Stat 532 Assignment 9

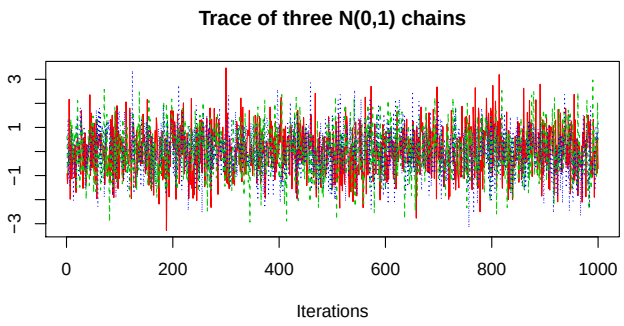
Kenny Flagg

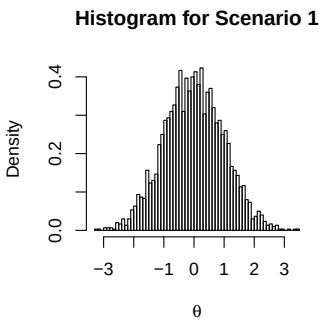
November 4, 2015

1. (a) I feel like I should report more than just sample path plots so I've also included histograms and the code to generate the chains for each scenario.

   i. Scenario 1: These chains are well-behaved and converged.

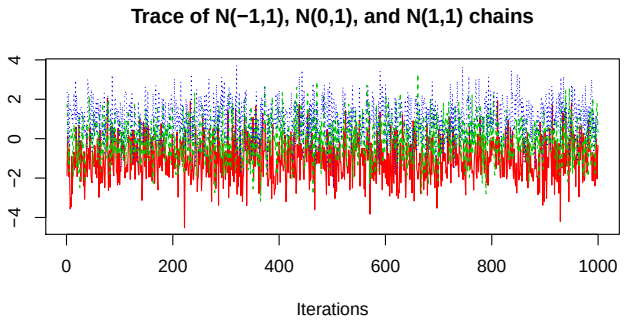
```
set.seed(8914)
chains1 <- mcmc.list(replicate(3, mcmc(rnorm(1000, 0, 1)), simplify = FALSE))
```

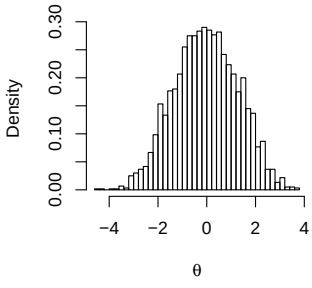




   ii. Scenario 2: These chains are very slowly mixing.

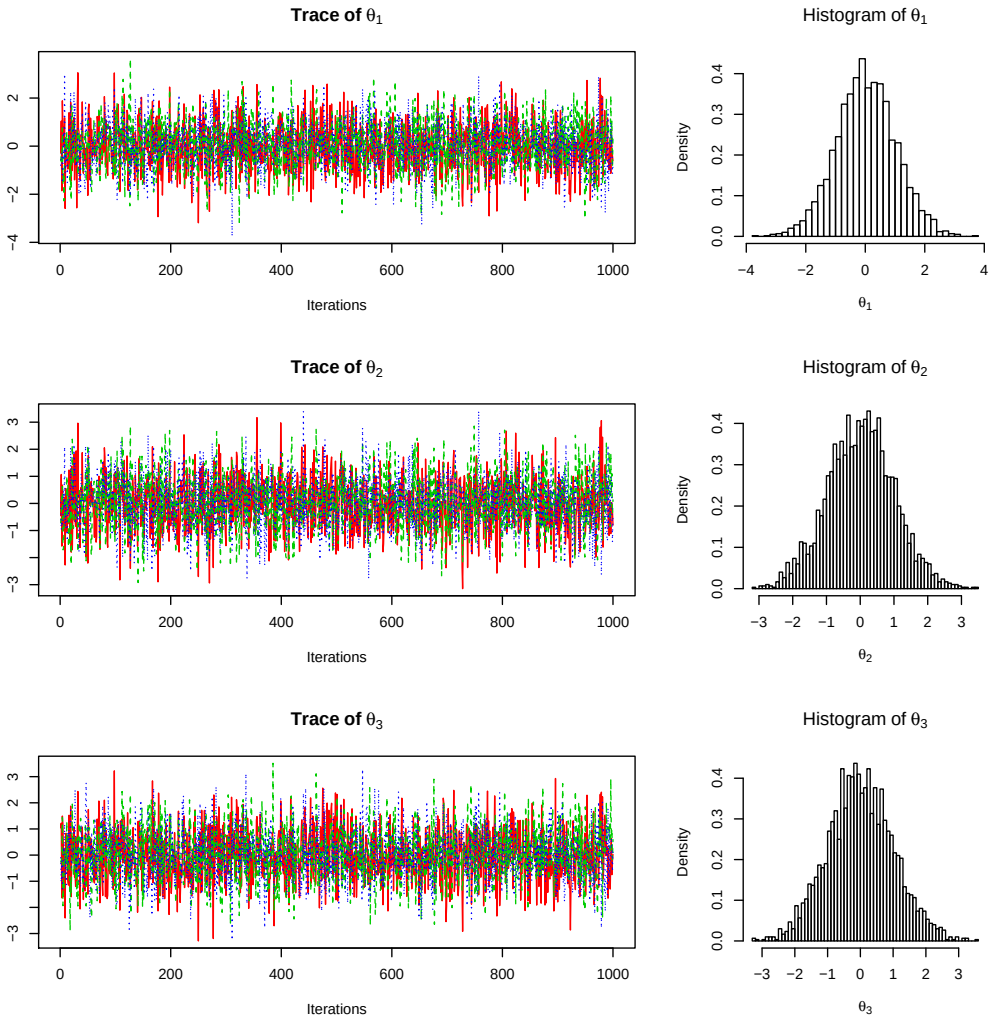
```
set.seed(2532)
chains2 <- mcmc.list(mcmc(rnorm(1000, -1, 1)),
                     mcmc(rnorm(1000, 0, 1)),
                     mcmc(rnorm(1000, 1, 1)))
```





