\left(x_2^{(t)} \mid x_1^{(t)}, x_3^{(t-1)}, \dots, x_d^{(t-1)}\right) \times \cdots$$
$$\times f_{X_d|X_{-d}}\left(x_d^{(t)} \mid x_1^{(t)}, \dots, x_{d-1}^{(t)}\right)$$

- admits f as stationary distribution (show that $\int k(x, y) f(x) dx = f(y)$)
- does not satisfy the detailed balance condition
- LLN applies if f satisfies positivity condition

Gibbs Sampler and Positivity Condition

Definition:

A distribution with density $f(x_1, x_2, \dots, x_d)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the positivity condition if for all $x_1, \dots, x_d$ such that $f_{X_i}(x_i) > 0$, we have $f(x_1, x_2, \dots, x_d) > 0$ (support of joint $= \prod$ support of margins)

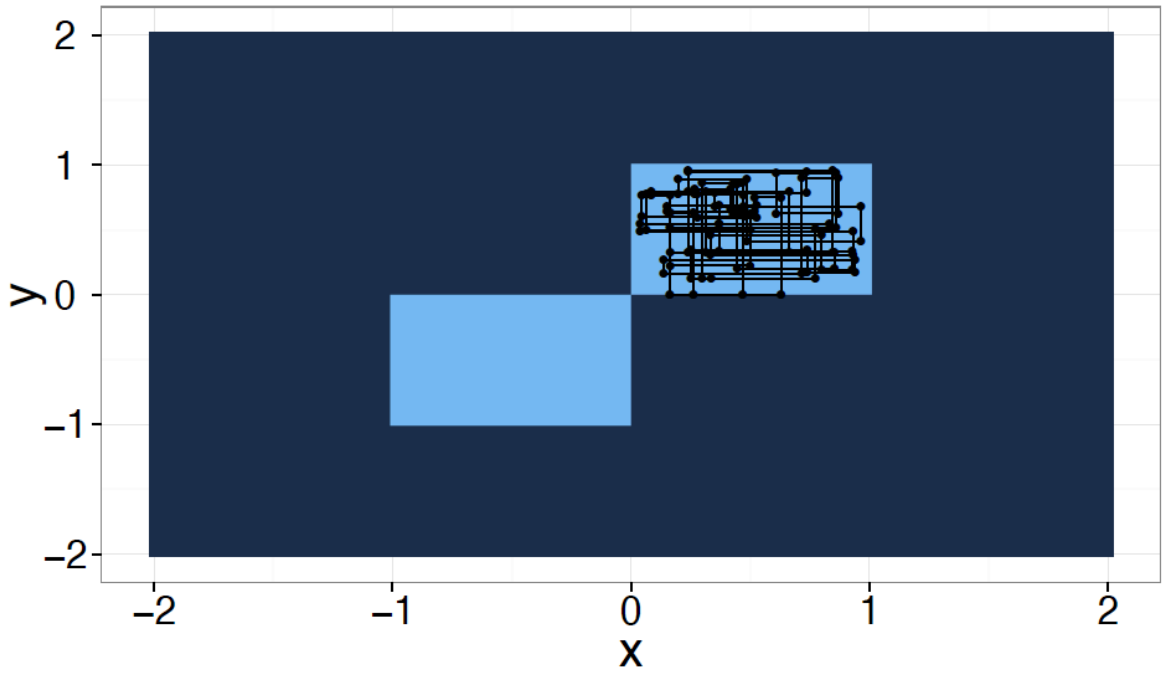
Result: If the target distribution f satisfies the positivity condition, then the MC generated by the systematic Gibbs sampler satisfies

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} h\left(X^{(t)}\right) = \int h(x) df(x)$$

for any integrable function $h : \mathbb{X} \to \mathbb{R}$

Positivity Condition Violated

Gibbs sampling targeting $\pi(x, y) \propto \mathbb{1}_{[-1,0]\times[-1,0]\cup[0,1]\times[0,1]}(x, y)$



Gibbs sampler can be reducible (we cannot get arbitrarily close to any of the points, by making moves parallel to the axes)

Remarks on Gibbs Sampler

Although the systematic Gibbs sampler does not satisfy detailed balance, each of its d components does

- Suppose we are at point $\mathbf{x}$ and decide to modify component j of $\mathbf{x}$ to take the value z
- Let $\mathbf{y}$ be the point with $y_j = z$ and $y_k = x_k$ for $k \neq j$
- If $\mathbf{y}$ is used as the proposal in Metropolis–Hastings, the M–H ratio is:

$$\alpha(\mathbf{x}, \mathbf{y}) = \frac{f(\mathbf{y})q(\mathbf{x} \mid \mathbf{y})}{f(\mathbf{x})q(\mathbf{y} \mid \mathbf{x})} = \frac{f(x_{-j})f(z \mid x_{-j})f(x_j \mid x_{-j})}{f(x_{-j})f(x_j \mid x_{-j})f(z \mid x_{-j})} = 1$$

$\Rightarrow$ Updating component j of $\mathbf{x}$ by sampling from its full conditional distribution can be viewed as a M–H proposal that is **never rejected**!

$\Rightarrow$ this motivates the *random scan Gibbs sampler*

Random Scan Gibbs Sampler

Algorithm: Random scan Gibbs sampler Let $\left(X_1^{(0)}, \ldots, X_d^{(0)}\right)^\top$ be the initial state then iterate for $t = 1, 2, \ldots$

1. sample an index j from a distribution on $\{1, \ldots, d\}$ (typically uniform)
2. sample $X_j^{(t)} \sim f_{X_j|X_{-j}}\left(\cdot \mid X_1^{(t-1)}, \ldots, X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, \ldots, X_d^{(t-1)}\right)$ and set $X_k^{(t)} := X_k^{(t-1)}$ for $k \neq j$

$\Rightarrow$ Random scan Gibbs is a multi-component Metropolis–Hastings sampler with acceptance probability equal to 1 and transition kernel

$$K\left(x^{(t-1)}, x^{(t)}\right) = \frac{1}{d} \sum_{j=1}^{d} f_{X_j|X_{-j}}\left(x_j^{(t)} \mid x_{-j}^{(t-1)}\right) \delta_{x_{-j}^{(t-1)}}\left(x_{-j}^{(t)}\right)$$

