Assignment 8

Sam Olson

Problem Description

Consider a problem of conducting a Bayesian analysis with a one-sample gamma model.

Assume that random variables $Y_1, \ldots, Y_n$ are independent and identically distributed with common probability density function (for $\alpha > 0$ and $\beta > 0$):

$$f(y \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y), \quad y > 0.$$

Suppose that we will assign a joint prior to α and β in product form, with particular values of $A > 0$, $\gamma_0 > 0$, and $\lambda_0 > 0$:

$$\pi_\alpha(\alpha) = \frac{1}{A} I(0 < \alpha < A), \quad \pi_\beta(\beta) = \frac{\lambda_0^{\gamma_0}}{\Gamma(\gamma_0)} \beta^{\gamma_0 - 1} e^{-\lambda_0 \beta}, \quad \beta > 0.$$

Recall that in the analysis of an actual data set, A, γ_0 and λ_0 will be given specific numerical values. Since this is a simulated example and we have no actual prior information, use the following hyperparameters:

$$A = 20, \quad \gamma_0 = 0.5, \quad \lambda_0 = 0.1.$$

This gives prior expectation of 5.0 and prior variance of 50. The prior does focus probability on smaller values, but still has $Pr(\beta > 10) = 0.16$.

1.

Consider using a Metropolis–Hastings algorithm with independent random-walk proposals for α and β.

Suppose that our current values are (α_m, β_m), and that the proposal (α^*, β^*) has been generated from

$$q(\alpha, \beta \mid \alpha_m, \beta_m)$$

which is the product of independent random walks.

Identify the appropriate acceptance probability for the jump proposal (α^*, β^*).

Answer

I believe this question is being asked in the abstract, i.e., not for the specific dataset in question. Under that pretense:

The target distribution for the Metropolis–Hastings algorithm is the joint posterior:

$$\pi(\alpha, \beta \mid y) \propto L(y \mid \alpha, \beta)\,\pi_\alpha(\alpha)\,\pi_\beta(\beta)$$

where L denotes the likelihood (not log-transformed), and with the likelihood for independent $Y_i \sim$ Gamma(α, β) given by:

$$L(y \mid \alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta y_i}$$

Suppose the proposal distribution is a product of independent symmetric random walks, such that:

$$q(\alpha^*, \beta^* \mid \alpha_m, \beta_m) = q_\alpha(\alpha^* \mid \alpha_m) q_\beta(\beta^* \mid \beta_m)$$

Because each random-walk is symmetric, the proposal satisfies:

$$q(\alpha^*, \beta^* \mid \alpha_m, \beta_m) = q_\alpha(\alpha^* \mid \alpha_m) q_\beta(\beta^* \mid \beta_m)$$

And the ratio:

$$\frac{q(\alpha_m, \beta_m \mid \alpha^*, \beta^*)}{q(\alpha^*, \beta^* \mid \alpha_m, \beta_m)} = 1$$

Consequently, the Hastings term"cancels and the acceptance probability depends only on the posterior ratio, specifically:

$$a\big((\alpha_m, \beta_m) \to (\alpha^*, \beta^*)\big) = \min\left\{1, \frac{p(\alpha^*, \beta^* \mid y)}{p(\alpha_m, \beta_m \mid y)}\right\} = \min\left\{1, \frac{L(y \mid \alpha^*, \beta^*)\,\pi_\alpha(\alpha^*)\,\pi_\beta(\beta^*)}{L(y \mid \alpha_m, \beta_m)\,\pi_\alpha(\alpha_m)\,\pi_\beta(\beta_m)}\right\}$$

Because $\pi_\alpha(\alpha)$ is uniform on $(0, A)$, this implies that any proposal with $\alpha^* \notin (0, A)$ or $\beta^* \leq 0$ is automatically rejected $(a = 0)$.

Knowing the data model comes from a one-sample Gamma, we know it comes from the exponential dispersion family. This is used primarily to motivate knowing the form of the sufficient statistics, $S_1 = \sum_{i=1}^{n} \log y_i$ and $S_2 = \sum_{i=1}^{n} y_i$. With this, we can write (taking the log of the posterior and then exponentiating at the end to get the original scale):

$$\log r = n\left[\alpha^* \log \beta^* - \log \Gamma(\alpha^*) - \alpha_m \log \beta_m + \log \Gamma(\alpha_m)\right] + (\alpha^* - \alpha_m)S_1 - (\beta^* - \beta_m)S_2 + (\gamma_0 - 1)\left[\log \beta^* - \log \beta_m\right] - \lambda_0(\beta^* - \beta_m)$$

Such that we have the acceptance probability for the Metropolis–Hastings update of (α, β) (exact) as:

$$a = \min\left\{1, \ \exp(\log r)\right\}$$

for a joint update of (α, β) under independent symmetric random-walk proposals.

2.

On the course web page is a data set called `gammadat_bayes.txt`.

Program a Metropolis–Hastings algorithm and simulate 50,000 values from the joint posterior of α, β, and $\mu = \alpha/\beta$.

Provide (with supporting evidence if appropriate):

- Information on how you selected a burn-in period. *NOTE: I do not expect you to compute Gelman-Rubin scale reduction factors for this assignment.*

- Information on how you tuned the algorithm for acceptance rate, including the random-walk variances and the final acceptance rate.

- Summaries of the marginal posterior distributions of α and β and $\mu = \alpha/\beta$, including histograms and five-number summaries, 95% central credible intervals, and correlation between α and β in the Markov chain.

Answer

To start, while not affecting the priors (in distribution or parameter value), we do want to look at the data to get our "initial" values.