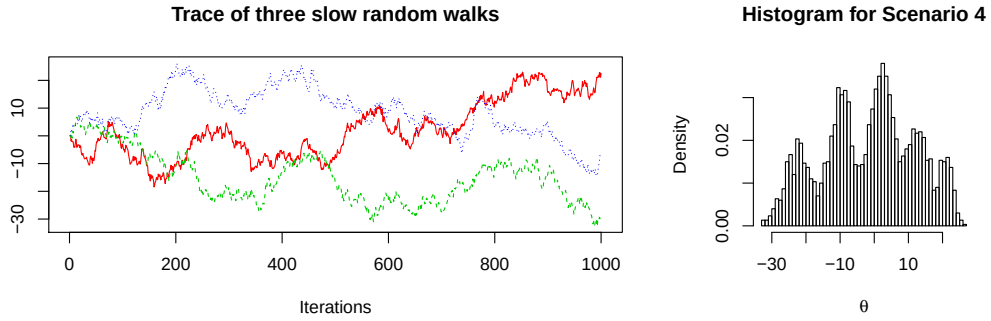
iii. Scenario 3: These chains have converged to the stationary distribution, but the parameters are strongly correlated.

```
set.seed(89362)
mvchain <- function(n){
  x <- rmvnorm(n, rep(0, 3), diag(0.2, 3) + 0.8)
  colnames(x) <- paste('theta', 1:3, sep = '')
  return(mcmc(x))
}
chains3 <- mcmc.list(replicate(3, mvchain(1000), simplify = FALSE))
```



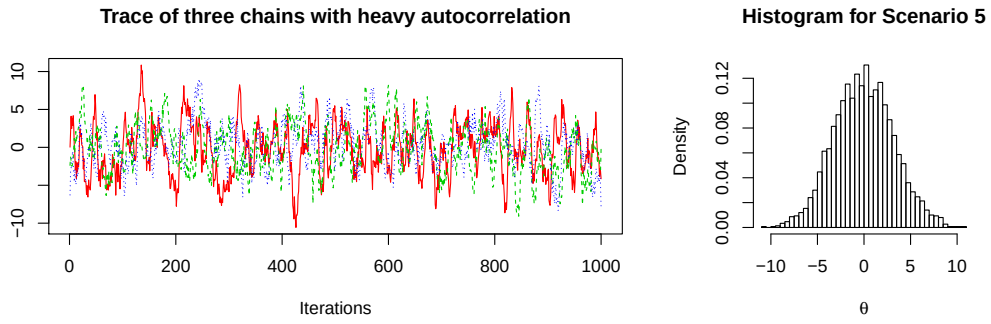
iv. Scenario 4: These chains are not converging.

```
set.seed(3624)
chains4 <- mcmc.list(replicate(3, mcmc(diffinv(rnorm(999, 0, 1))), simplify = FALSE))
```



v. Scenario 5: These chains have converged to the stationary distribution, but have high autocorrelation.

```
set.seed(4532)
chains5 <- mcmc.list(replicate(3,
            mcmc(filter(rnorm(1000), filter = rep(1, 10), circular = TRUE)),
            simplify = FALSE))
```



(b) I have `coda` version 0.16-1, which computes $\widehat{R}$ and n_{eff} differently from the methods described in the text.

The `gelman.diag` help page and page 284 of BDA3 agree that the estimated posterior variance is

$$\widehat{\text{var}}^{+}(\psi|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

where W and B are, respectively, the within-chain and between chain sample variances. The formulas for $\widehat{R}$ itself differ. On page 285, the book gives the form

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}(\psi|y)}{W}},$$

while the `gelman.diag` function computes

$$\widehat{R} = \sqrt{\frac{(d+3)\widehat{\text{var}}^{+}(\psi|y)}{(d+1)W}},$$

where

$$d = \frac{2\left(\widehat{\text{var}}^+(\psi|y)\right)^2}{\text{Var}\left(\widehat{\text{var}}^+(\psi|y)\right)}$$

is a method of moments estimator of the degrees of freedom. When d is small, the $\widehat{R}$ values will differ.

On pages 286-287, the textbook describes estimating the effective sample size as

$$\hat{n}_{\text{eff}} = \frac{mn}{1 + 2\sum_{t=1}^{T} \widehat{\rho}_t},$$

where $\widehat{\rho}_t$ is the estimated autocorrelation for lag t and T is set so that the sum stops when the autocorrelation estimates become too noisy. The help page for `effectiveSize` is vague and unhelpful, but I looked at the code for the function and saw that it computes

$$n_{\text{eff}} = \frac{n\text{Var}(\psi)}{S(0)}.$$

$S(0)$ is the spectral density at frequency 0, estimated by the `spectrum0.ar` function which in turn fits autoregressive models. I don't know enough about time series to understand any of that last sentence, but I think it is reasonable to assume that the BDA3 method and the `coda` method are similar but not equivalent.

