Math 390.03-02 / 650.03-01 Spring 2016
Final Examination

Professor Adam Kapelner

Wednesday, May 25, 2016

Full Name _____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using a cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

_____ _____

signature date

Instructions

This exam is 120 minutes and closed-book. You are allowed three pages (front and back) of "cheat sheets." You may use a graphing calculator of your choice. Please read the questions carefully. If the question reads "compute," this means the solution will be a number otherwise you can leave the answer in *any* widely accepted mathematical notation which could be resolved to an exact or approximate number with the use of a computer. I advise you to skip problems marked "[Extra Credit]" until you have finished the other questions on the exam, then loop back and plug in all the holes. I also advise you to use pencil. The exam is 100 points total plus extra credit. Partial credit will be granted for incomplete answers on most of the questions. $\boxed{\text{Box}}$ in your final answers. Good luck!

This exam assumes you will be able to compute some answers numerically given a computer. Thus, you can leave your answer in terms of any of the following functions as long as the question does not say "compute explicitly." However, you must make clear the numerical values that are parameters for these functions.

Distribution	Quantile Function	PMF / PDF function	CDF function	Sampling Function
beta	`qbeta`(p, α, β)	`d-`(x, α, β)	`p-`(x, α, β)	`r-`(α, β)
betabinomial	`qbetabinom`(p, n, α, β)	`d-`(x, n, α, β)	`p-`(x, n, α, β)	`r-`(n, α, β)
betanegativebinomial	`qbeta_nbinom`(p, r, α, β)	`d-`(x, r, α, β)	`p-`(x, r, α, β)	`r-`(r, α, β)
binomial	`qbinom`(p, n, θ)	`d-`(x, n, θ)	`p-`(x, n, θ)	`r-`(n, θ)
exponential	`qexp`(p, θ)	`d-`(x, θ)	`p-`(x, θ)	`r-`(θ)
gamma	`qgamma`(p, α, β)	`d-`(x, α, β)	`p-`(x, α, β)	`r-`(α, β)
geometric	`qgeom`(p, θ)	`d-`(x, θ)	`p-`(x, θ)	`r-`(θ)
inversegamma	`qinvgamma`(p, α, β)	`d-`(x, α, β)	`p-`(x, α, β)	`r-`(α, β)
negative-binomial	`qnbinom`(p, r, θ)	`d-`(x, r, θ)	`p-`(x, r, θ)	`r-`(r, θ)
normal (univariate)	`qnorm`(p, θ, σ)	`d-`(x, θ, σ)	`p-`(x, θ, σ)	`r-`(θ, σ)
normal (multivariate)		`dmvnorm`($\boldsymbol{x}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$)		`r-`($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$)
poisson	`qpois`(p, θ)	`d-`(x, θ)	`p-`(x, θ)	`r-`(θ)
T (standard)	`qt`(p, ν)	`d-`(x, ν)	`p-`(x, ν)	`r-`(ν)
T (nonstandard)	`qt.scaled`(p, ν, μ, σ)	`d-`(x, ν, μ, σ)	`p-`(x, ν, μ, σ)	`r-`(ν, μ, σ)
uniform	`qunif`(p, a, b)	`d-`(x, a, b)	`p-`(x, a, b)	`r-`(a, b)

Table 1: Functions from `R` (in alphabetical order) that can be used on this exam. The hyphen in colums 3, 4 and 5 is shorthand notation for the full text of the r.v. which can be found in column 2.

- The quantile function finds the minimum x which satisfies $\sum_{x_0=-\infty}^{x} p(x_0) \geq p$ if the r.v. is discrete or solves for x in the following equation: $\int_{-\infty}^{x} f(x)dx = p$ if the r.v. is continuous.
- The PMF / PDF function computes $p(x)$ if the r.v. is discrete or $f(x)$ if the r.v. is continuous.
- The CDF function calculates the following sum: $\sum_{x_0=-\infty}^{x} p(x_0)$ if the r.v. is discrete or it calculates the following integral $\int_{-\infty}^{x} f(x)dx$ if the r.v. is continuous.
- The sampling function will draw a random realization from the distribution of interest.

