Week 13: Bayesian Computations (continued)

MATH-517 Statistical Computation and Visualization

Tomas Masak

December 16th 2022

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-Hasting: extremely versatile approach to MCMC
  - but a good proposal (leading to a good acceptance rate) can be hard to find for multidimensional problems

Gibbs Sampling

Idea: decompose the multidimensional distribution into *full conditionals* and draw from those in a cyclic manner

- not as universal as MH, since the calculation of the conditional distributions not always possible

Assuming

- $\mathcal{X} = \Pi_{i=1}^{p} \mathcal{X}_i$ (typically $\mathcal{X}_i = \mathbb{R}$ or $\mathcal{X}_i = (0, \infty)$),
- the target distribution belongs to some vector $X = (X_1, \ldots, X_p)^\top$ (notation),

the Gibbs sampler proceeds as follows from initial $x^{(0)} = (x_1^{(0)}, \ldots, x_p^{(0)})^\top$:

- **for** $t = 1, 2, \ldots$
  - generate $x_1^{(t)}$ from $X_1 \mid X_2 = x_2^{(t-1)}, X_3 = x_3^{(t-1)}, \ldots, X_p = x_p^{(t-1)}$
  - generate $x_2^{(t)}$ from $X_2 \mid X_1 = x_1^{(t)}, X_3 = x_3^{(t-1)}, \ldots, X_p = x_p^{(t-1)}$
  - $\ldots$
  - generate $x_p^{(t)}$ from $X_p \mid X_1 = x_1^{(t)}, X_2 = x_2^{(t)}, \ldots, X_{p-1} = x_{p-1}^{(t-1)}$

Remarks on Gibbs Sampling

- full conditionals specify the joint distribution
- it can be shown that the Gibbs sampler above actually samples from this joint distribution even though detailed balance does not hold
  - this is a fundamental result in random field theory
- the algorithm above is called *systematic Gibbs sampler*
  - with random reordering of indices in every step, it becomes the *random Gibbs sampler*
  - for the random Gibbs sampler, one often updates only a subset of indices in every step
  - in the special case of updating only one random index in a step, the random Gibbs sampler corresponds to MH (with proposal such that the acceptance prob. is 1) and hence detailed balance holds

Toy Example

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

$$X = (X_1, X_2)^\top \sim \mathcal{N}\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2\end{pmatrix}\right).$$

Easy, since Gaussian conditionals are Gaussian:

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

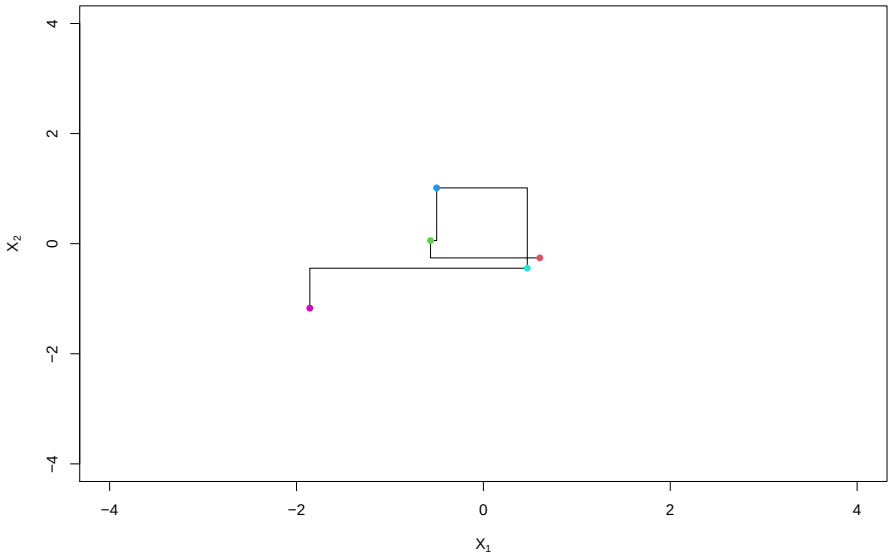
E.g. for $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$ and $\sigma_{12} = 0.5$, we have...