(c) The $\widehat{R}$ and n_{eff} values appear in Table 1. The default behavior of the `gelman.diag` function is to discard the first half of each chain. There is no reason to do that for these examples, so I had it use the full chains.

	$\widehat{R}$	n_{eff}
Scenario 1	1.0004	3000
Scenario 2	1.7543	3000
Scenario 3, θ_1	1.0017	3394
Scenario 3, θ_2	1.0021	2966
Scenario 3, θ_3	1.0015	3326
Scenario 4	2.1302	13
Scenario 5	1.0080	384

Table 1: PSRF values and effective sample sizes.

(d) This was good excuse to get to know the diagnostics I've been assigned for the one-pagers. The Heidelberger and Welch diagnostic works on one single chain, so for each scenario I concatenated the three chains into one in the order $(\text{Chain}_1, \text{Chain}_2, \text{Chain3})$. I ran the `heidel.diag` function from `coda` with $\alpha = 0.05$ for the stationarity test and $\epsilon = 5$ as the required halfwidth ratio. I collected the results on the next page in Table 2.

Michael provided his `QuantileEquivalenceMCMC` library. Quantile equivalence plots for $p = 0.025$, 0.5, and 0.975 are presented in Figure 1 on page 6. The "converged" or "not converged" text was generated from output of the `qed` function. I will consider chains converged if the empirical probabilities are within $b = 0.025$ of the truth, so I chose $\epsilon = 0.021$ and $\alpha = 0.05$.

	Stationarity Test	Start Iter.	p-value	Halfwidth Test	Mean	Halfwidth
Scenario 1	Passed	1	0.141	Failed	-0.005	0.035
Scenario 2	Failed		0.001			
Scenario 3, θ_1	Passed	1	0.142	Passed	0.029	0.034
Scenario 3, θ_2	Passed	1	0.084	Passed	0.017	0.035
Scenario 3, θ_3	Passed	1	0.330	Passed	0.009	0.033
Scenario 4	Passed	1	0.902	Failed	-1.658	10.674
Scenario 5	Passed	1	0.669	Failed	-0.033	0.385

Table 2: Heidelberger and Welch results using $\epsilon = 5$ and $\alpha = 0.05$.

(e) I consider Scenarios 1, 3, and 5 to be converged, and Scenarios 2 and 4 to be unconverged. $\widehat{R}$ performed well, and QED worked for tail probabilities. Other results were mixed.

Scenarios 2 and 4 had large $\widehat{R}$ values; the others all had $\widehat{R}$ values near 1. If we follow Gelman's protocol of saying chains are converged when $\widehat{R} < 1.1$ then we would correctly classify all of these scenarios.

The n_{eff} results are interesting, but their usefulness is questionable. Scenarios 1 and 2 had effective sample sizes of 3000 because they are sequences of independent draws. This is not related to convergence. In Scenario 3, two of the parameters had n_{eff} larger than the actual number of samples. I suspect this is an anomaly due to the correlation between parameters. It certainly is does not provide meaningful information. Scenarios 4 and 5 have small n_{eff} values because they contain little independent information. Running 3,000 simulations and getting an effective sample size of 13 is certainly a situation that warrants inspection, but a small n_{eff} is not reason to label a chain as not converged.

Results from the Heidelberger and Welch stationarity test are puzzling. It classified all scenarios as stationary except for Scenario 2. The three chains were appended end-to-end, so in Scenario 2 the test was done on a chain that jumped between three different distributions. I am surprised that that was classified as non-stationary when the wandering chains of Scenario 4 or the autocorrelated sequences of Scenario 5 were classified as stationary. I should also note that the halfwidth test is context-dependent. The test is passed if the ratio of the margin of error to the mean is below a user-specified threshold. In all of these scenarios the mean is near 0, so the halfwidth test is unreliable.

The quantile equivalence tests for the probabilities 0.025 and 0.975 gave the correct results in all cases except Scenario 5. Oddly, none of the medians were found to be converged. I have no explanation for this. The plots are useful tools for assessing strength of evidence. For example, the plots for Scenario 5 show that two chains are in close agreement, but the third has less variability than the other two. There may be a problem with that one chain, but they are probably near convergence.

My preferred way to assess convergence would be to use a both multi-chain diagnostic on all chains and a single-chain diagnostic on each chain individually. QED and $\widehat{R}$ are good choices for multi-chain diagnostics. QED has a meaning that directly relates to inference, and the plots are informative. $\widehat{R}$ is reliable for symmetric distributions. I would not use Heidelberg and Welch until I study the theory behind it and I come to understand how and when to use it. Effective sample size should be checked to ensure it is not suspiciously small, but it is not worth reporting.