All other parameters are the function parameters in the density / PMF that you should be familiar with from the class notes (and your cheat sheet)

Problem 1 This question is about ratings on youtube. Each video which is voted on is either up-voted or down-voted. A video's rating looks like the following:

(ignore the number of views). This video for instance has a 5800/(5800 + 960) = 85.8% approval rating out of 6,760 total votes.

But there is a question: how should we order videos by approval rating? For example, here is a table of four videos we wish to order about a topic we are interested in:

Video Name	# Up votes	# Down votes	n	Approval Rating
A	5080	960	6040	84.1%
B	3	0	3	100.0%
C	25	1	26	96.2%
D	0	1	1	0.0%

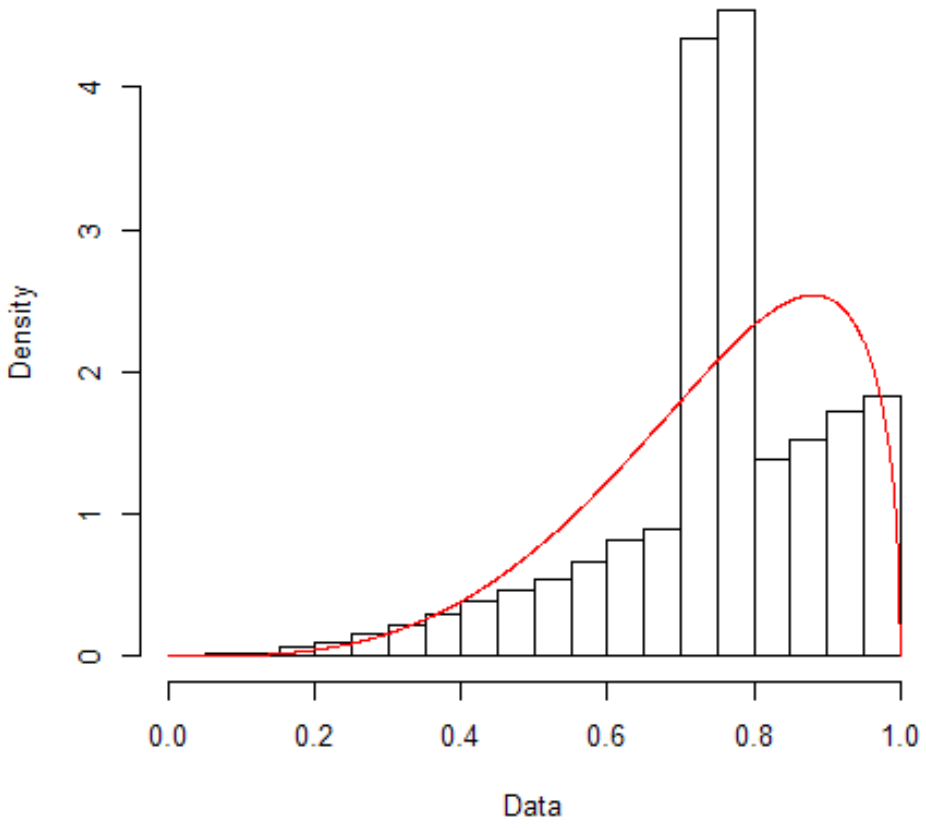
Table 2: Table of videos with their youtube ratings.

(a) [2 pt / 2 pts] Order the movies by name from best to worst using the MLE estimate of its true approval rating.

(b) [3 pt / 5 pts] Why is what you did in (a) a poor way to order the four movies?

(c) [1 pt / 6 pts] We are now going to use some previous data to create a prior for the true approval rating. What is this kind of procedure is called? Two words.