Based on method of moments estimators, we have:

$\alpha_0 = \frac{\bar{y}^2}{s_y^2}$, $\beta_0 = \frac{\bar{y}}{s_y^2}$ (for shape-rate parametrization).

Where $\bar{y}$ and s_y^2 are the sample mean and variance, respectively (such that $E[Y] = \frac{\alpha}{\beta}$ and $V[Y] = \frac{\alpha}{\beta^2}$).

For both α_0 and β_0, we will "impose" that any estimates or updates retain $\alpha_0, \beta_0 > 0$.

Additionally, given the hyperparameters to this problem, we also "impose" that $\alpha_0 < 20$.

We run a sampling procedure without any tuning, first to determine a suitable Burn-In period.

Note: Throughout the sampling procedure, we simulate α and β simultaneously, which is important to keep later estimates of μ more comparable ("apples-to-apples"). Furthermore, the actual sampling is done using log likelihood, again, for the purpose of stability. However, we built-in a check for "invalid" parameter proposals (negative-valued); during the procedure, after identifying a "candidate", if the proposed parameter is negative-valued, it is "reverted" to the prior value. This ensures appropriateness of the method.

We also consider a handful of starting values for α, β to ensure the trace plots achieve stability around the same number of iterations.
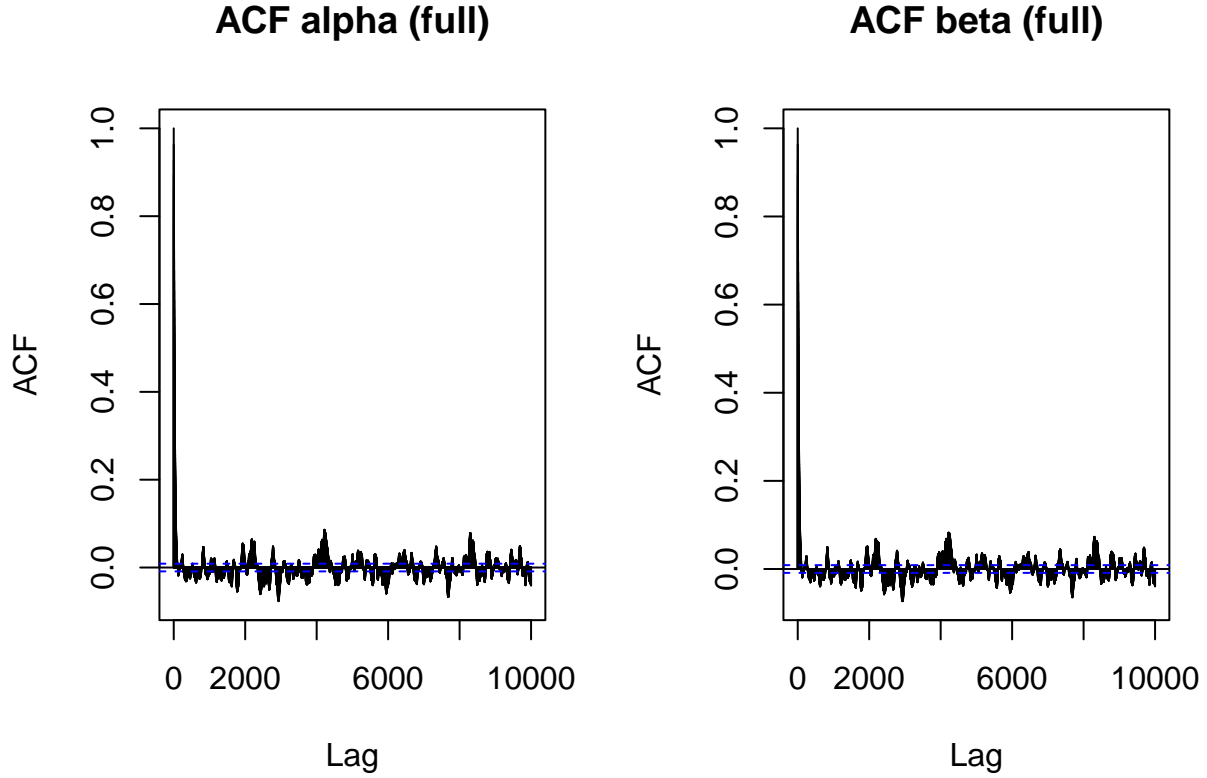
Our sample values provide a basis for determining our Burn-In Period. We use our Metropolis-Hastings algorithm for a rather large number (50,000) without *any* Burn-In, and then look at both the trace plots and autocorrelation plots (ACFs). With that, we have:

alpha trace (B=0)

beta trace (B=0)