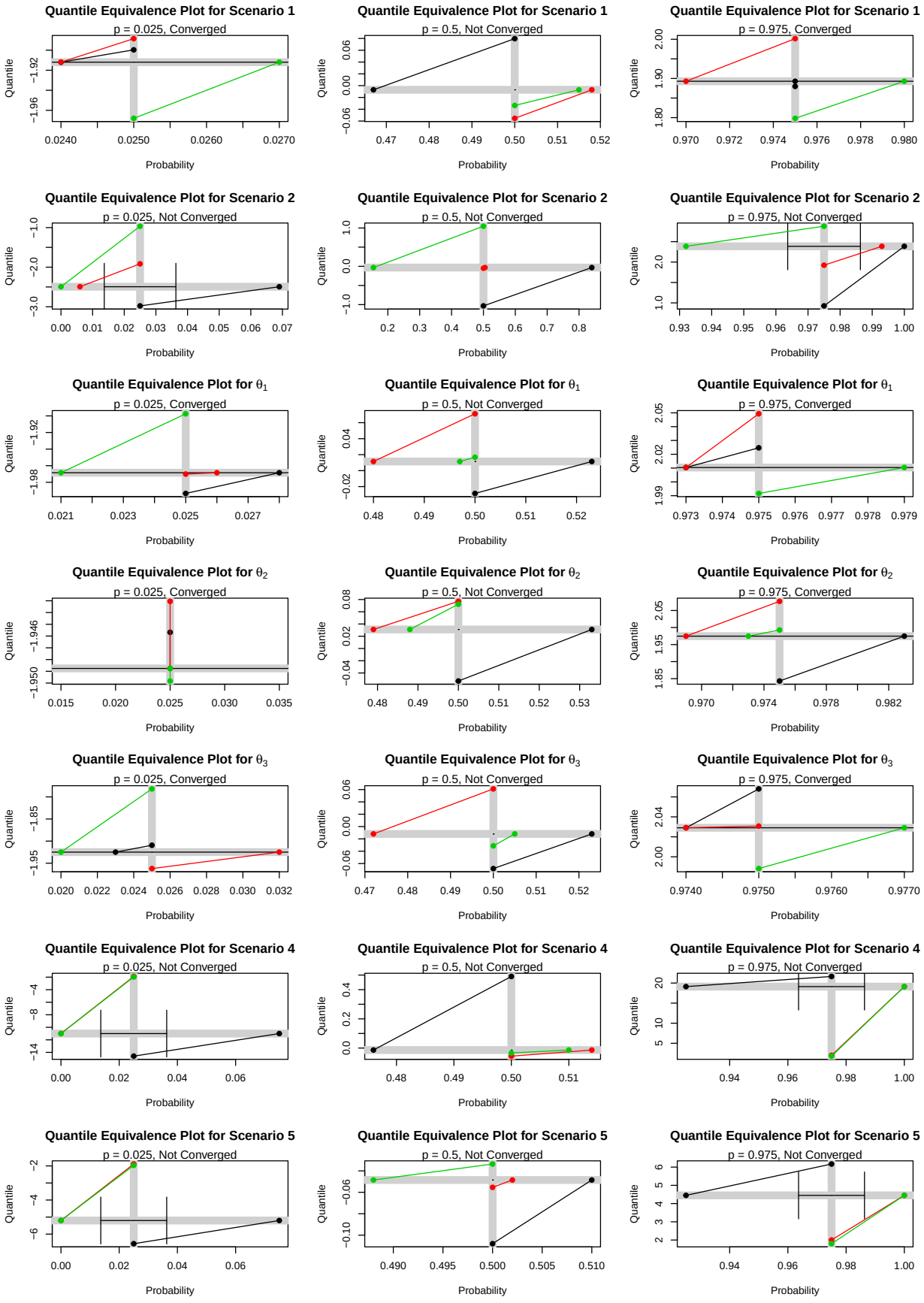


Figure 1: Quantile equivalence plots and tests using $\epsilon = 0.021$ and $\alpha = 0.05$.

2. (a) Using the identity $\alpha = \frac{\nu}{2}$, $\beta = \frac{\nu}{2}s^2$, inverse-Gamma$(0.001, 0.001)$ is equivalent to scaled-inverse-$\chi^2(0.0005, 5 \times 10^{-10})$.

   (b) I plotted all of the distributions in Figure 2. For Prior D, I chose 1 degree of freedom to make the distribution rather wide.