Below is a histogram of the approval ratings of $n_0 = 30,000$ videos of which there are more than 200 votes each. The curve displayed atop the histogram is the best fit beta density. I used `R`'s `fitdistrplus` package which creates a fit via the MLE's of α and β. I include estimates in output from `R` below the plot.




```
Parameters:
       estimate Std. Error
shape1 4.283762 0.03567291
shape2 1.442157 0.01073980
```

Despite the fact that this fit does not seem to be appropriate, we will use it for now.

(d) [4 pt / 10 pts] Besides the fact that the curve does not fit the empirical distribution (given by the histogram), what is wrong with the estimates of α and β given above? Hint: think about pseudocounts.

(e) [3 pt / 13 pts] Given that a movie has n total votes and x of those are thumbs up, what is the distribution of the true approval rating θ given the data coupled with the prior constructed above question (d)?

(f) [5 pt / 18 pts] Order the movies from best to worst using the Bayesian estimate which minimizes mean squared error. Compute explicitly. No credit unless work is shown.

(g) [1 pt / 19 pts] We will now attempt to improve the model by improving the prior by modeling the prior as a sum of two beta distributions. What kind of model would this be called? Two words.

(h) [3 pt / 22 pts] The likelihood function of the two-beta model is below. Remember there are n_0 data points in the prior data which I've denoted $y_1, \ldots, y_{n_0}$. Remember that n which denotes the sample size of a single video which we are not estimating yet.

$$\mathbb{P}\left(Y_1, \ldots, Y_{n_0} \mid \alpha_1, \beta_1, \alpha_2, \beta_2, \rho\right) = \prod_{i=1}^{n_0} \rho f_1(y_i) + (1-\rho) f_2(y_i)$$

$$= \prod_{i=1}^{n_0} \rho \left(\frac{1}{B(\alpha_1, \beta_1)} y_i^{\alpha_1 - 1} (1-y_i)^{\beta_1 - 1} \right) + (1-\rho) \left(\frac{1}{B(\alpha_2, \beta_2)} y_i^{\alpha_2 - 1} (1-y_i)^{\beta_2 - 1} \right)$$

Why would estimating $\alpha_1, \beta_1, \alpha_2, \beta_2, \rho$ be difficult if you were to use the MLE method?

(i) [4 pt / 26 pts] We will instead use the E-M algorithm. Write the likelihood now of the data-augmented model where the I_i's are indicators for the ith observation belonging to the f_1 distribution.

$\mathbb{P}\left(Y_1, \ldots, Y_{n_0}, I_1, \ldots, I_{n_0} \mid \alpha_1, \beta_1, \alpha_2, \beta_2, \rho\right) =$

(j) [4 pt / 30 pts] In the E step, write an expression for $\hat{I}_i$ Use hat symbols (the caret on top) to denote estimates. Do not compute.

(k) [4 pt / 34 pts] In the M step, write an expression for $\hat{\rho}_{MLE}$. Use hat symbols (the caret on top) to denote estimates. Do not compute.