iter

iter

ACF alpha (full)   **ACF beta (full)**
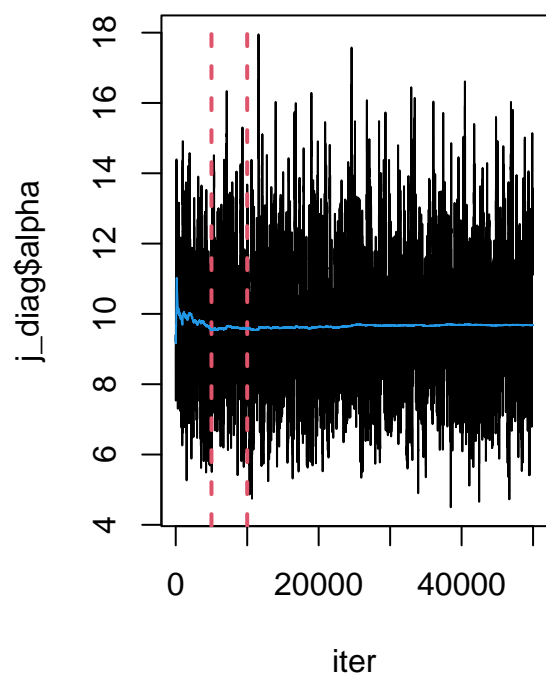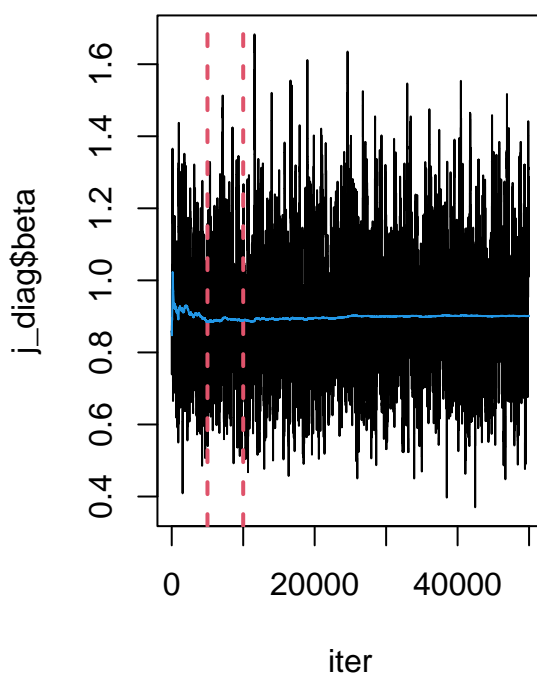
To that end, we use a "running mean" (blue) to get a sense of when variability in the parameter "stabilizes" in the Trace Plot graphs. We see that by iteration 5,000, α and β samples tend to stabilize. Additionally, the ACFs decay rapidly (well before the 5,000); to stay on the conservative side, we double this and set that equal to our Burn-In period.

Ultimately, we settle on a Burn-In period of 10,000.

Note: Although the ACF plot displays values that remain slightly outside the "white-noise" bounds around the proposed Burn-In period, the autocorrelation clearly decays toward zero and alternates in sign, indicating that dependence weakens over time. Therefore, it is still reasonable to suppose the proposed Burn-In period is appropriate.

We then have some additional tuning. The general procedure is to "scale" the jump variance by some proportion of the original sample variance. To begin with, we settle on multipling the variance by some proportion, initially around 0.25 to 2, and then after iterating deciding on 0.5, 0.75, 1.0, 1.25, and 1.5. The acceptance probabilities for each of these scales are reported, using the Burn-In of 10,000.

scale	valpha	vbeta	accept
0.50	0.8473	0.0073	0.4264
0.75	1.9063	0.0165	0.2917
1.00	3.3891	0.0293	0.2166
1.25	5.2954	0.0458	0.1677
1.50	7.6254	0.0660	0.1338

Our goal in tuning is to reach a Metropolis-Hastings acceptance rate somewhere in the range of 20-60%, which is a heuristic noted in the Chapter 7 notes on Simulation. Ultimately, we decided on using a "scale" of

1.25, which given the seed and method noted, gives us an acceptance probability around 16.5%. Note: While this is on the lower end, I ultimately wanted to scale the jump variance somewhat higher than the sample to allow for greater exploration of a wider range of possible values for the "jumps".

This then leads us to run the whole simulation procedure again, now using both the Burn-In period of 5,000 and the tuned parameters.

Table 2: (MH) Posterior summaries: five-number statistics and 95 central credible intervals

Method	Parameter	Posterior Summary						
		Min	Q1	Median	Q3	Max	CI.2.5	CI.97.5
MH	alpha	4.4454	8.2681	9.3900	10.6457	17.4904	6.3353	13.3748
MH	beta	0.3904	0.7641	0.8763	0.9926	1.6698	0.5804	1.2527
MH	mu	8.8090	10.4279	10.7508	11.0919	13.2251	9.8048	11.8102