$\Rightarrow$ satisfies detailed balance and admits f as stationary distribution

Toy Example

Using the systematic Gibbs sampler, calculate $P(X_1 \geq 0, X_2 \geq 0)$ for

$$X = (X_1, X_2)^\top \sim \mathcal{N}\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho \\ \rho & \sigma_2^2\end{pmatrix}\right)$$

Easy, since Gaussian conditionals are Gaussian:

$$X_i \mid X_j = x_j \sim \mathcal{N}\left(\mu_i + \frac{\rho}{\sigma_j^2}(x_j - \mu_j), \sigma_i^2 - \frac{\rho^2}{\sigma_j^2}\right)$$

The Gibbs sampler proceeds as follows in this case

1. Sample $X_1^{(t)} \sim \mathcal{N}\left(\mu_1 + \rho/\sigma_2^2\left(X_2^{(t-1)} - \mu_2\right), \sigma_1^2 - \rho^2/\sigma_2^2\right)$
2. Sample $X_2^{(t)} \sim \mathcal{N}\left(\mu_2 + \rho/\sigma_1^2\left(X_1^{(t)} - \mu_1\right), \sigma_2^2 - \rho^2/\sigma_1^2\right)$

E.g., for $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ and $\rho = 0.3$, we have...

Toy Example

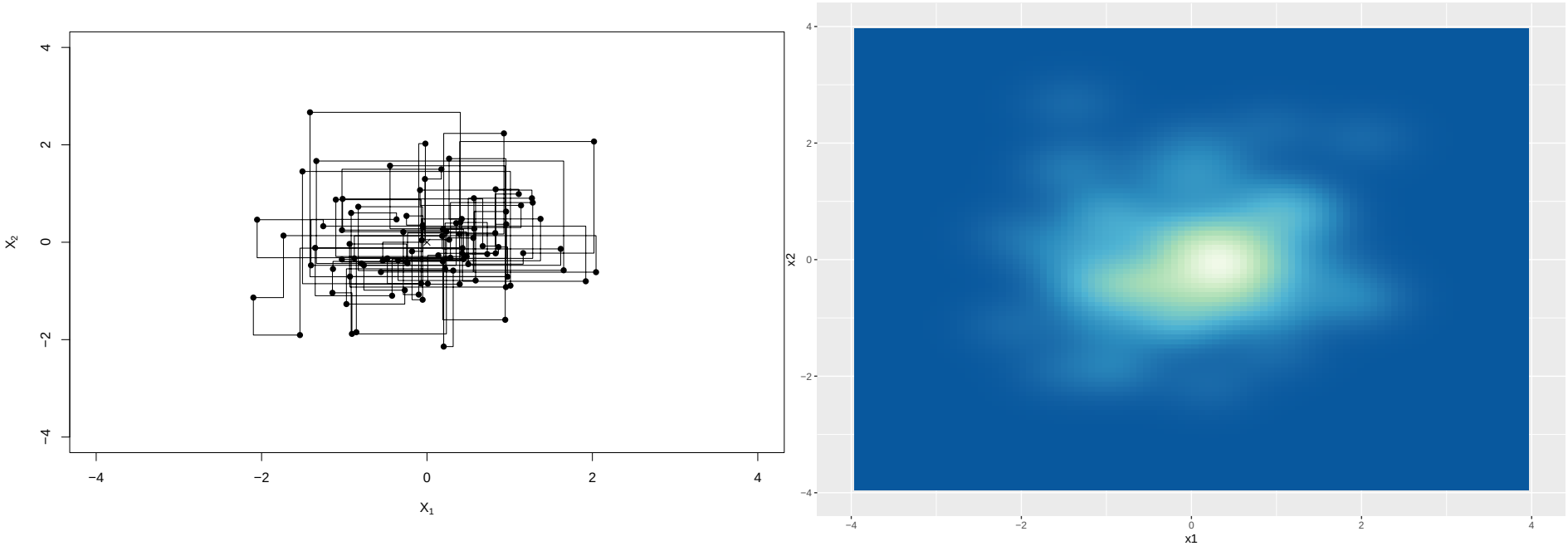
```
set.seed(123)
burnin <- 1000
TT  <- 2000
X1  <- rep(0, burnin+TT)
X2  <- rep(0, burnin+TT)
rho <- 0.3
X1[1] <- 0
X2[1] <- 0
for(t in 2:(burnin+TT)){
  X1[t] <- rnorm(1,0+rho/1*(X2[t-1]-0), sqrt(1-rho^2/1))
  X2[t] <- rnorm(1,0+rho/1*(X1[t]-0), sqrt(1-rho^2/1))
}
X1 <- X1[-(1:burnin)]
X2 <- X2[-(1:burnin)]

sum((X1 >= 0 & X2 >= 0 ))/TT # empirical P(X1 >= 0, X2 >= 0)
```

Toy Example

Markov chain $X^{(t)}$ has correlated successive samples

First 100 steps (with $\rho = 0.3$)