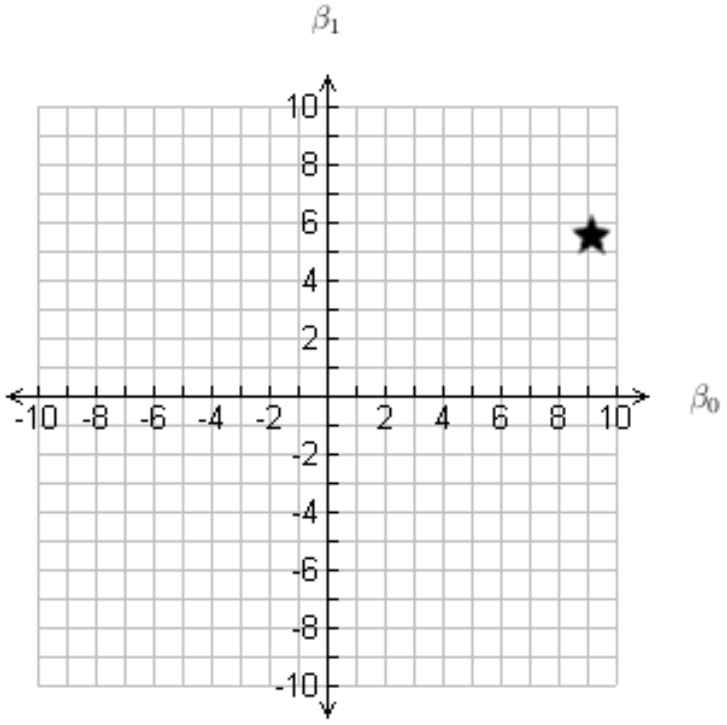
(l) [8 pt / 42 pts] Luckily, the `R` package `betareg` can run the E-M algorithm for us until it converges. In 33 iterations we got the following:

$$\alpha_1 = 2.97,\ \beta_1 = 1.00,\ \alpha_2 = 751.27,\ \beta_2 = 250.44 \text{ and } \rho = 0.666.$$

Find the Bayesian estimate for video B which minimizes mean squared error for the prior given by these estimates from E-M above. No need to compute explicitly.

(m) [5 pt / 47 pts] [Extra credit] Video B gets 100 more votes. Without looking at whether they are thumbs up or thumbs down, what is the probability that 65 of them or less are thumbs up? Do not compute explicitly. If you can do this, you may want to work for Google.

Problem 2 This question is about ridge regression. Given n samples from a bivariate distribution, a best fit line was estimated with intercept b_0 and slope b_1. This estimate is shown as a star on the graph below. The axes represent the entire parameter space $\text{Supp}[\beta_0] \times \text{Supp}[\beta_1]$.



(a) [3 pt / 50 pts] Imagine a ridge regression was run with ridge penalty m being very large. Mark the graph above with the letter "a" with one place where you think this ridge estimate could be.

(b) [3 pt / 53 pts] Imagine a ridge regression was run with ridge penalty m being very small. Mark the graph above with the letter "b" with one place where you think this ridge estimate could be.

(c) [3 pt / 56 pts] Imagine a ridge regression was run with ridge penalty $m = 1$. Mark the graph above with the letter "c" with one place where you think this ridge estimate could be.

Problem 3 We now build a toy Bayesian model where the data points are normal:

$$X_1, \ldots, X_n \overset{exch}{\sim} \mathcal{N}\left(\theta, \sigma^2\right)$$

and we have the standard conjugate prior on the mean:

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

but we have the following prior on the variance:

$$\sigma^2 \sim \text{Gamma}\left(\frac{n_0}{2}, \beta\right)$$

(note that this is a gamma and not an inverse gamma and the gamma has the same support).

(a) [5 pt / 61 pts] Write the posterior exactly below. I have already included the denominator so you do not have to specify that. You only have to specify the numerator. Note: this is an equals sign and not a proportionality sign. You will do the proportionality in the next problem. Hint: I have left it in three pieces for a reason.

$$\mathbb{P}\left(\theta, \sigma^2 \mid X_1, \ldots, X_n\right) = \frac{1}{\mathbb{P}\left(X_1, \ldots, X_n\right)} \times \left(\prod_{i=1}^{n} \right) \times \left(\right) \times \left(\right)$$

(b) [4 pt / 65 pts] Write the kernel of the posterior below. Use your answer from (a). Collect like terms and simplify.

$$\mathbb{P}\left(\theta, \sigma^2 \mid X_1, \ldots, X_n\right) \propto$$

(c) [2 pt / 67 pts] Is the kernel proportional to a kernel from a known random variable which can be sampled? (yes / no)

(d) [6 pt / 73 pts] Explain in English how you would use grid sampling to sample n draws from the posterior in (a). Say "Step 1," "Step 2," etc. Do this problem last!