Table 3: (MH) Acceptance rates

Component	Acceptance.Rate
Overall	16.8%

Table 4: (MH) Correlation matrix for alpha and beta

Parameter	alpha	beta
alpha	1.0000	0.9713
beta	0.9713	1.0000
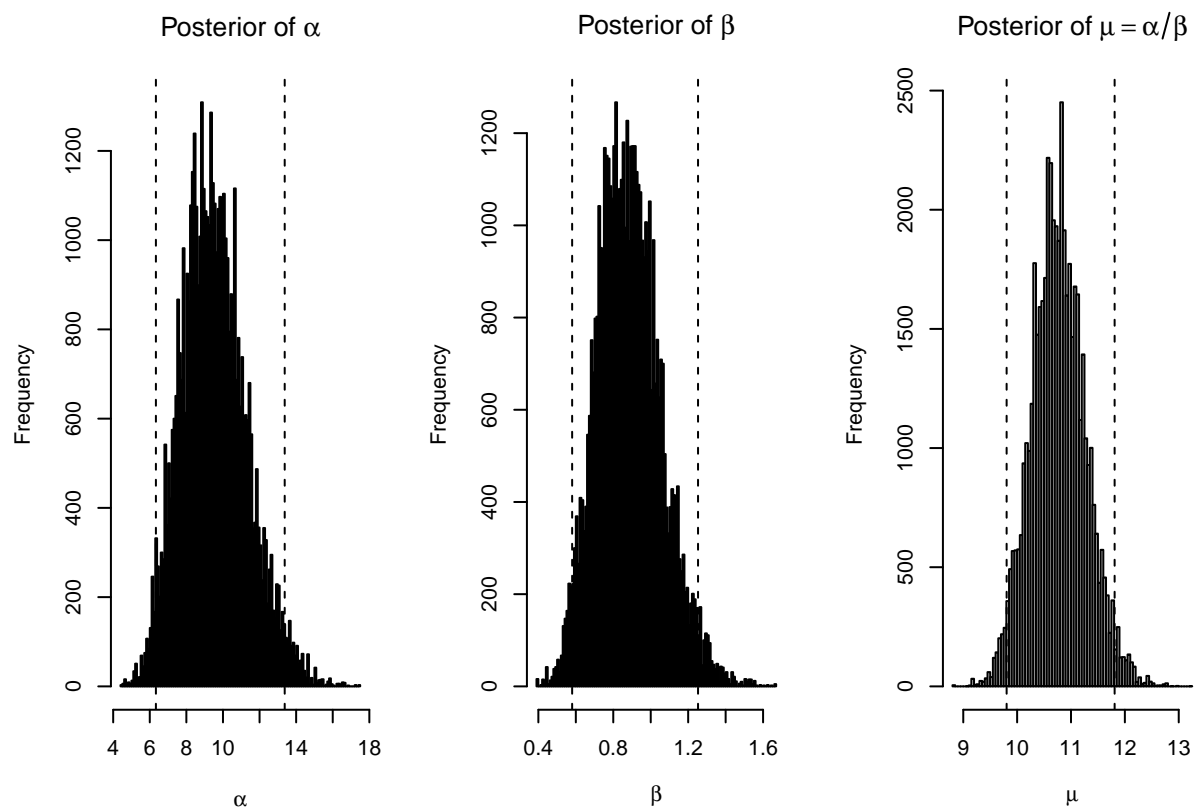
Note: The Credible Intervals are included in Table 2 ("CI. 2.5, CI. 97.5")

3.

Using both the 75th percentile and the range as data characteristics of potential interest, compute posterior predictive p-values from 10,000 posterior predictive datasets.

Answer

We then do additional posterior predictive checks to validate the results of our simulation. Using the kept MH draws from the prior Question, we simulated 10,000 posterior predictive datasets.

Table 5: Posterior predictive p-values (upper) 10,000 datasets

Statistic	Observed	PPP
75th percentile	12.9317	0.4255
Range	14.9500	0.5737

Generally, we want posterior predictive p-values that are not too large and not too small (so somewhere in the range of 0.2 to 0.7, or so); the values we observe for the statistics of interest seem suitable.

4.

Now consider the use of a Gibbs Sampling algorithm to simulate from the joint posterior of α and β and μ. Derive full conditional posterior densities for α and β Using these distributions, program a Gibbs Sampling algorithm and simulate 50,000 values from the joint posterior. Provide (with supporting evidence if appropriate),

- information on how you selected a burn-in period. Again, there is no need to compute Gelman-Rubin scale reduction factors for this assignment.

- summaries of the marginal posterior distributions of α and β, including histograms and five-number summaries, 95% central credible intervals, and correlation between α and β in the Markov chain.

Answer

Let $Y_i \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$ with density

$$f(y \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, y^{\alpha-1} e^{-\beta y}$$

for $y > 0$

Priors specified as $\alpha \sim \text{Uniform}(0, A)$ and $\beta \sim \text{Gamma}(\gamma_0, \lambda_0)$ using the rate parametrization.

Denote $S_1 = \sum_{i=1}^n \log y_i$ and $S_2 = \sum_{i=1}^n y_i$.

The joint posterior (up to a constant) is

$$\pi(\alpha, \beta \mid y) \propto \beta^{n\alpha} e^{-\beta S_2} \frac{e^{(\alpha-1)S_1}}{[\Gamma(\alpha)]^n} \beta^{\gamma_0-1} e^{-\lambda_0 \beta} \mathbb{1}_{(0,A)}(\alpha)$$

Collecting all terms in β from the joint posterior gives the kernel for the conditional posterior:

$$\beta^{n\alpha+\gamma_0-1} \exp\big(-(\lambda_0 + S_2)\beta\big)$$

so

$$\beta \mid \alpha, y \sim \text{Gamma}\big(\gamma_0 + n\alpha, \ \lambda_0 + S_2\big)$$

Collecting terms in α yields

$$\pi(\alpha \mid \beta, y) \propto \exp\Big(n\alpha \log\beta - n\log\Gamma(\alpha) + (\alpha-1)S_1\Big)\mathbf{1}_{(0,A)}(\alpha)$$

Where S_1 and S_2 are the sufficient statistics for the Gamma family, which simplifies the expression of the conditional posteriors.

Note: The conditional posterior for α does not belong to any standard family due to the presence of the $\log\Gamma(\alpha)$ term. Consequently, α must be updated using a random-walk Metropolis step within the Gibbs sampler (i.e., a Metropolis-within-Gibbs update), with proposals restricted to $(0, A)$.

With the full conditionals derived, we now proceed to simulation. As in Question 2, we assess an appropriate Burn-In period and tuning strategy. However, unlike in Question 2, since β is conditionally conjugate, tuning is required only for the Metropolis step on α.

Convergence diagnostics such as trace plots and autocorrelation functions are examined using zero Burn-In initially. The method primarily focuses on jointly simulating (α, β) and employs the log-likelihood for improved numerical stability, again, similar to Question 2.