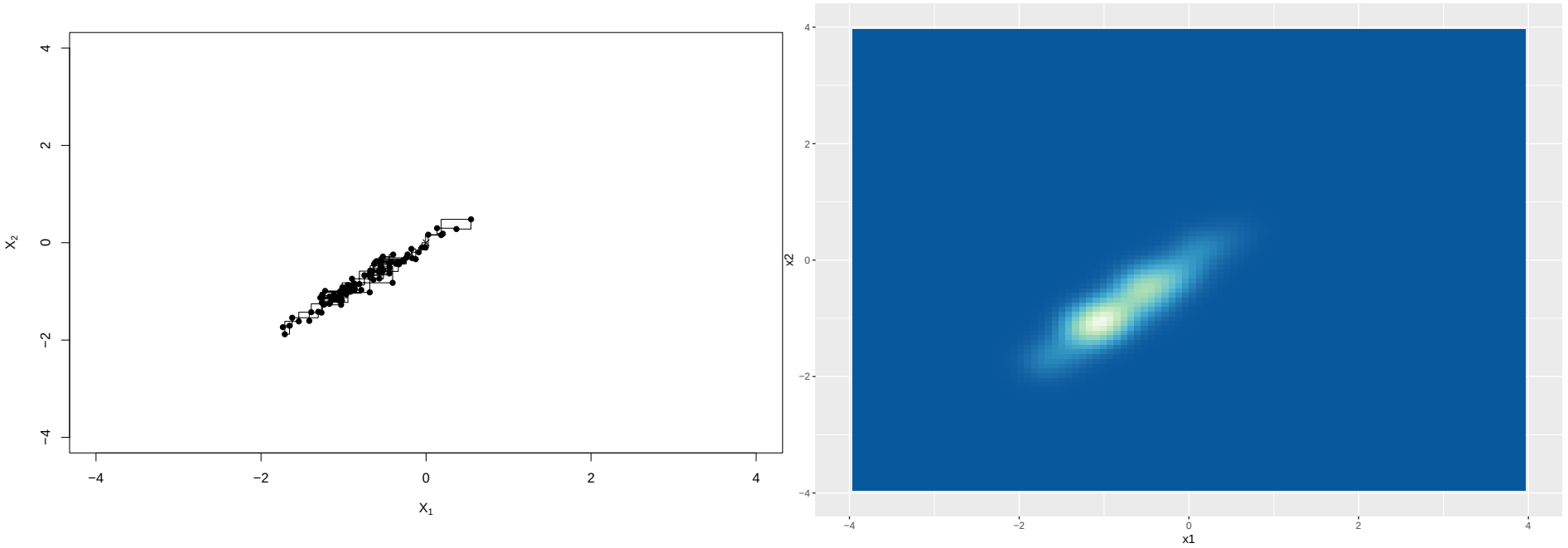


$P(X_1 \geq 0, X_2 \geq 0)$ is estimated at 0.298 (true ≈ 0.2984)

Toy Example

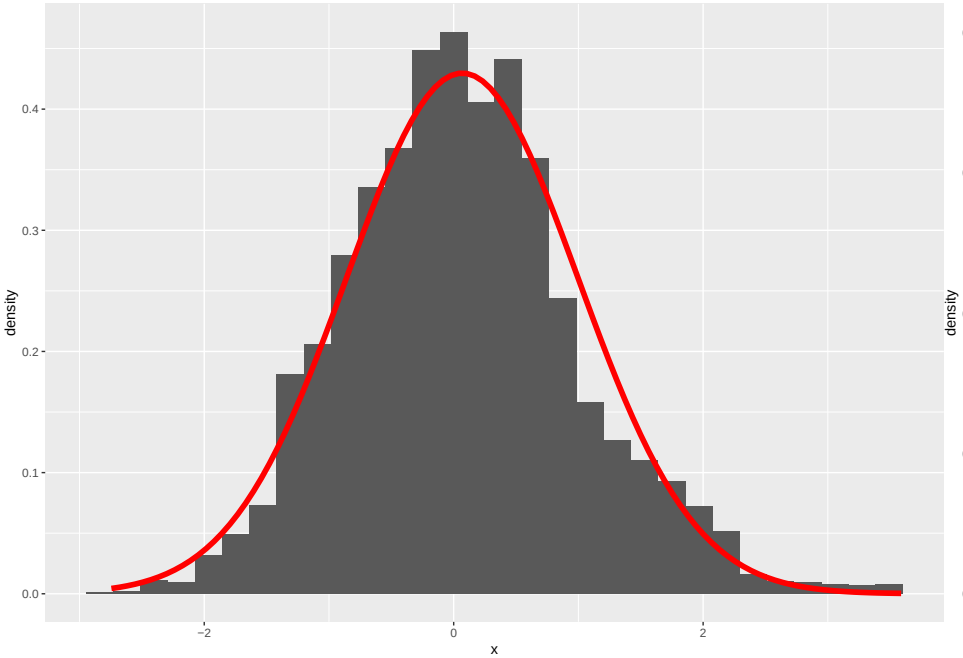
Markov chain $X^{(t)}$ has strongly correlated successive samples $\Rightarrow$ chain mixes slowly

First 100 steps (with $\rho = 0.99$)

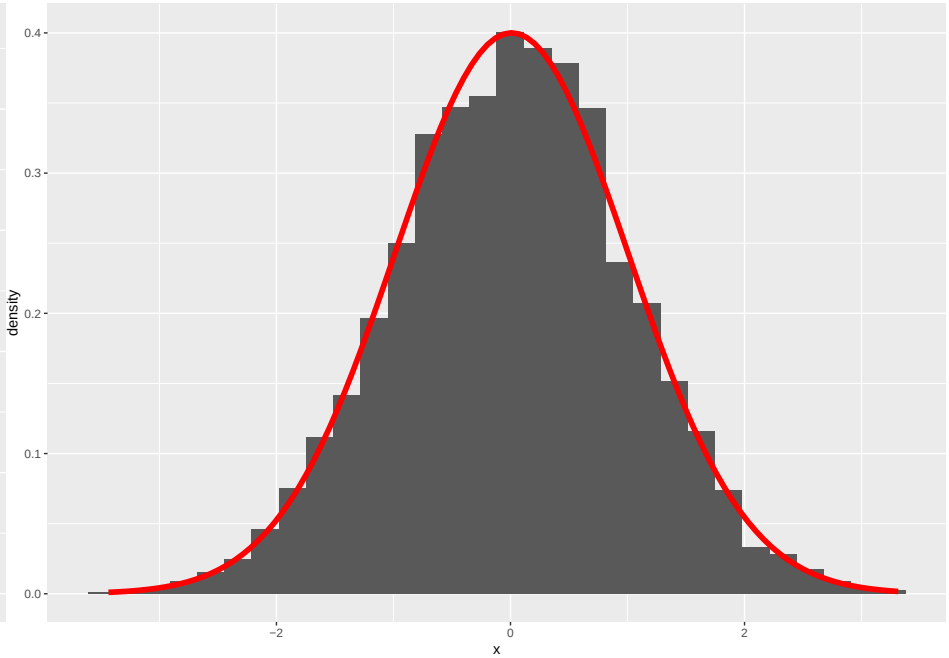


$P(X_1 \geq 0, X_2 \geq 0)$ is estimated at 0.5635 (true ≈ 0.4775)