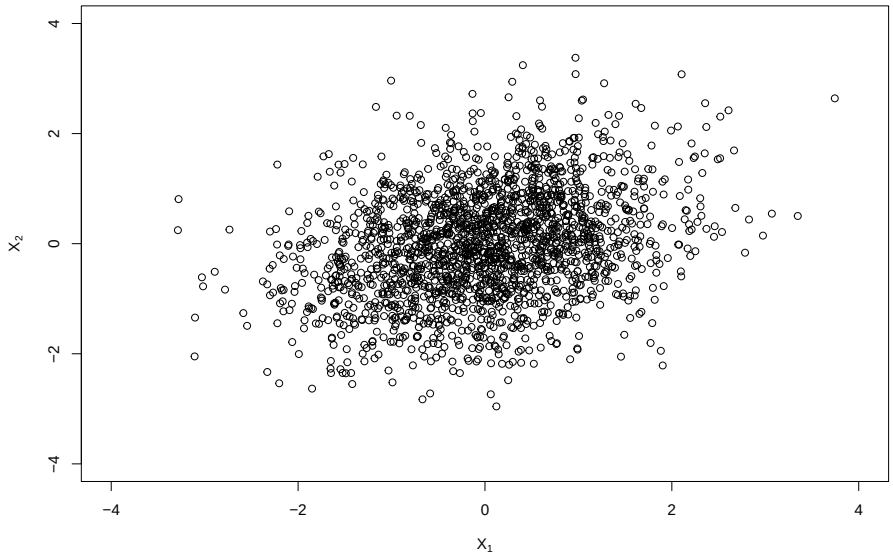
Toy Example

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

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

Toy Example

```
## [1] 0.298
```





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*
  - potentially via an MH step

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 (trace plots, ACF, etc.)
  - optionally, calculate some diagnostic statistics
- in reality, we can never know
  - silent failure?!

Simple but Real Example

- the height 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 observation fallin into $(62, 64]$ is $\Phi_{\mu,\sigma}(62) - \Phi_{\mu,\sigma}(60)$
- likelihood:

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

- prior: $f(\mu, \sigma) = \frac{1}{\sigma}$
  - improper and uninformative (though the latter is a misnomer)
  - changing variables $\lambda = \log(\sigma)$ removes $1/\sigma$ from the posterior

Posterior:

$$f(\mu, \sigma \mid D = d) \propto L(\mu, \exp(\lambda))$$

Real but Simple Example

- we want to sample from the posterior using random walk MH using the `LearnBayes` library, which uses

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

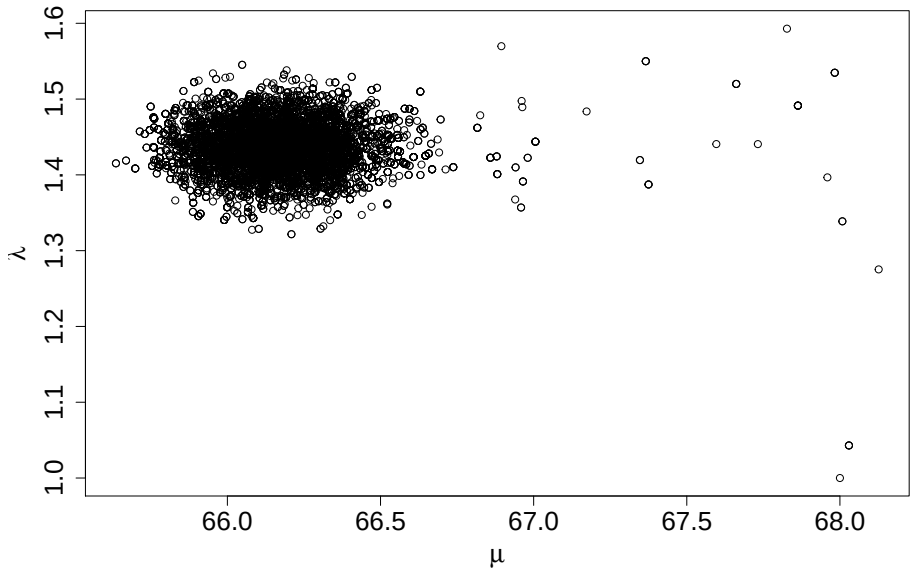
  where $Z \sim (0, \Sigma)$ and $s > 0$ is a scale parameter
  - overparametrization for the sake of convenience (debatable)
- for MH we have to choose
  - starting point $(\mu^{(0)}, \lambda^{(0)})^\top$
  - scale s
  - covariace Σ