(e) [3 pt / 76 pts] Despite what you wrote in (d), we are going to attempt to use MCMC to sample from the posterior in (a). Write the kernel of the conditional distribution of θ given σ^2 and the data below. Use your answer from (b). Then show that it is proportional to a known random variable which can be sampled (make sure to give its parameters).

$\mathbb{P}\left(\theta \mid X_1, \ldots, X_n,\ \sigma^2\right) \propto$

(f) [5 pt / 81 pts] Write the kernel of the conditional distribution of σ^2 given θ and the data below. Use your answer from (b). You can use the notation from class $n\hat{\sigma}^2 := \sum_{i=1}^{n}(x_i - \theta)^2$

$\mathbb{P}\left(\sigma^2 \mid X_1, \ldots, X_n,\ \theta\right) \propto$

(g) [2 pt / 83 pts] Is the kernel proportional to a kernel from a known random variable which can be sampled? (yes / no) Can you use a Gibbs iteration to sample $\mathbb{P}\left(\sigma^2 \mid X_1, \ldots, X_n,\ \theta\right)$? (yes / no)

(h) [2 pt / 85 pts] We will now put together (e) and (f) to create a Metropolis-within-Gibbs sampler. We have $n = 50$ samples from the true data generating process. We will use a relatively uninformative prior on θ which is $\mu_0 = 0$ and $m = 1$. For the gamma prior, we will use an "uninformative" $n_0 = 1$ and $\beta = 1/2$. We will run it for 10,000 iterations. Figure 1 shows the first 2,000 iterations. At about what iteration number t did the $\mathbb{P}\left(\theta \mid X_1, \ldots, X_n,\ \sigma^2\right)$ chain converge?

(i) [2 pt / 87 pts] At about what iteration number t did the $\mathbb{P}\left(\sigma^2 \mid X_1, \ldots, X_n,\ \theta\right)$ chain converge?

(j) [1 pt / 88 pts] Using your answers from (h) and (i), what do you think the burn-in B iteration should be going forward?

(k) [2 pt / 90 pts] Figure 2 shows the autocorrelation estimates. At what iteration t would you first be able to thin the $\mathbb{P}\left(\theta \mid X_1, \ldots, X_n,\ \sigma^2\right)$ chain?

(l) [2 pt / 92 pts] Figure 2 shows the autocorrelation estimates. At what iteration t would you first be able to thin the $\mathbb{P}\left(\sigma^2 \mid X_1, \ldots, X_n,\ \theta\right)$ chain?

(m) [1 pt / 93 pts] Using your answers from (k) and (l), what do you think the thin-mod value T should be going forward?

(n) [5 pt / 98 pts] Figure 3 shows samples from the posterior after the chains were burned and thinned. What integral does the bottom histogram approximate?

(o) [3 pt / 101 pts] Estimate the posterior expectations, $\mathbb{E}[\theta \mid X_1, \ldots, X_n]$ as well as $\mathbb{E}[\sigma^2 \mid X_1, \ldots, X_n]$ using Figure 3.

(p) [2 pt / 103 pts] Use Figure 3 to estimate a 95% credible region for θ.

(q) [2 pt / 105 pts] Use Figure 3 to accept or reject $H_0 : \theta \leq 12.5$ by estimating the p-value.

(r) [5 pt / 110 pts] [Extra Credit] What is the probability that 10 new observations will have an average greater than 14? Write an algorithm below that would approximate this. You can reference figure 3 in your answer.