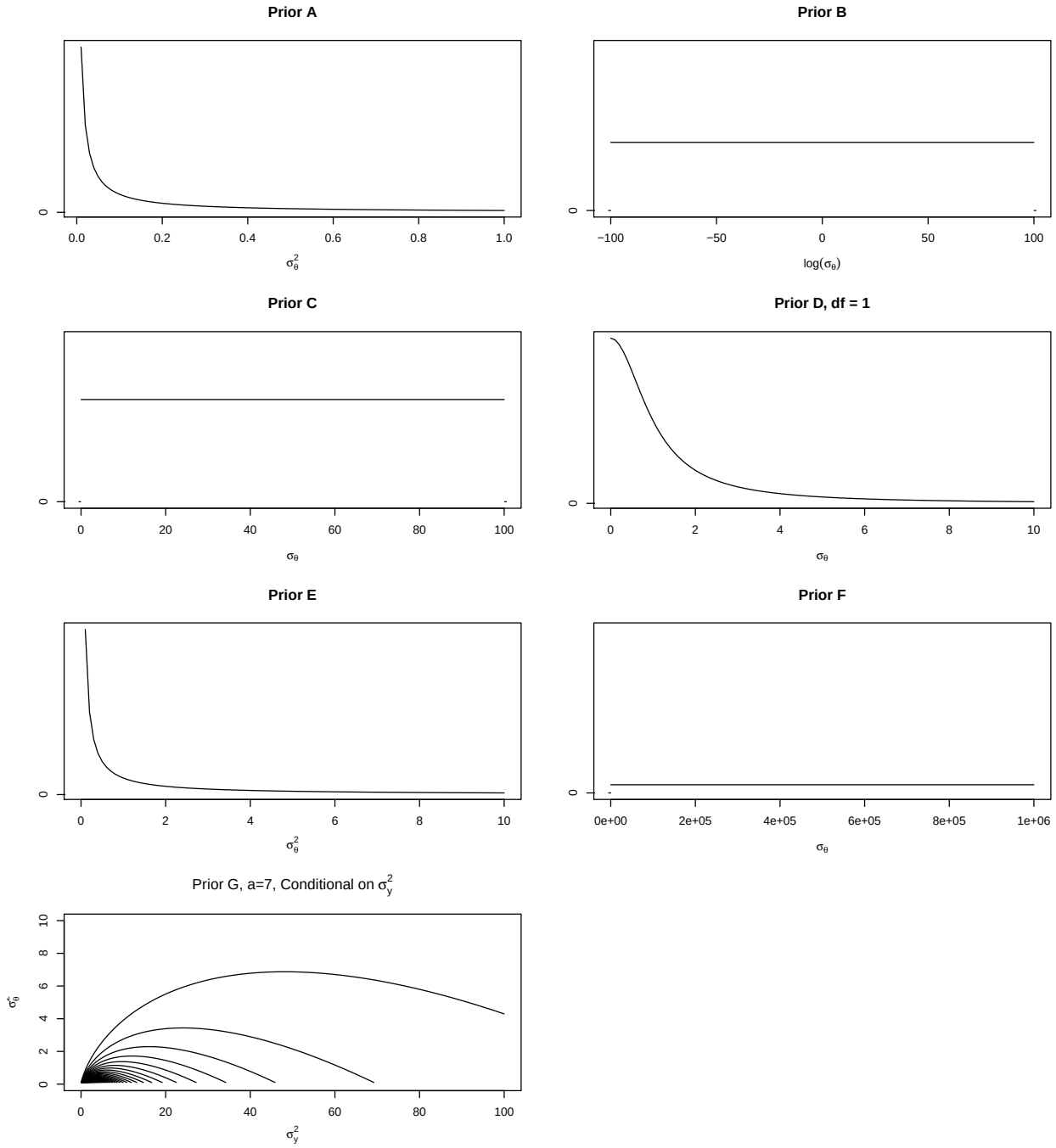


Figure 2: Seven possible prior distributions for hierarchical scale parameters.

(c) With a simple Jacobian calculation, I found that Prior B becomes $p(\sigma_\theta) = \frac{1}{200\sigma_\theta}$ for $e^{-100} < \sigma_\theta < e^{100}$. The density blows up near the lower bound and is very flat everywhere else. Since $e^{100} \approx 2.7 \times 10^{43}$, Prior B allows for occasional mind-bogglingly large values. Any graphical presentation of this will make Prior C appear invisible, so Figure 3 shows 100 draws from Prior B displayed alone and compares the density functions on a scale more appropriate for Prior C.

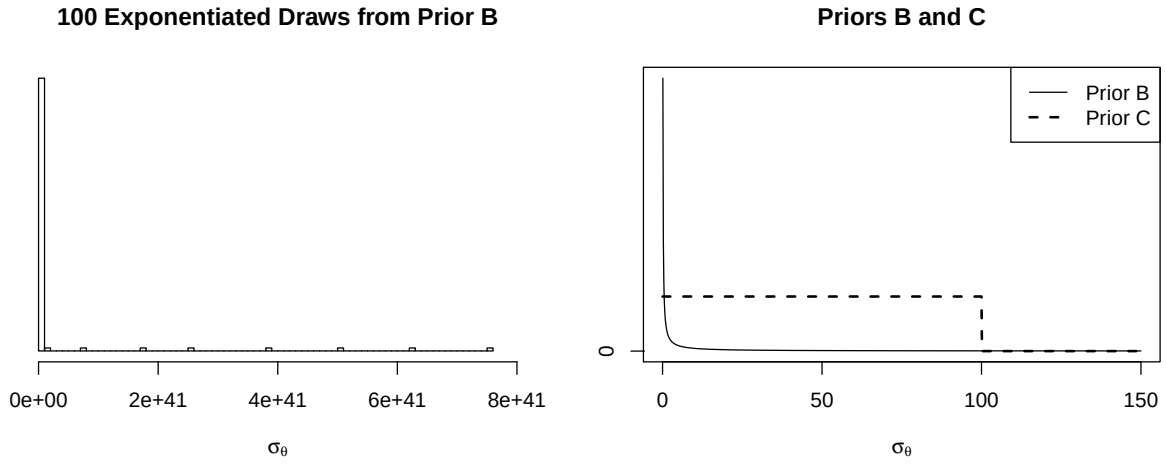


Figure 3: Left: 100 draws from Prior B. Right: Priors B and C on the scale of Prior C.

(d) Prior C becomes $p\left(\sigma_\theta^2\right) = \frac{1}{200\sqrt{\sigma_\theta^2}}$ for $0 < \sigma_\theta^2 < 10{,}000$. Figure 4 shows that this is well-constrained compared to Prior A. The philosophy of modeling a complete lack of a priori knowledge should allow the variance to be entirely unbounded, but when prior draws on the order of 10^{306} occur, I think the analyst needs to justify the decision *not* to use a weakly informative prior.