Real but Simple Example

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

Acceptance rate:

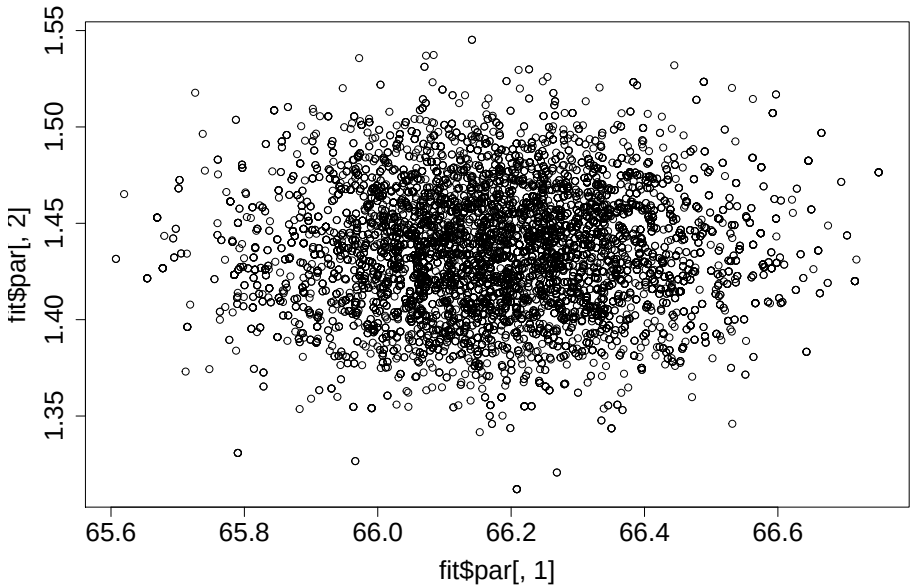
```
## [1] 0.3134
```



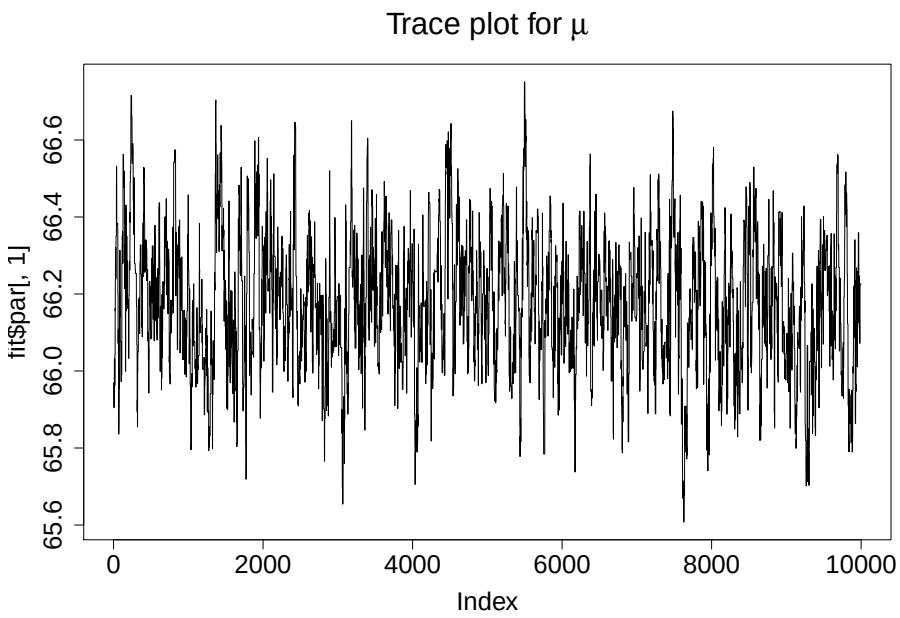