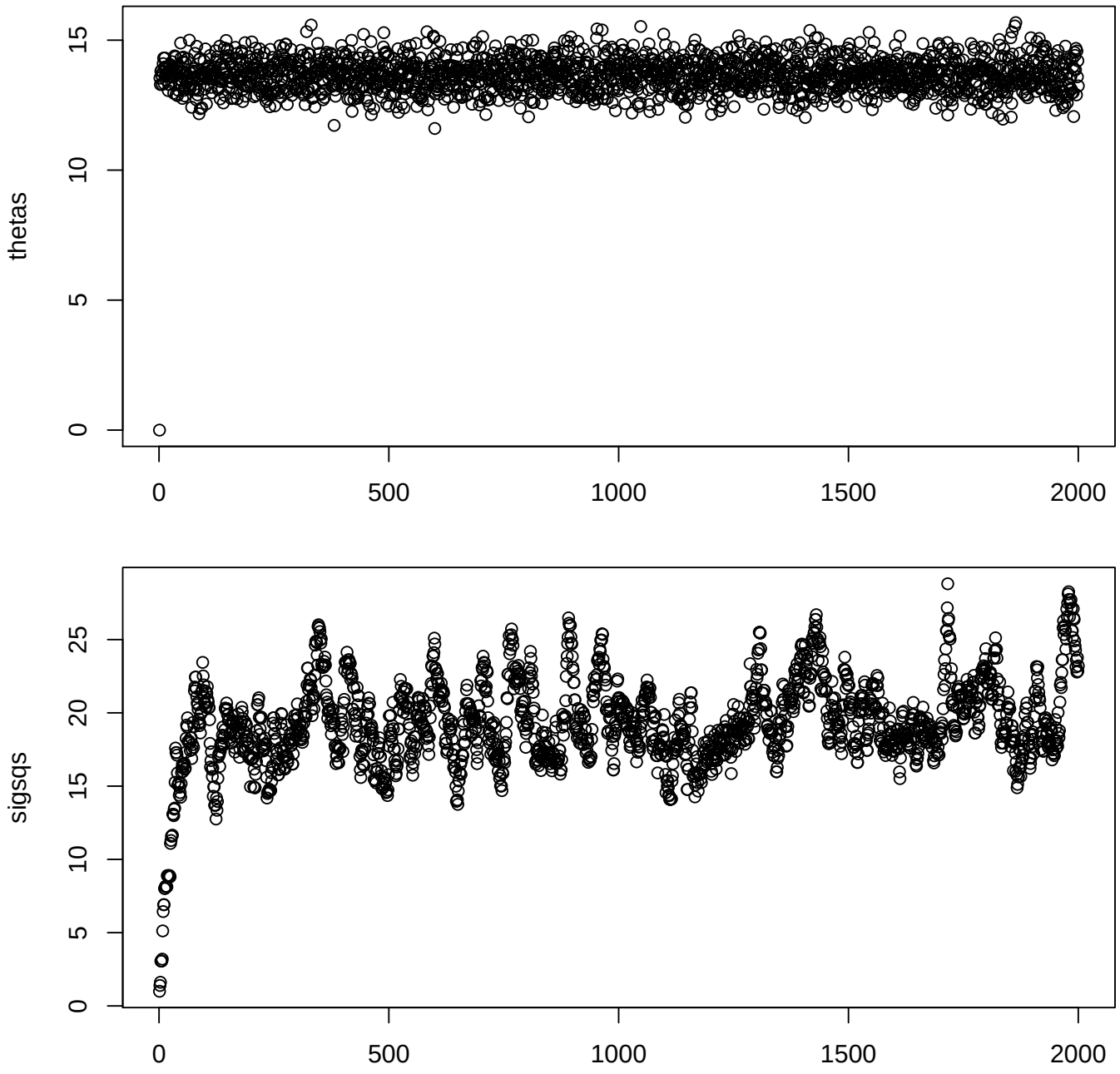


Figure 1: First 2,000 iterations of the Metropolis-within-Gibbs sampler for this data.