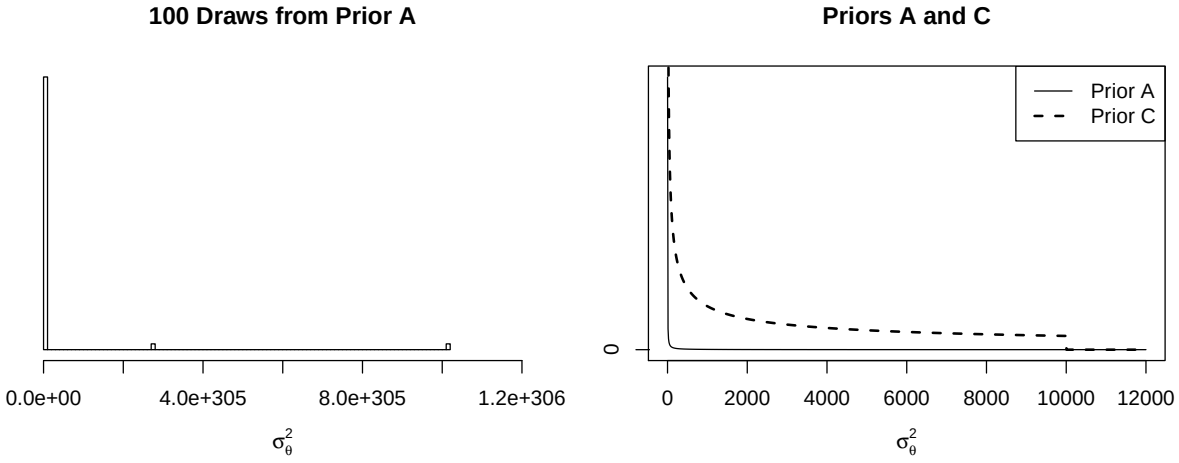


Figure 4: Left: 100 draws from Prior A. Right: Priors A and C on the scale of Prior A.

(e) The t distribution with 1 degree of freedom is the same as the Cauchy distribution, so prior D is a half-Cauchy distribution if df = 1.

3. (a) i. It seemed easiest to make use of base R functions. For $\sigma > 0$, the folded-noncentral-t density is simply the sum of the noncentral-t densities at $\pm\sigma$. This is illustrated in Figure 5.
Since R does not include a scaled-t distribution, I modified the central t distribution into a location-scale distribution.

```
# n       number of draws
# df      degrees of freedom
# scale   scale parameter
# center  normal center
dft <- function(t, df, center = 0, scale = 1){
  return(ifelse(t > 0,
                (dt((t-center)/scale, df = df) + dt((-t-center)/scale, df = df)) / scale,
                0)
         )
}
```

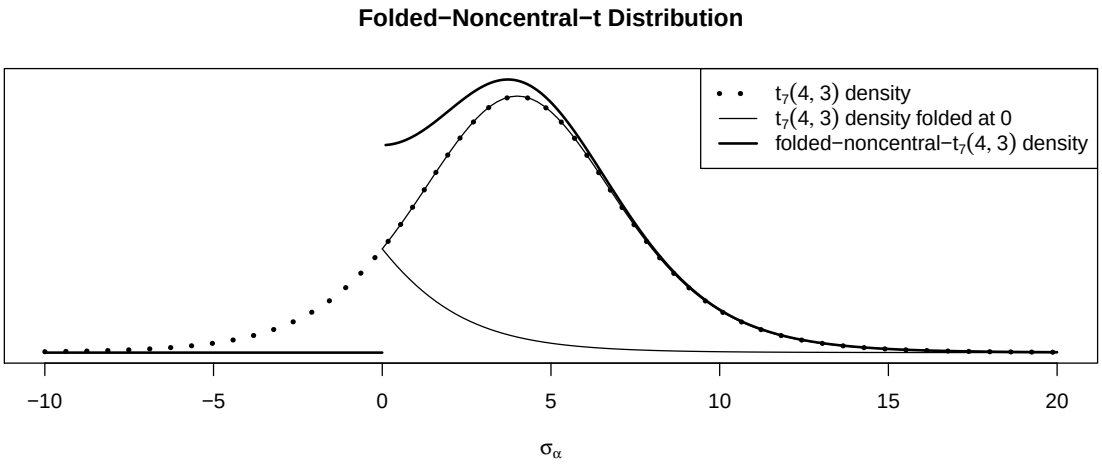


Figure 5: Illustration of the construction of the folded-noncentral-t distribution.

ii. Gelman comments that we can set the Normal scale at 1 and specify the scale of the square-root-inverse-χ^2. I prefer to rescale the Normal component and use R's built in functions without having to think about the inverse-scale parameter of a square-root-χ^2 distribution, although fixing both scales at 1 and multiplying by the draw by the chosen scale for the t distribution would achieve the same result.

```
# n       number of draws
# df      degrees of freedom
# scale   scale parameter
# center  normal center
rft <- function(n, df, scale = 1, center = 0){
  return(abs(rnorm(n, mean = center, sd = scale)) * sqrt(df / rchisq(n, df = df)))
}
```

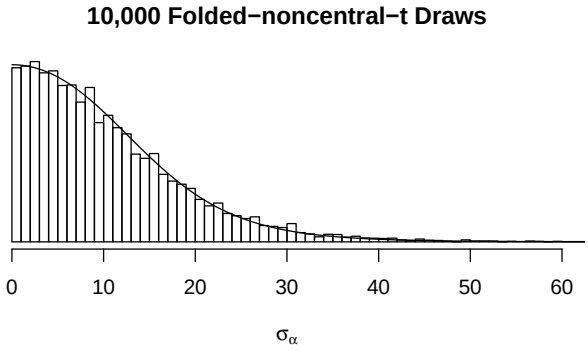


Figure 6: Histogram of 10,000 draws from a folded-noncentral-$t_7(6, 10)$ distribution, with the density curve overlaid.

(b) The basic hierarchical model is

$$y_{ij} = \mathrm{N}(\mu + \alpha_j, \sigma_y^2);$$
$$\alpha_j = \mathrm{N}(0, \sigma_\alpha^2)$$

for groups $j = 1, \ldots, J$ and individuals $i = 1, \ldots, n_j$ in group j.