Real but Simple Example

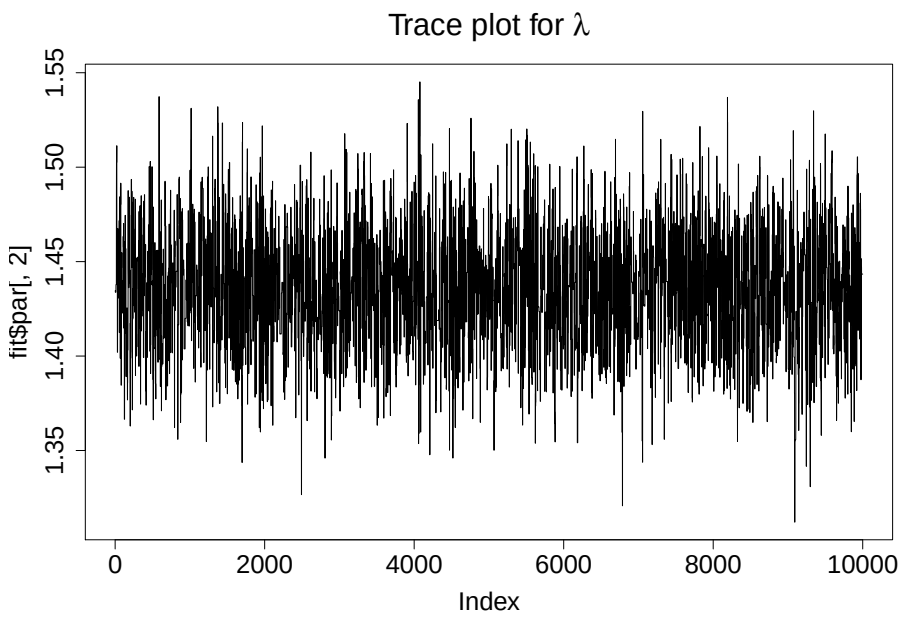
- above starting point chosen badly
- normally taken care of by burnin, here let's re-run
  - $(\mu^{(0)}, \lambda^{(0)})^\top = (66, 1.4)^\top$