ACF alpha (full)

ACF beta (full)
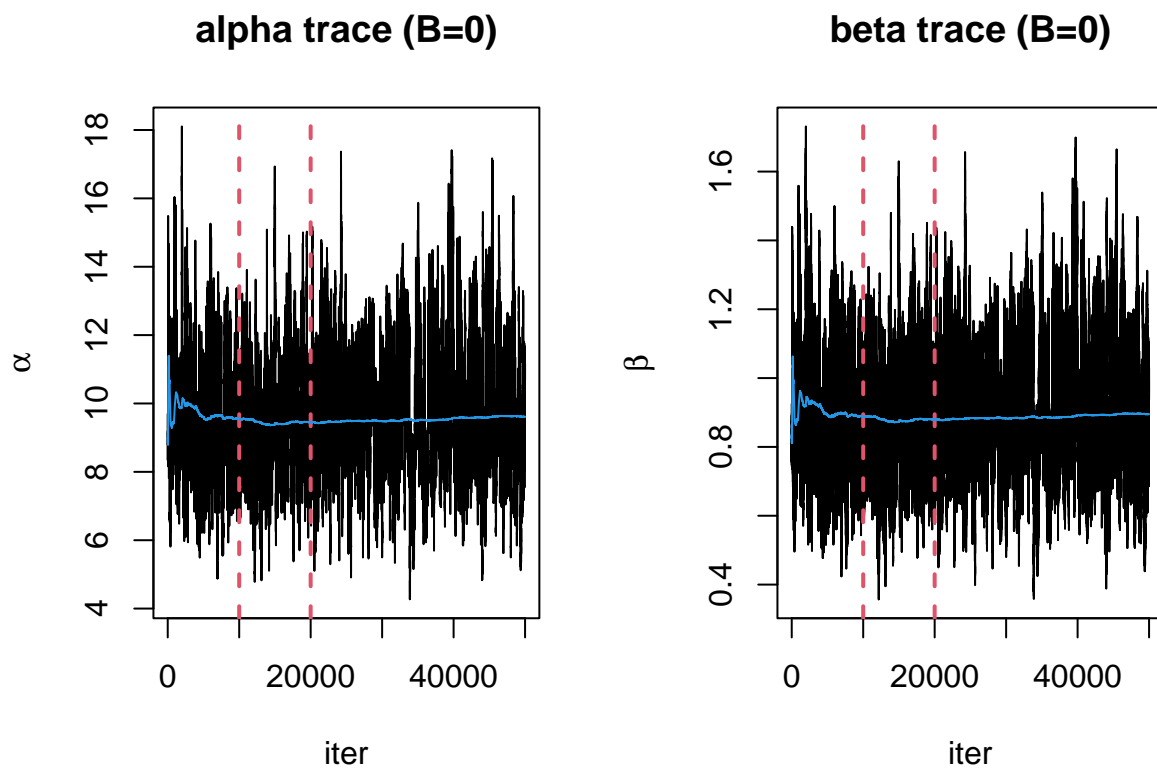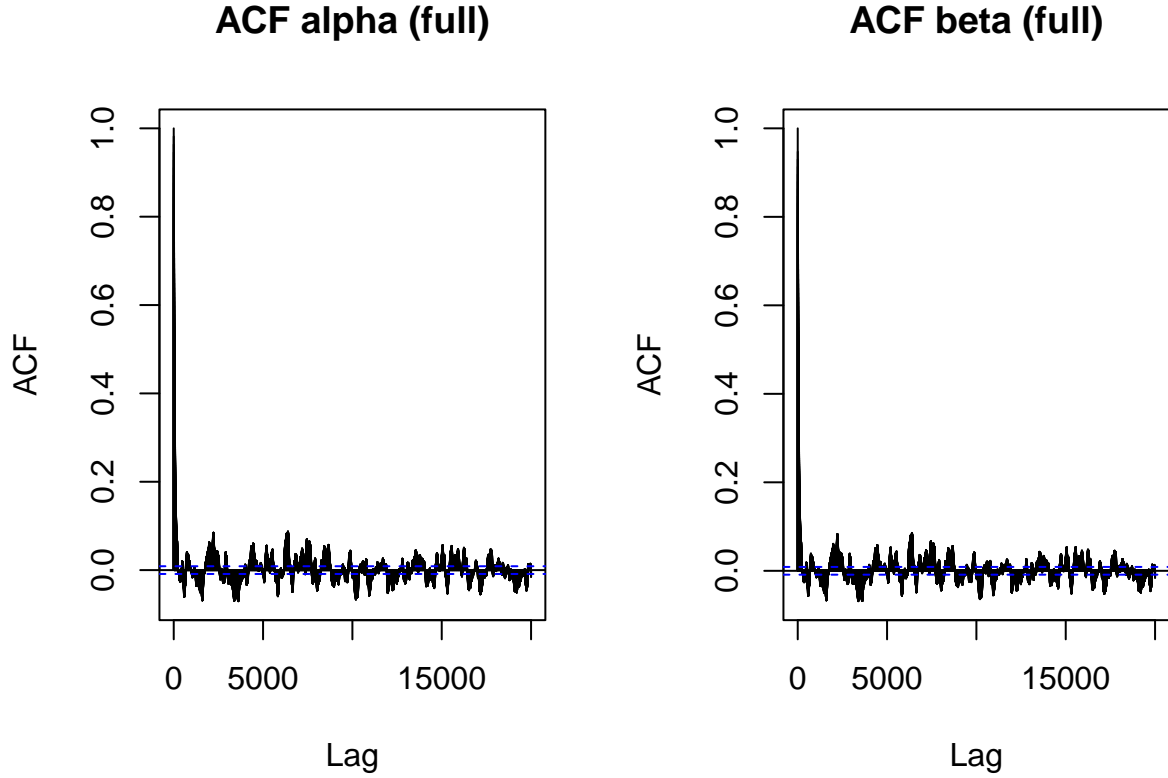
Similar to Question 2, we use a "running mean" (blue) to get a sense of when variability in the parameter "stabilizes". And again, similar (but different!) to Question 2, by iteration 10,000, the ACFs has decayed rapidly, and the trace plot has generally stabilized; to stay on the conservative side, we double this and set that equal to our Burn-In period, settling on a Burn-In period of 20,000.