Toy Example



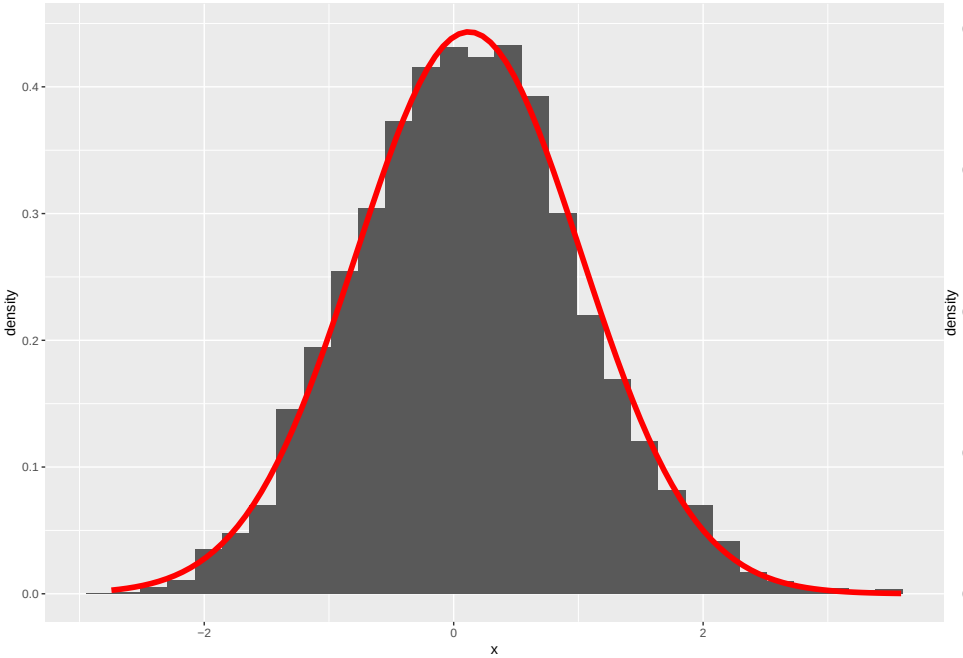
(a) large correlation



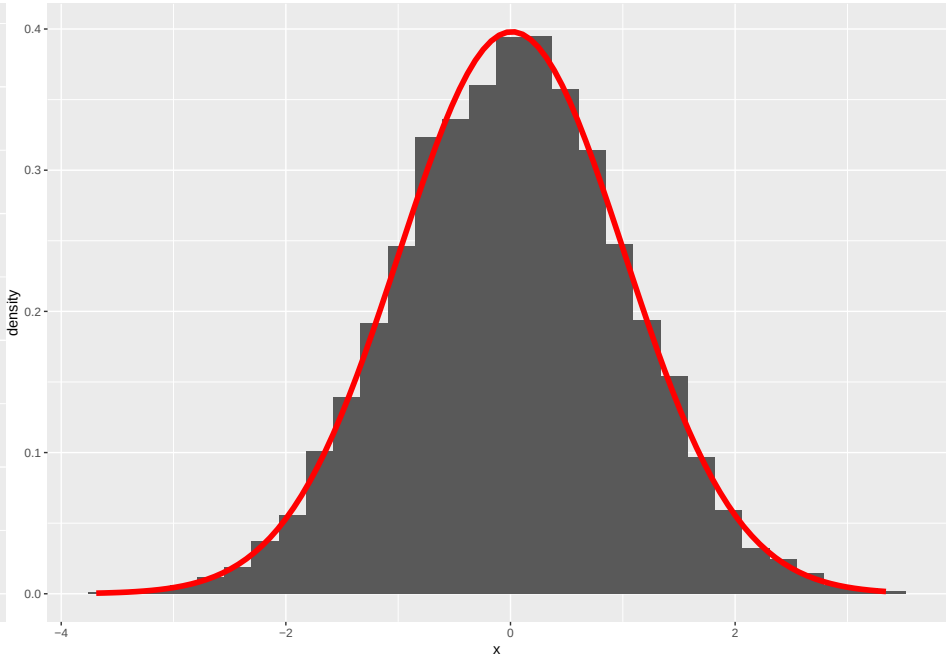
(a) small correlation

Histogram of first component after 4000 iterations

Toy Example



(a) large correlation



(a) small correlation

Histogram of first component after 10000 iterations

Metropolis-within-Gibbs

What if sampling from full conditionals isn't easy for Gibbs?

- do a single Metropolis–Hastings step instead

What if parameters are naturally grouped in a real application?

- e.g., some parameters correspond to location and others to scale
- location parameters can usually be sampled at once, conditionally on all the other parameters
  - *blocked Gibbs sampler*: blocks of variables are updated by sampling from their joint conditional on all other variables
  - potentially via a M–H step

Limitations of the Gibbs sampler

- limits the choice of target distributions
- requires some knowledge of f
- is multi-dimensional, by construction

Gibbs Sampling in Practice

- Many posterior distributions can be automatically decomposed into conditional distributions by computer programs

→ This is the idea behind `BUGS` (Bayesian inference Using Gibbs Sampling) or `JAGS` (Just another Gibbs Sampler) with `R` packages

- `rjags` (see the JAGS user manual)
- `runjags` Denwood (2016): for additional functionalities, including parallel computing

The `Stan` platform implements MCMC sampling using the Hamiltonian Monte Carlo (transitions rely on derivatives of the target) and its adaptive variant NUTS

→ available in different languages (`R`, `Python`, `Julia`)

Output Analysis

MCMC compared to MC:

- sacrifices independence for more versatility
  - ergodic theory: independence not really needed in the long run
- in practice, the question is: what is a long enough run?
- just inspect the samples drawn (after discarding burnin)
  - check whether the acceptance rate is reasonable
  - visualize graphical outputs (to follow)
  - calculate diagnostic statistics (to follow)
- in reality, we can never know
  - silent failure?! E.g., careless use of Gibbs (conditional distributions are well defined but their combination does not correspond to any joint distribution...), or positivity condition violated
  - but sometimes, we can know for sure that there is a problem!