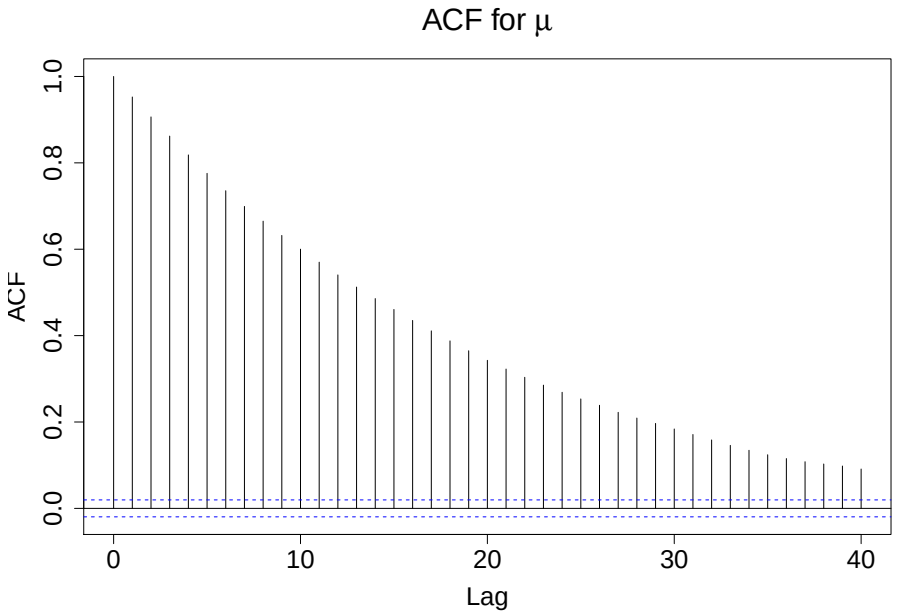
```
## [1] 0.3196
```

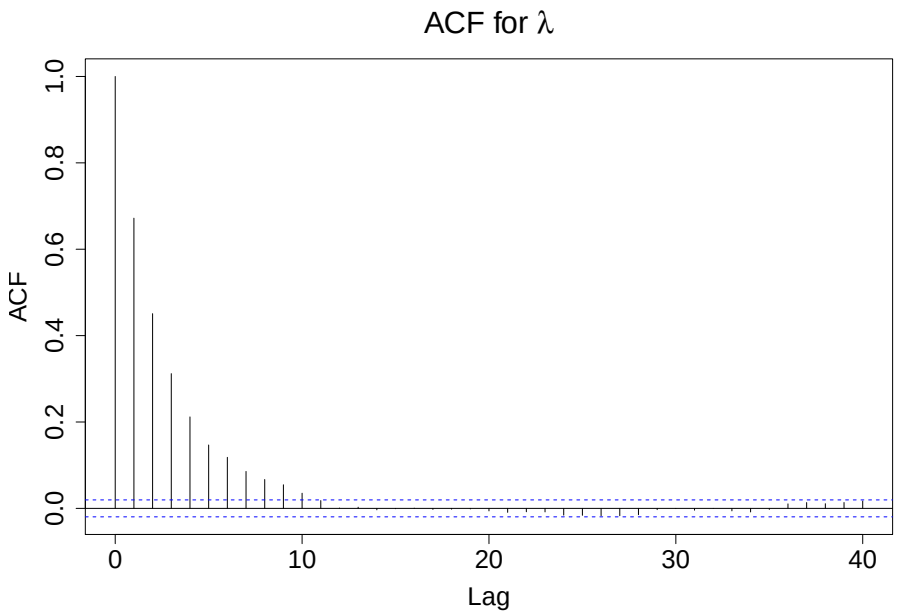


Real but Simple Example - Ouput Check









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 than for λ
- actually, why not to take Σ estimated from our previous run

```
var(fit$par)
```

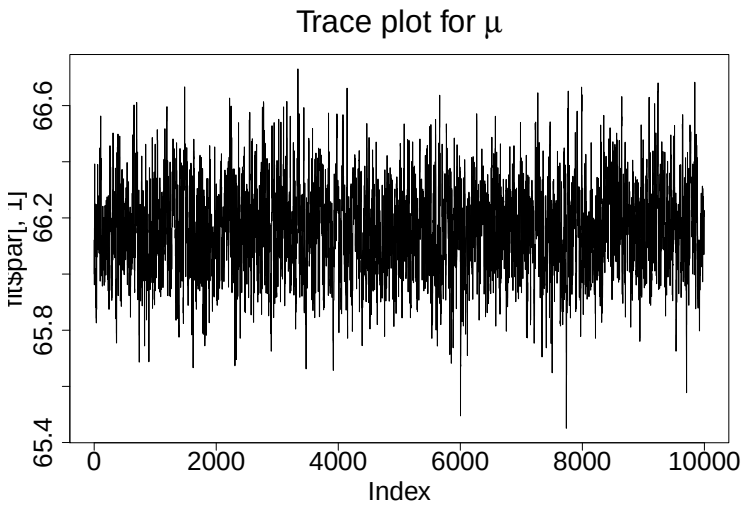
```
##                [,1]          [,2]
## [1,] 3.035891e-02 7.329492e-05
## [2,] 7.329492e-05 9.306388e-04
```