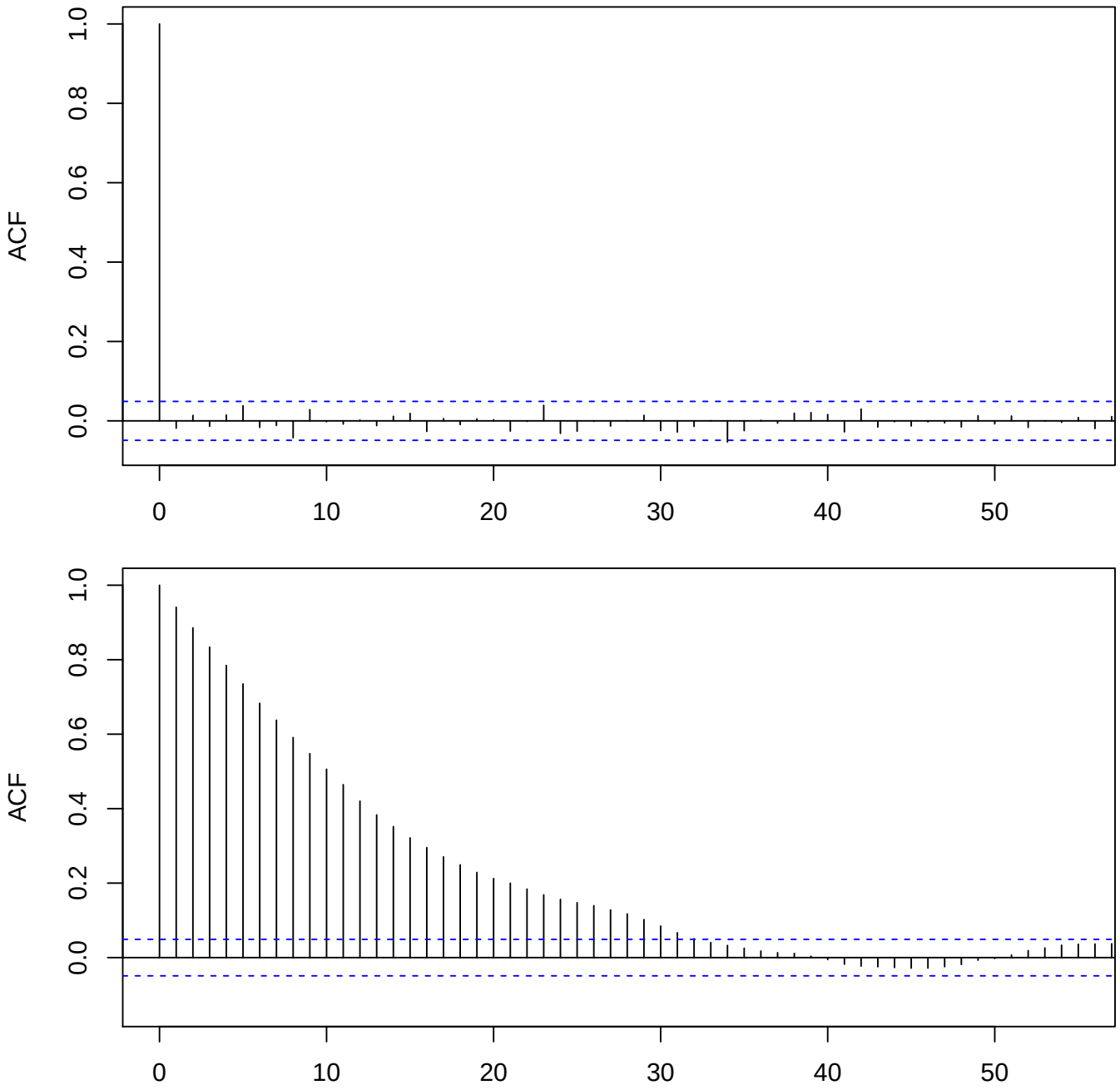


Figure 2: Autocorrelation plots of the Metropolis-within-Gibbs sampler for this data.