Output Analysis: Multiple starting points

Simple ideas such as running multiple chains and checking that they are converging to similar distributions are often employed in practice

- We start M chains from various (dispersed) starting points
- After enough iterations, the starting point should not matter and hence we should obtain the same results based on each chain
- We have the classical "sum of squares" decomposition in "intra group" and "inter group" terms:

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\left(X_{m,t}-\bar{X}_{\cdot,\cdot}\right)^2 = \sum_{m=1}^{M}\sum_{t=1}^{T}\left(\bar{X}_{m,\cdot}-\bar{X}_{\cdot,\cdot}\right)^2 + \sum_{m=1}^{M}\sum_{t=1}^{T}\left(X_{m,t}-\bar{X}_{m,\cdot}\right)^2$$

Output Analysis: Multiple starting points

- This leads to considering

$$W = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T-1}\sum_{t=1}^{T}\left(X_{m,t} - \bar{X}_{m,.}\right)^2$$

$$B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\bar{X}_{m,\cdot} - \bar{X}_{.,.}\right)^2$$

$$V = \left(1 - \frac{1}{T}\right)W + B$$

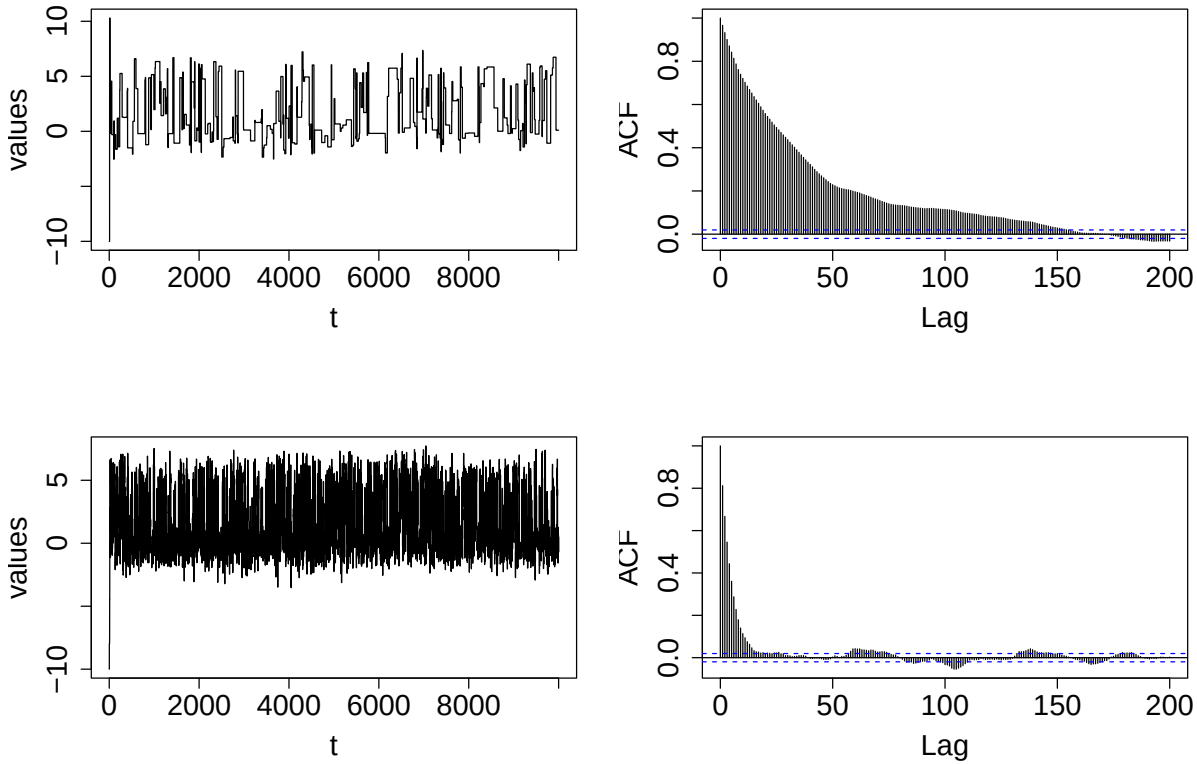
  In principle W (mean of empirical variance within each chain) and V (empirical variance from all chains) should both converge to the true variance of the target distribution $\rightarrow$ plot $\sqrt{V/W}$ (for different iterations) and compare it to 1 (version in `R` is slightly different)

This leads to the shrink factor of Gelman–Rubin: variance between chains relative to variance within chains (if multiple chains reached the target then this factor should be 1)

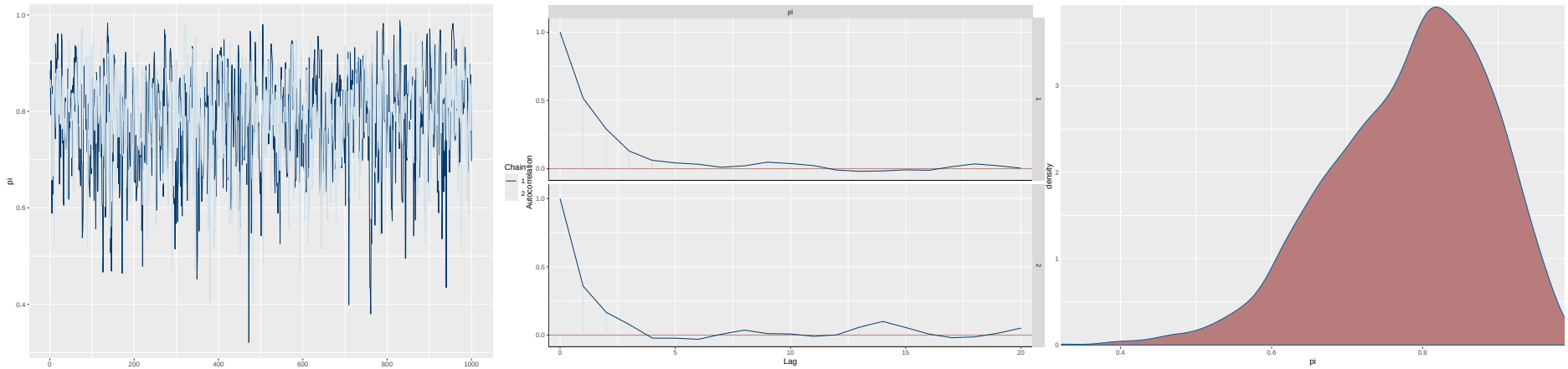
- > 1 indicates instability, with variability in the combined chains exceeding that within the chains
- rule of thumb: red flag if > 1.05

Output Analysis

- trace plots are often used to informally assess stochastic convergence
  - if MCMC is working, they should look like a "fat, hairy caterpillar"
- ACF (autocorrelation function) plots display the autocorrelation within a chain as a function of the lag
  - if the ACF takes too long to decay to 0, the chain exhibits a high degree of dependence and will tend to get stuck