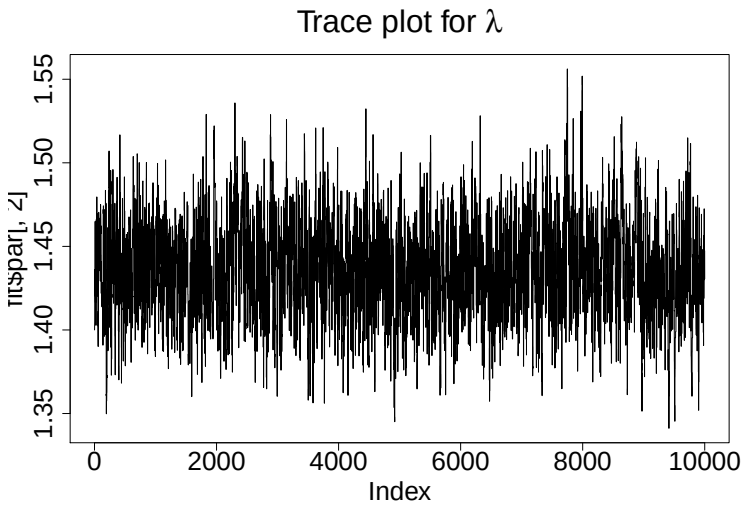
- acceptance too high with our $s = 0.1$ now, let's increase s
  - $s = 1$ gives 58%
  - let's take $s = 2$

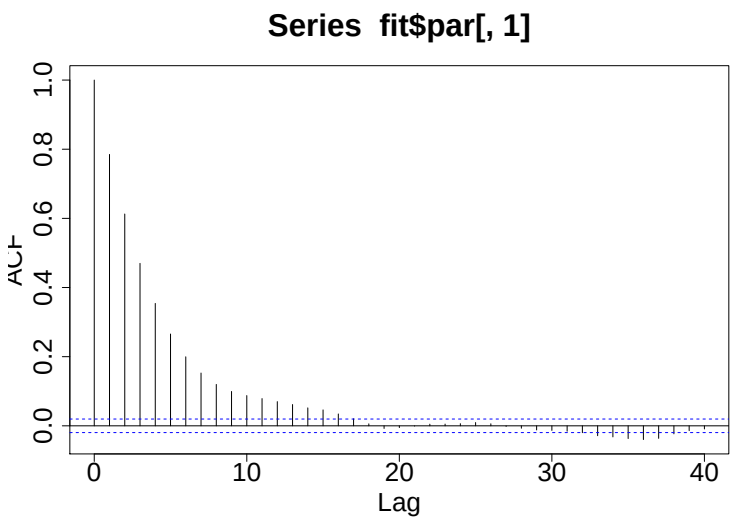
Real but Simple Example - Final Run

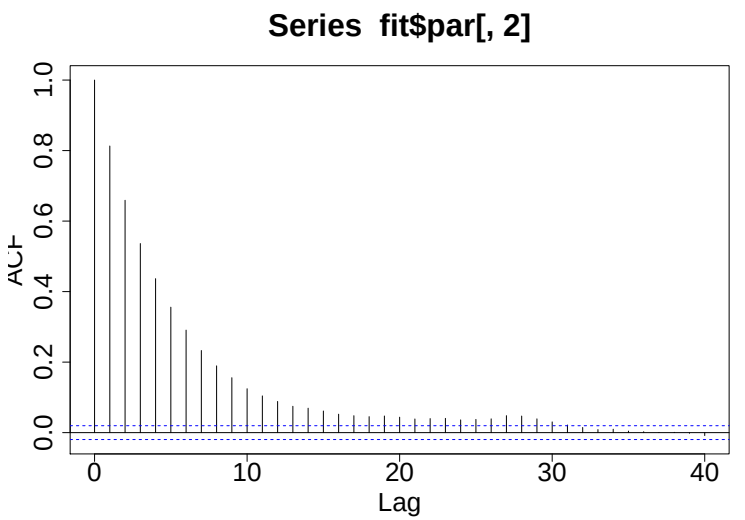
Lets analyze the output again.

```
## [1] 0.5386
```