Note: Similar to Question 2, although the ACF plot displays values that remain slightly outside the "white-noise" bounds around the proposed Burn-In period, the autocorrelation clearly decays toward zero and alternates in sign, indicating that dependence weakens over time. Therefore, it is still reasonable to suppose the proposed Burn-In period is appropriate.

For tuning then, we again scale our α variance in increments of 0.25, considering the same range of scales, and targeting an acceptance probability around 20-60%. However, given this is using the Gibbs sampling method, I am cautious about not properly "mixing", so apriori I will prioritize lower acceptance proabilities (we will accept more sampling that don't make a "jump", where by comparisons the MH algorithm will actively discard unsuitable "jump" candidates).

scale	valpha	acc_alpha
0.50	0.4766	0.5625
0.75	1.0723	0.4366
1.00	1.9063	0.3539
1.25	2.9787	0.2896
1.50	4.2893	0.2483
1.75	5.8382	0.2168
2.00	7.6254	0.1909

Given the scaling of 2.0, with acceptance probability around 19.1%, we now have both a suitable Burn-In

Period and tuned parameters for our "full simulation" via Gibbs.

Table 7: (Gibbs) Posterior summaries: five-number statistics and 95 central credible intervals

| Method | Parameter | Posterior Summary | | | | | | |
		Min	Q1	Median	Q3	Max	CI.2.5	CI.97.5
Gibbs	alpha	4.1562	8.2506	9.3828	10.6414	19.1854	6.3985	13.4101
Gibbs	beta	0.3475	0.7650	0.8738	0.9948	1.8442	0.5866	1.2588
Gibbs	mu	8.7153	10.4238	10.7501	11.0897	13.4660	9.8168	11.8063

Table 8: (Gibbs) Acceptance rates

Component	Acceptance Rate
Overall	19.1%

Table 9: (Gibbs) Correlation matrix for alpha and beta (Gibbs)

Parameter	alpha	beta
alpha	1.0000	0.9719
beta	0.9719	1.0000

Posterior of α Posterior of β Posterior of $\mu = \alpha/\beta$

13
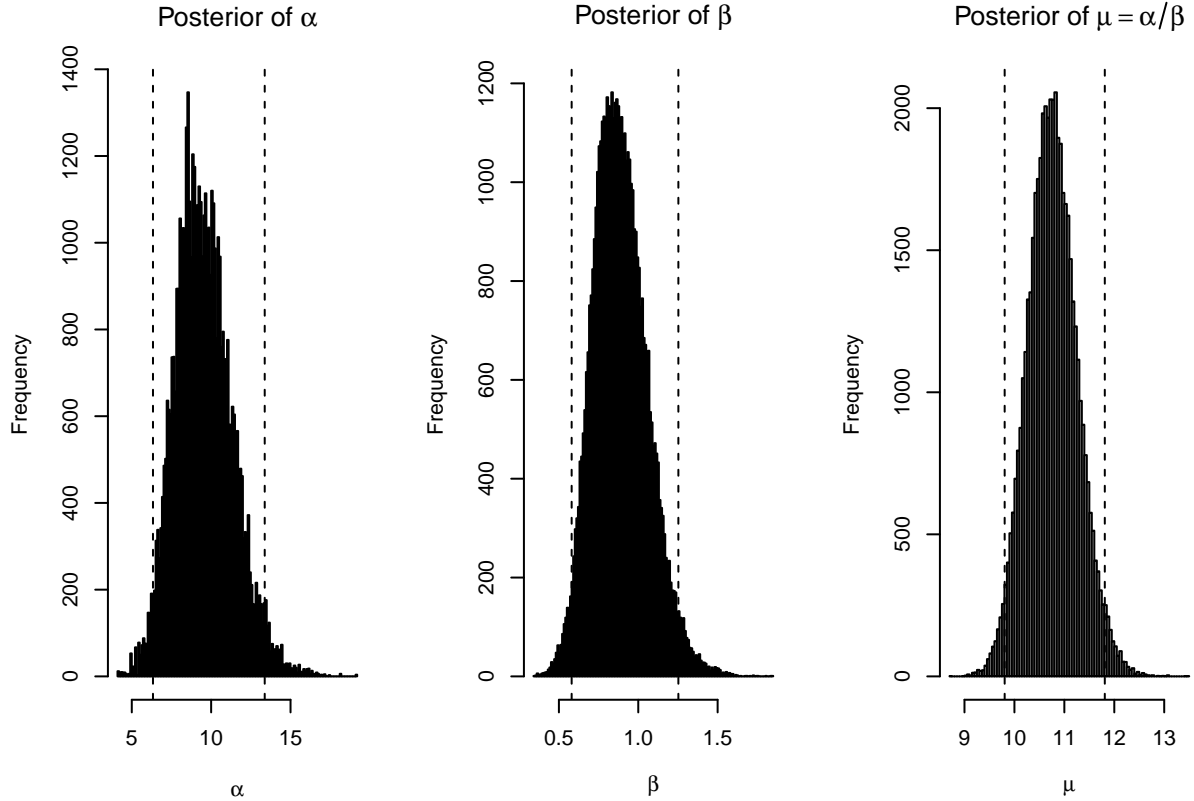
Note: While the phrasing of the question doesn't explicitly ask for Credible Intervals or 5-number summaries of μ, I felt this might have been an unintentional omission, given they are explicitly asked for previously for the MH method (and these quantities are of interest when comparing methods).

5.

Using both the 75th percentile and the range as data characteristics of potential interest, compute posterior predictive p-values from 10,000 simulated posterior predictive data sets.