(c) The following code generated the data, which appear in Figure 7.

```
set.seed(8725)
n <- c(5, 10, 30, 30, 20, 25, 50, 10)
J <- length(n)
sigsq.y <- 4
mu <- 20
sigsq.a <- 2

# Generate alphas and a list of y vectors
alpha <- rnorm(J, 0, sqrt(sigsq.a))
y <- lapply(1:J, function(j){rnorm(n[j], mu+alpha[j], sqrt(sigsq.y))})
```

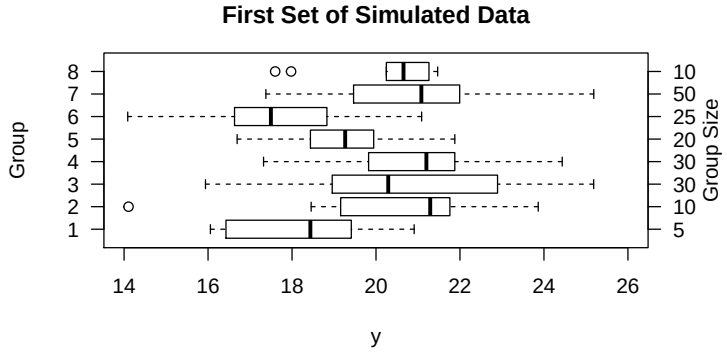


Figure 7: Fake data from the hierarchical model with $\mu = 20$, $\sigma_y^2 = 4$, and $\sigma_\alpha^2 = 2$.

(d) We could use $p(\mu, \sigma_y) \propto \frac{1}{\sigma_y}$ for a weakly informative prior. I tend to prefer it over the uniform prior on μ, σ_y because I like to constrain σ_y to be small.

After reading Gelman's paper, I would first consider a half-Cauchy prior for σ_α. Gelman suggests that a uniform prior will work when there are 8 groups, but a half-Cauchy with large scale is also quite flat.

(e) My fake data has a range of a little less than 12 which suggests an overall standard deviation for y (ignoring groups) of at most 6. The value of σ_α is certainly smaller than this. To be conservative, I set $A = 10$.

Keeping problem 2 in mind, I did not want to set ϵ too small. I chose $\epsilon = 0.5$, which allows large values but has most of its mass below $\sigma_\alpha^2 = 100$.

I fit the model in Stan. For each prior, I ran four chains for 1,000 iterations of warmup and then simulated 2,500 samples from each chain. Stan selected random initial values.

The posterior distributions and some quantiles of σ_α and μ are compared in Figure 9 on page 12. The $\sigma_\alpha^2 \sim$ Inv-Gamma$(0.5, 0.5)$ prior was less spread out than the other priors and resulted in the posterior distribution of σ_α having a shorter right tail compared to the other posteriors. The posterior distributions from the uniform priors are practically identical.

For μ, the medians and quartiles for are nearly identical. The 95% posterior interval from the inverse-Gamma prior is slightly narrower than the others, but the posterior inferences are essentially the same in all three cases.

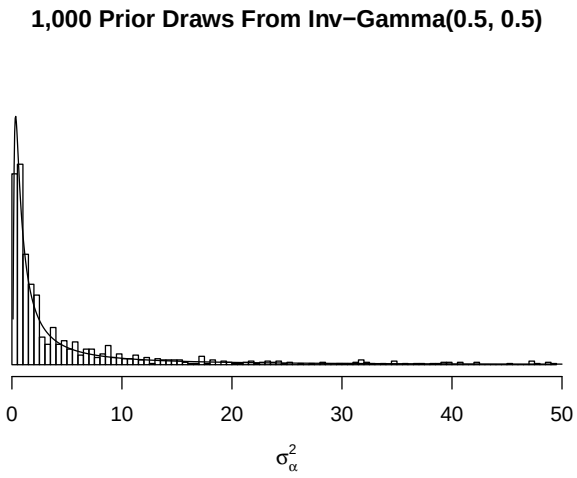


Figure 8: Inverse-Gamma prior with $\epsilon = 0.5$. 104 draws exceeded 50 and do not appear on this plot. 79 draws exceeded 100. The maximum was 2,112,847.3.