Output Analysis: Beta-Binomial Model



- the chains mix quickly (move quickly around plausible values of the posterior)
- the autocorrelation quickly drops off
- shrink factor ≈ 1 (stability across parallel chains)

$\Rightarrow$ if not, use more iterations or try thinning, i.e., use every k-th observation (reduces correlation) or different scaling of proposal (if M–H)

Simple but Real Example

- the height (in inches) of college students has $\mathcal{N}(\mu, \sigma^2)$
  - we work with σ, i.e., the standard deviation instead of variance
- only binned data available

```
X
(-Inf,60]   (60,62]   (62,64]   (64,66]   (66,68]   (68,70]   (70,72]   (72,74] (74, Inf]
       32        77       110       108       107        78        81        34        20
```

- multinomial data, probabilities depend on μ and σ
  - e.g., prob. of an obs. falling into $(60, 62]$ is $\Phi_{\mu,\sigma}(62) - \Phi_{\mu,\sigma}(60)$
- likelihood:

$$f(d \mid \mu, \sigma) \propto \prod_{j=1}^{9} \{\Phi_{\mu,\sigma}(a_j) - \Phi_{\mu,\sigma}(a_{j-1})\}^{d_j} =: \ell(\mu, \sigma)$$

- prior: $f(\mu, \sigma) = 1/\sigma$
  - improper prior (Jeffrey's prior)
  - changing variable $\lambda = \log(\sigma)$ removes $1/\sigma$ from the posterior

Posterior:

$$f(\mu, \sigma \mid D = d) \propto \ell\{\mu, \exp(\lambda)\}$$

Real but Simple Example

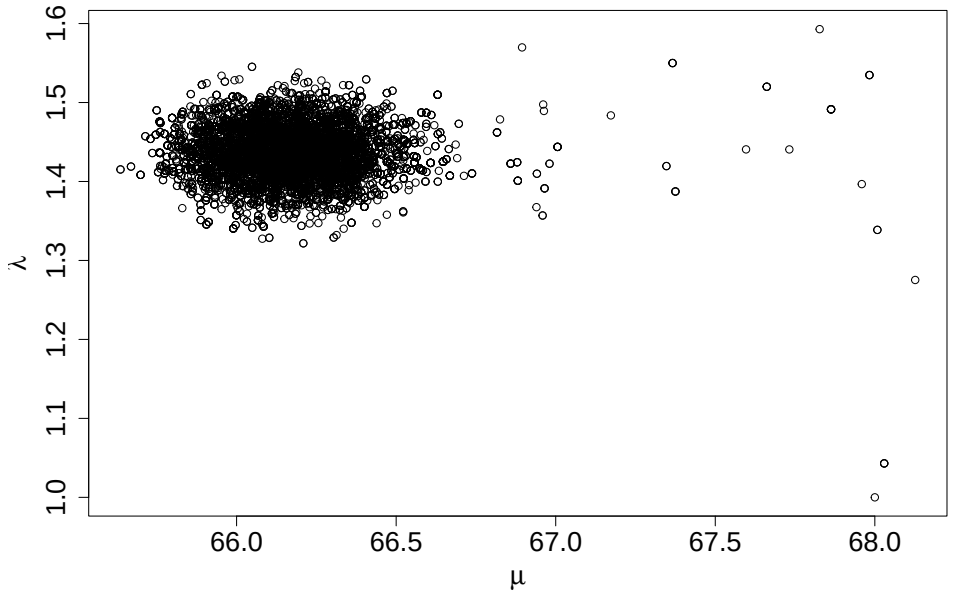
- **Aim**: sample from posterior using normal random walk M–H

$$U^{(t+1)} = X^{(t)} + sZ$$

  where $Z \sim \mathcal{N}(0, \Sigma)$ and $s > 0$ is a scale parameter
  - overparametrization for the sake of convenience (debatable)
- for MH we have to choose
  - starting point $(\mu^{(0)}, \lambda^{(0)})^\top$
  - scale s
  - covariance Σ

Real but Simple Example

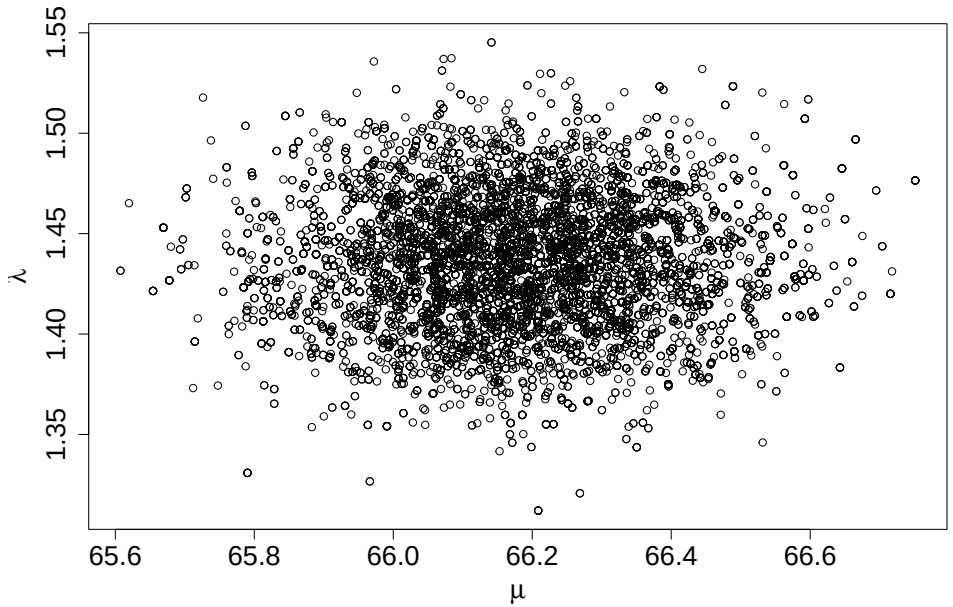
- Looking at the binned data, why not take
  - $(\mu^{(0)}, \lambda^{(0)})^\top = (68, 1)^\top$
  - scale $s = 1 \Rightarrow$ acceptance too low (0.009), so let's take $s = 0.1$
  - covariance $\Sigma = I_{2\times 2}$