Answer

We then do additional posterior predictive checks to validate the results of our simulation. Using the Gibbs draws from the prior Question, we simulated 10,000 posterior predictive datasets.

Table 10: Posterior predictive p-values (upper) 10,000 datasets

Statistic	Posterior Predictive Check	
	Observed	PPP
75th percentile	12.9317	0.4416
Range	14.9500	0.5720

Generally, we want posterior predictive p-values that are not too large and not too small (so somewhere in the range of 0.2 to 0.7, or so); the values we observe for the statistics of interest seem suitable, though potentially a bit large. Further comparison and analysis is discussed in Question 6.

6.

Compare your results from the use of Metropolis-Hastings and Gibbs Sampling.

Answer

The two sampling approaches, Metropolis–Hastings (MH) and Gibbs (Metropolis-within-Gibbs), produce highly consistent (across methods and against the observed sample) posterior inferences for all parameters of interest, α, β, and μ. Both have nearly identical posterior locations and spreads for α, β, and the derived parameter $\mu = \alpha/\beta$.

Specifically: For the MH method, the 95% central credible intervals were:

$$\alpha \in [6.32, \ 13.72], \quad \beta \in [0.578, \ 1.294], \quad \mu \in [9.823, \ 11.788],$$

While the Gibbs method had:

$$\alpha \in [6.40, \ 13.41], \quad \beta \in [0.587, \ 1.259], \quad \mu \in [9.810, \ 11.792].$$

The pairwise posterior correlation between α and β is also very similar, and strongly positive, at approximately 0.97 for both methods. This is a rather unsurprising result, and while we did not take advantage of this correlation, we perhaps could adapt our sampling methods to be more efficient by incorporating this relationship.

From a more descriptive perspective, the MH posterior histograms appear slightly more "jagged," reflecting that its random-walk updates move both parameters jointly (possibly yielding higher autocorrelation). By comparison the Gibbs method, which draws β directly from its conjugate full conditional and updates α by a Metropolis step, produces smoother histograms and traces. This may also relate to how proposed candidates are treated, as MH proceeds by accepting or rejecting entire candidate pairs, while Gibbs continually updates by drawing directly from conditional distributions, along with their respective acceptance probabilities (16.8% for MH vs. 19.1% for Gibbs).

Posterior predictive checks are also consistent across methods. Both yield one-sided posterior predictive p-values between approximately 0.4 and 0.6 for both the 75th percentile and range statistics, with:

$$p_{\text{ppp}}^{(Q_{0.75})} \approx 0.43 \text{ or } 0.44, \quad p_{\text{ppp}}^{(\text{range})} \approx 0.57$$
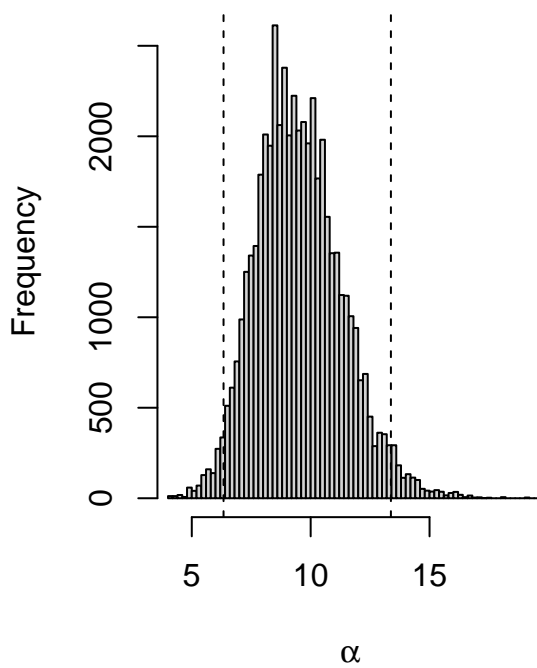
Since the ideal results are posterior predictive p-values around 0.4 to 0.6 (since these are not interpreted like traditional p-values), our results indicate that the simulations provide an adequate job at simulating these statistics of interest (statistics which are notably **not** the sufficient statistics). Another way to say this is: The simulated datasets from each posterior reproduce the empirical features of the data without obvious systematic bias.