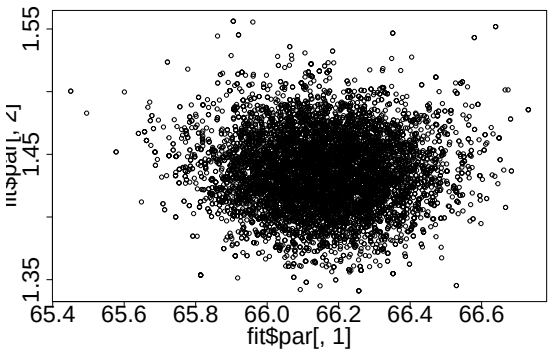


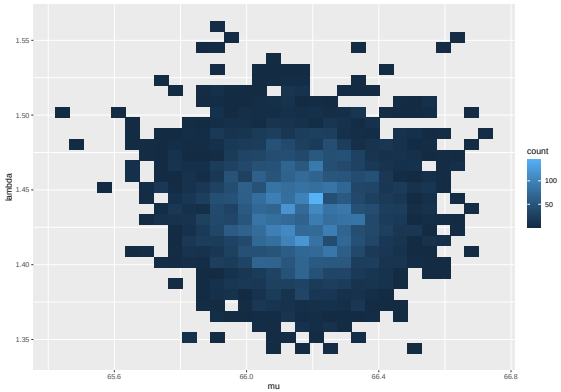


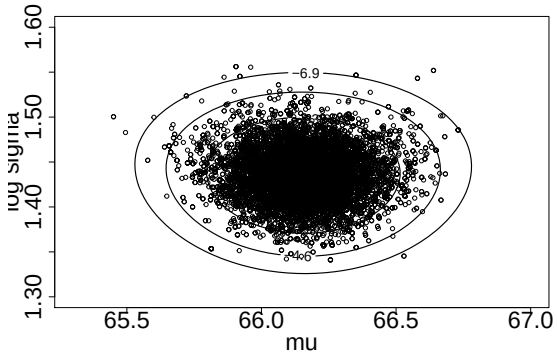
Real but Simple Example - Estimated Posterior

```
op <- par(ps=30)
plot(fit$par[,1],fit$par[,2])

sims <- data.frame(mu = fit$par[,1], lambda = fit$par[,2])
ggplot(sims, aes(x = mu, y = lambda)) +
  geom_bin2d()

mycontour(groupeddatapost,c(65.3,67,1.3,1.6),d,xlab="mu",ylab="log sigma")
points(fit$par[,1],fit$par[,2])
apply(fit$par,2,mean)
```

```
## [1] 66.158565  1.435971
```







Final Thoughts

- Bayesianism is a different way of thinking about problems
  - e.g. hierarchical models
- prior vs. no prior
- MLE vs. MAP (or rather posterior mean)
- 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 as a whole (like in MH) but endorse moving around
- 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 $\Rightarrow$ somewhat weak Bayesian theory

- 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

References

J Albert (2007) Bayesian Computations with R

C Robert & G Casella (2010) Introducing Monte Carlo Methods with R

K Murphy (2012) Machine Learning: a Probabilistic Perspective

Few things to mention if we have time at the end of this final lecture:

- studentized Bootstrap CI
- final projects
  - lessons learned
  - statisticians are plumbers
  - good work ethics
- feedback
  - EPFL, new course, first course
- Applied Stats course next semester
  - follow-up course
  - workflow established here
  - project-based (well-defined), methods-oriented