Acceptance rate: 0.3134

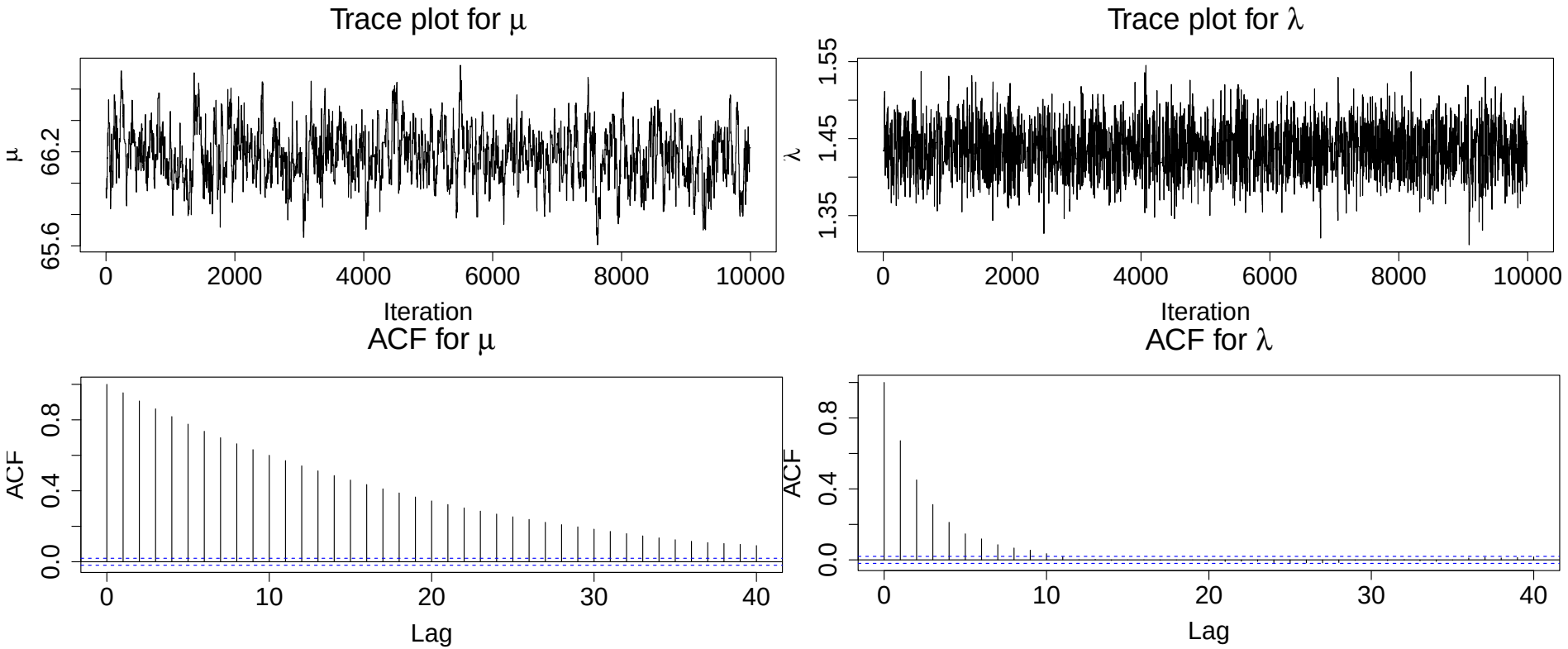
Real but Simple Example

- above starting point chosen badly
- normally taken care of by burnin, here let's re-run
  - $(\mu^{(0)}, \lambda^{(0)})^\top = (66, 1.4)^\top$



Acceptance rate: 0.3196

Real but Simple Example - Ouput Check



Real but Simple Example - Ouput Check

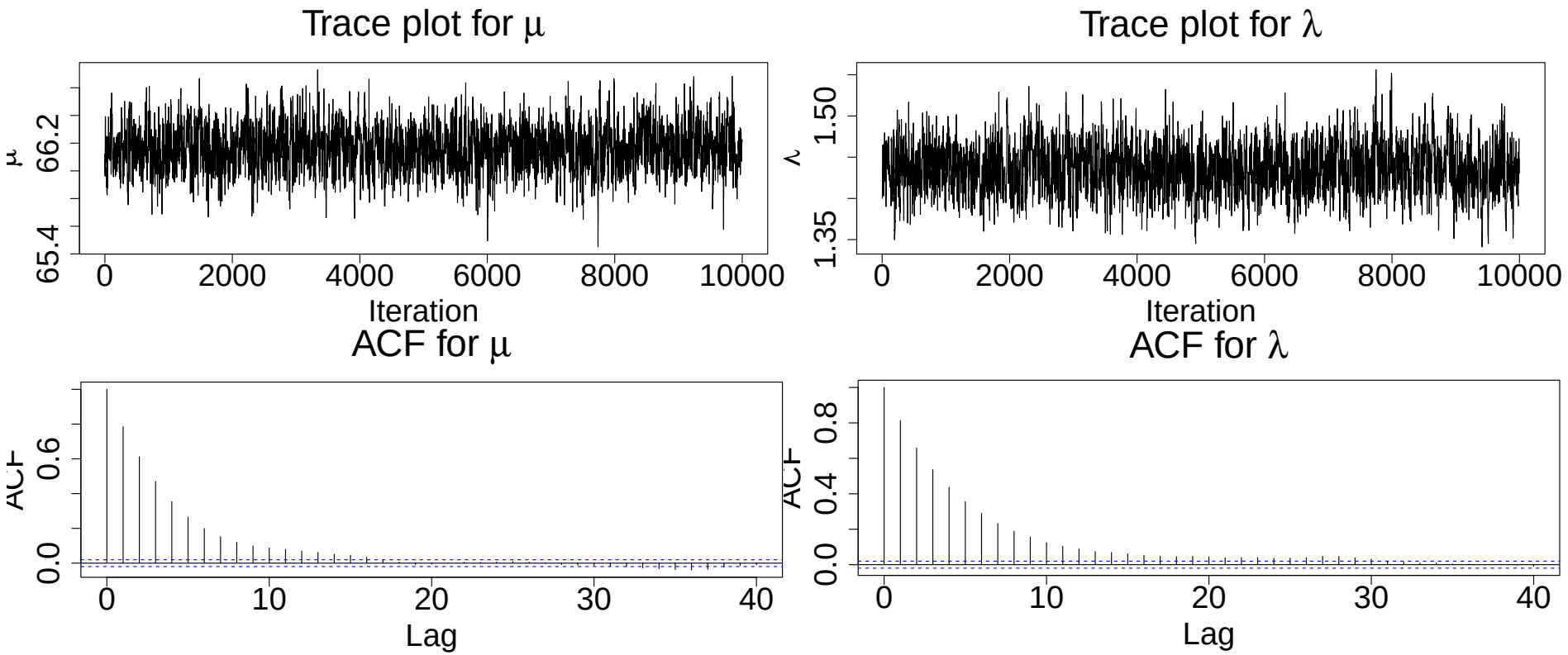
- the plots above look good, but values of μ are correlated for too long
- their correlation can be reduced by taking Σ diagonal with the variance for μ higher than that for λ
- actually, why not take Σ estimated from our previous run

```
             [,1]         [,2]
[1,] 3.035891e-02 7.329492e-05
[2,] 7.329492e-05 9.306388e-04
```