Though unfortunately there are not contextual details to draw upon in concluding, there are a few remarks to summarize. To start, both algorithms appear to converge to very similar (if not the same) posterior distributions for α, β, and μ. In comparing the two methods, the MH method shows slightly rougher marginal behavior but comparable central estimates, while the Gibbs method achieves smoother mixing. In both cases, we do an adequate job of simulating the original data, such that, after identifying and applying an appropriate Burn-In and tuning, the two methods give nearly indistinguishable posterior summaries and predictive performance.
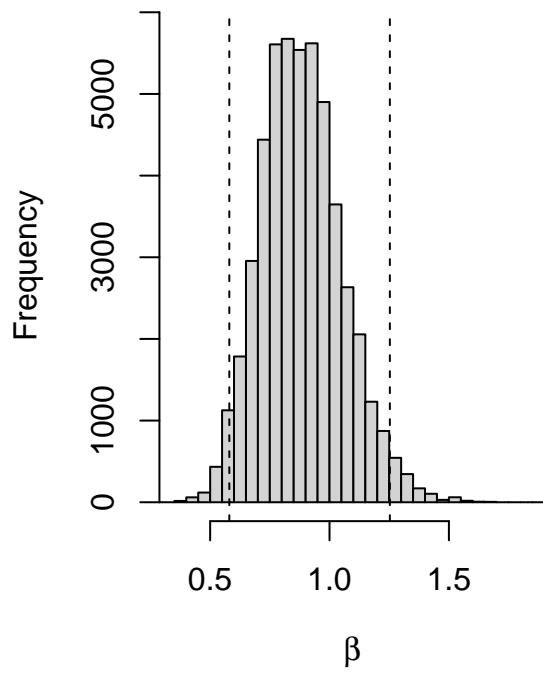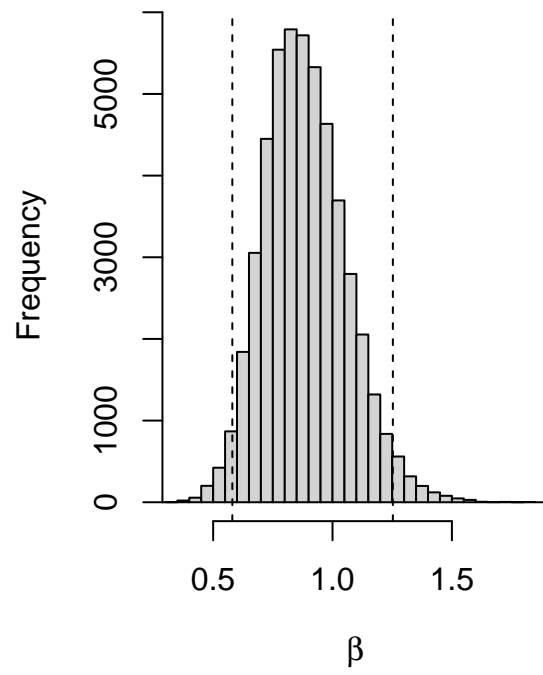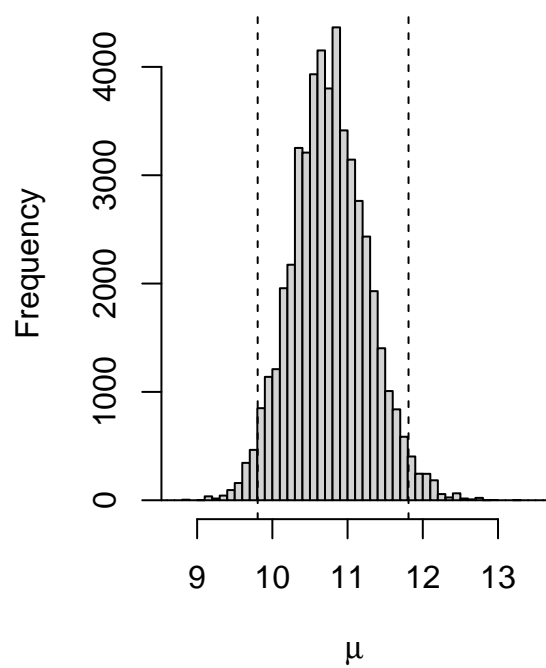
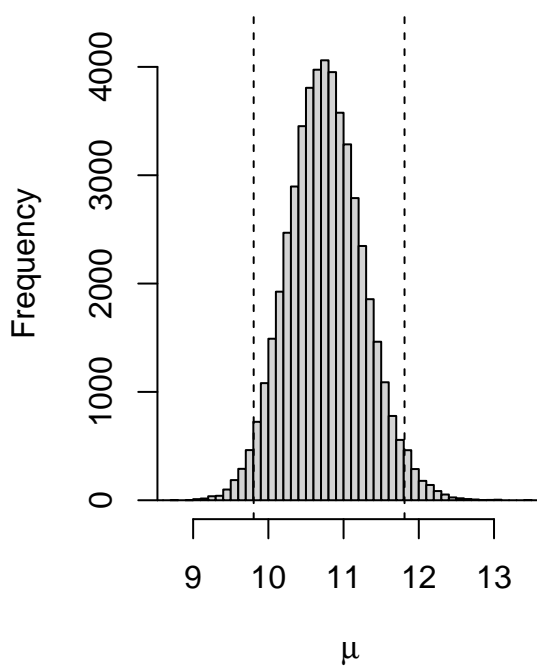MH – Posterior of α   Gibbs – Posterior of α
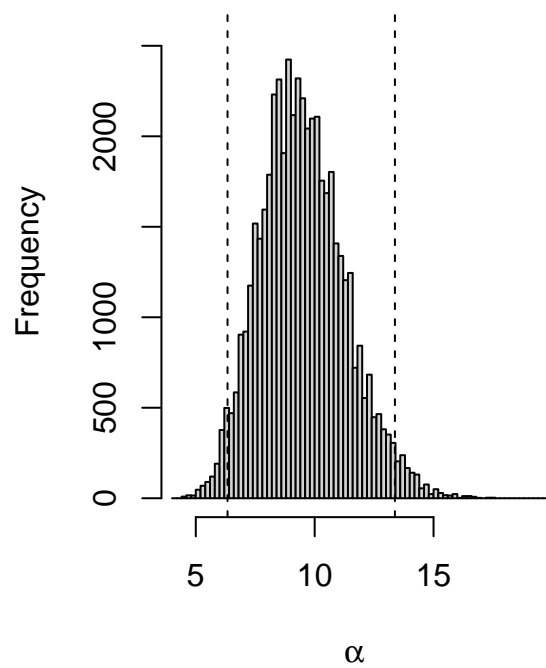
MH – Posterior of β

Gibbs – Posterior of β

MH – Posterior of $\mu = \alpha/\beta$ ⋅ Gibbs – Posterior of $\mu = \alpha/\beta$

7.

On this particular assignment, attach your R code for functions you programmed to do the necessary computations as an APPENDIX – not part of the body of your answer.

Answer

R Code attached separately. Note: R code functions are shown in the attached, and then their respective call(s) are noted.