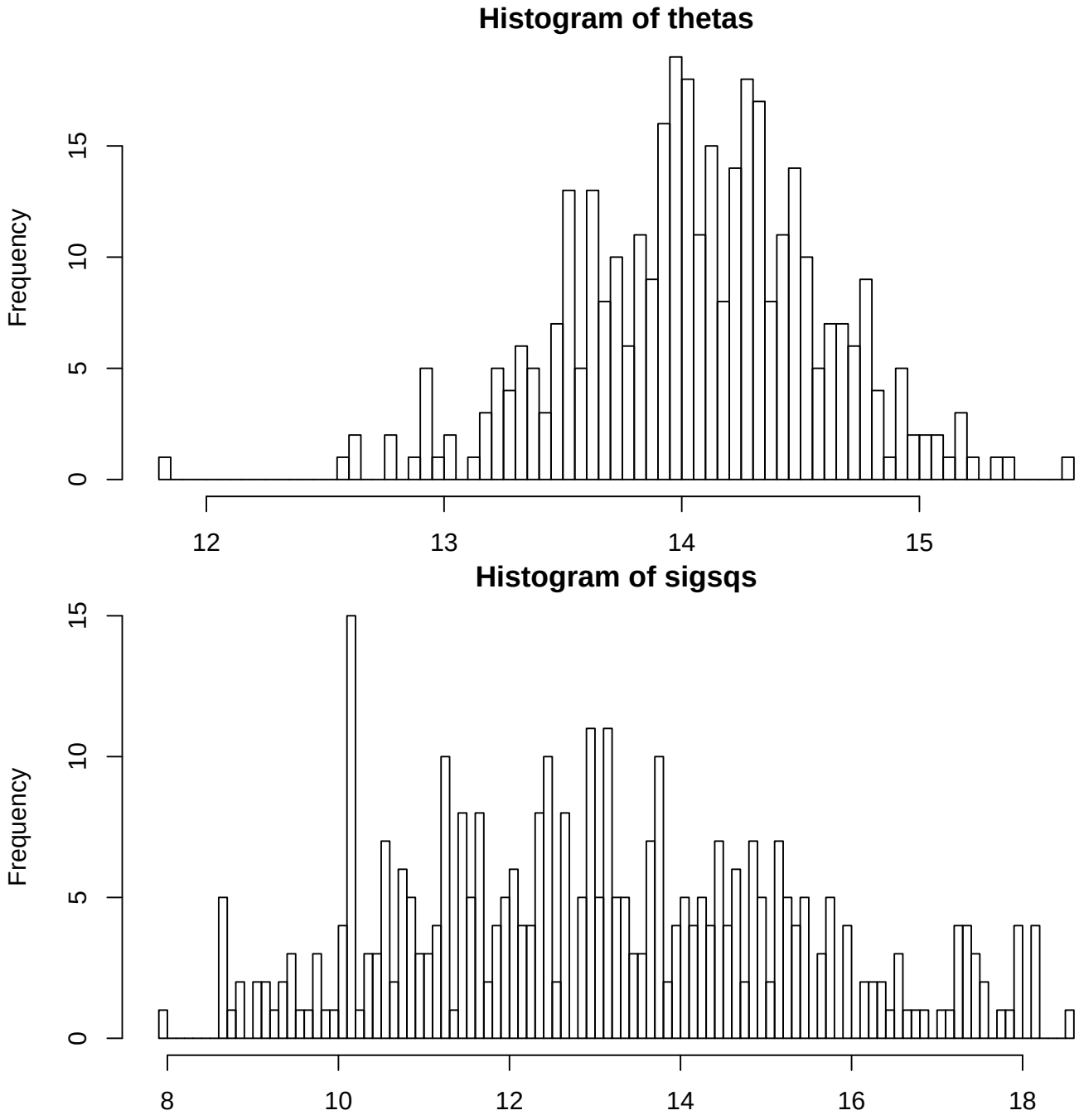


Figure 3: Histograms of the Metropolis-within-Gibbs sampler for this data.