- acceptance too high with our $s = 0.1$ now, let's increase s
  - $s = 1$ gives 58%
  - let's take $s = 2$

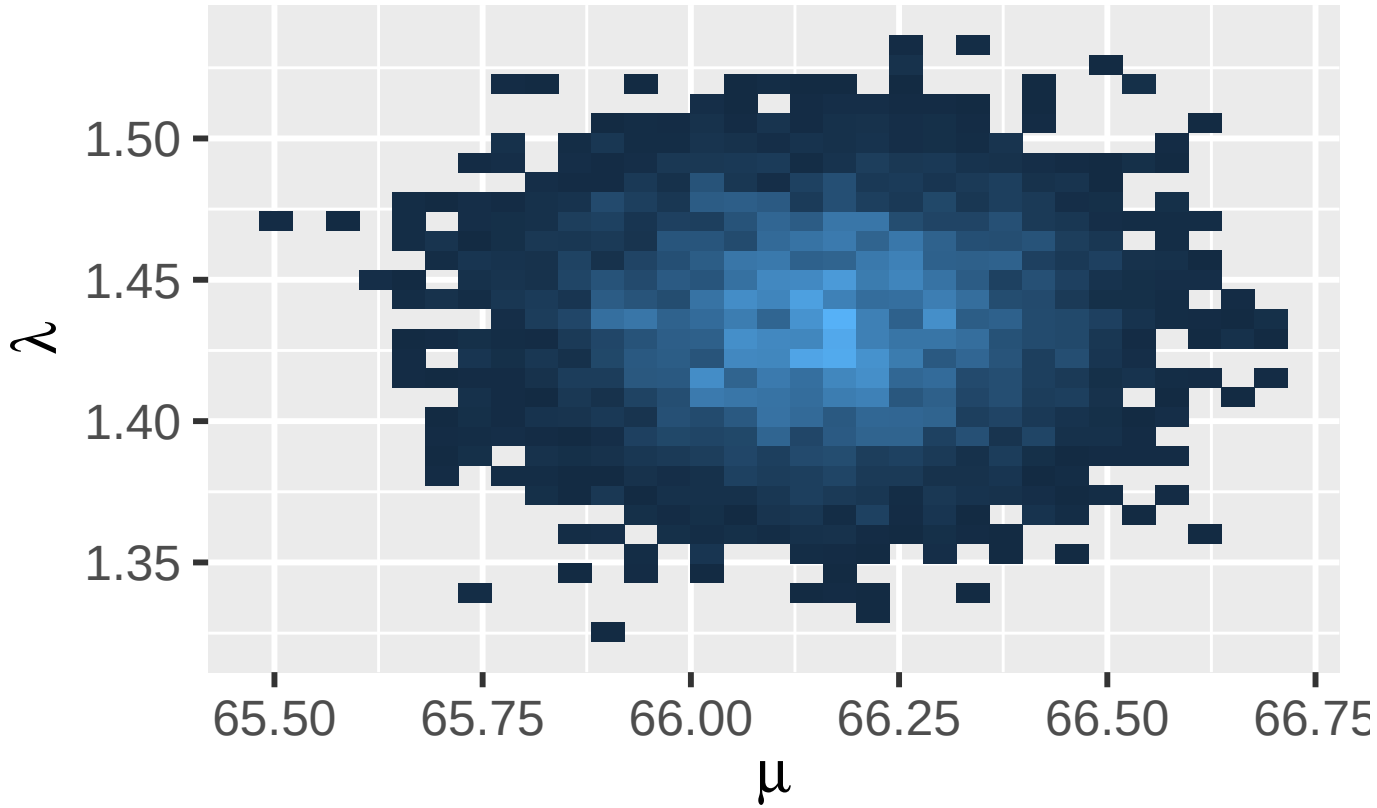
Real but Simple Example - Final Run

Let's analyze the output again



Acceptance rate: 0.5386

Real but Simple Example - Estimated Posterior



The posterior mean estimates are $\hat{\mu} = 66.159$ and $\hat{\lambda} = 1.435$

Final Thoughts

- Bayesianism is a different way of thinking about problems
  - e.g., hierarchical models
- prior versus no prior
- MLE versus MAP
- sampling not the only way to be Bayesian
  - *variational* methods (back to optimization)
  - *empirical* Bayes (back to frequentism)
- Hamiltonian MC and NUTS
  - explore the space in an adaptive way
- BUGS & JAGS
  - packages for Bayesian computations (JAGS has R interface `rjags`)
  - uses model structure and Gibbs sampling whenever possible
- STAN
  - a package with R interface `rstan`
  - uses NUTS
- silent failure!?
  - multimodal distributions problematic for sampling
  - plateau regions problematic for optimization

Final Thoughts

- as sample size $|D|$ grows:
  - at first, we are going away from the prior, and the posterior is getting complicated
  - then, the posterior becomes more and more regular (courtesy of CLT) and the prior serves as a bit of regularization
  - eventually, the prior stops mattering
    - back to frequentism in the large sample limit
- in every statistical task, there are three sources of error:
  - data is random (vanishes with increasing data set)
  - my model is wrong (never goes away)
  - inference is inexact (vanishes with investing more computational resources)

Far better an approximate answer to the right question, than the exact answer to the wrong question.

– John W. Tukey

Assignment 8

No more assignments! But, I would appreciate your feedback on specific aspects of the course

See the Moodle page of the course