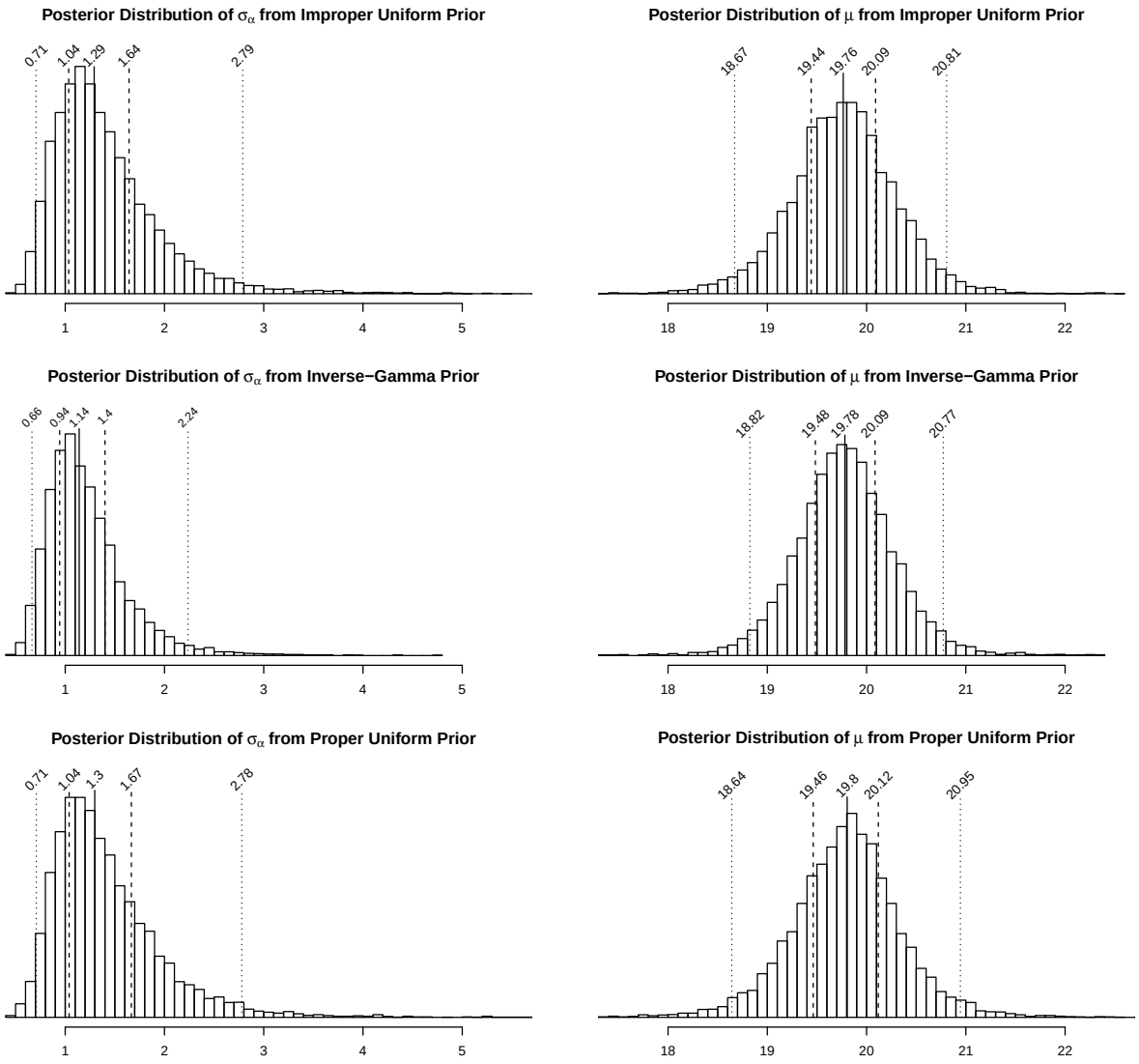


Figure 9: Comparison of the posterior distributions of σ_α and μ from three different priors, for the data with large between-group variation. Vertical lines mark the 0.025, 0.25, 0.5, 0.75, and 0.975 quantiles. Top: $\sigma_\alpha \sim \text{Unif}(0, \infty)$ prior. Center: $\sigma_\alpha^2 \sim \text{Inv-Gamma}(0.5, 0.5)$ prior. Bottom: $\sigma_\alpha \sim \text{Unif}(0, 10)$ prior.

(f) The model with redundant parameters is

$$y_{ij} = \text{N}(\mu + \xi\eta_j, \sigma_y^2);$$
$$\eta_j = \text{N}(0, \sigma_\eta^2)$$

where $\alpha_j = \xi\eta_j$ and $\sigma_\alpha = |\xi|\sigma_\eta$. The folded Normal prior on σ_α is equivalent to $\xi \sim \text{N}(0, \sigma_0^2)$, $\sigma_\eta = 1$ constant. The half-Cauchy prior of σ_α is $\xi \sim \text{N}(0, \sigma_0^2)$, $\sigma_\eta \sim \text{Inv-}\chi_1^2$. Again, I intended to use weakly informative priors to keep the posterior distributions

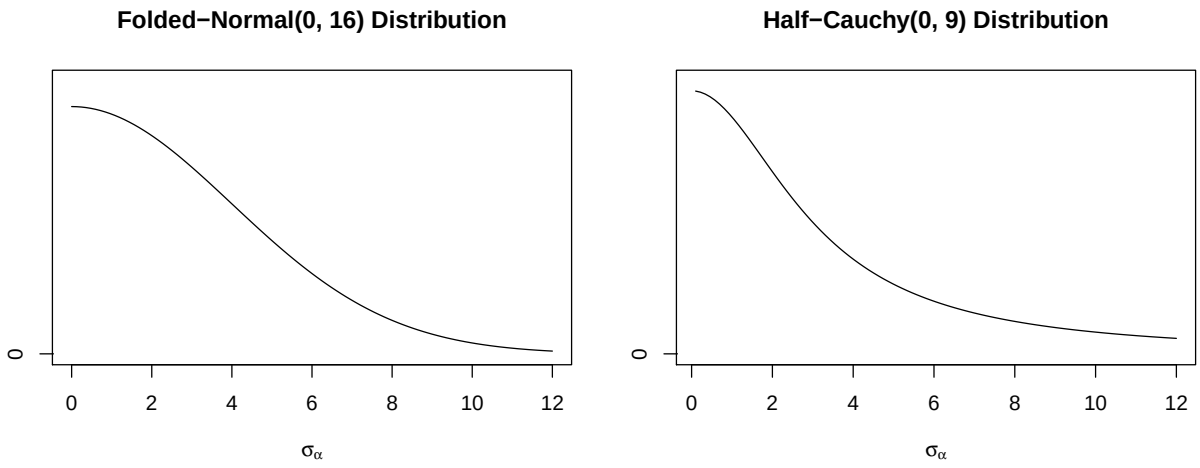


Figure 10: Priors for the redundant-parameter model.

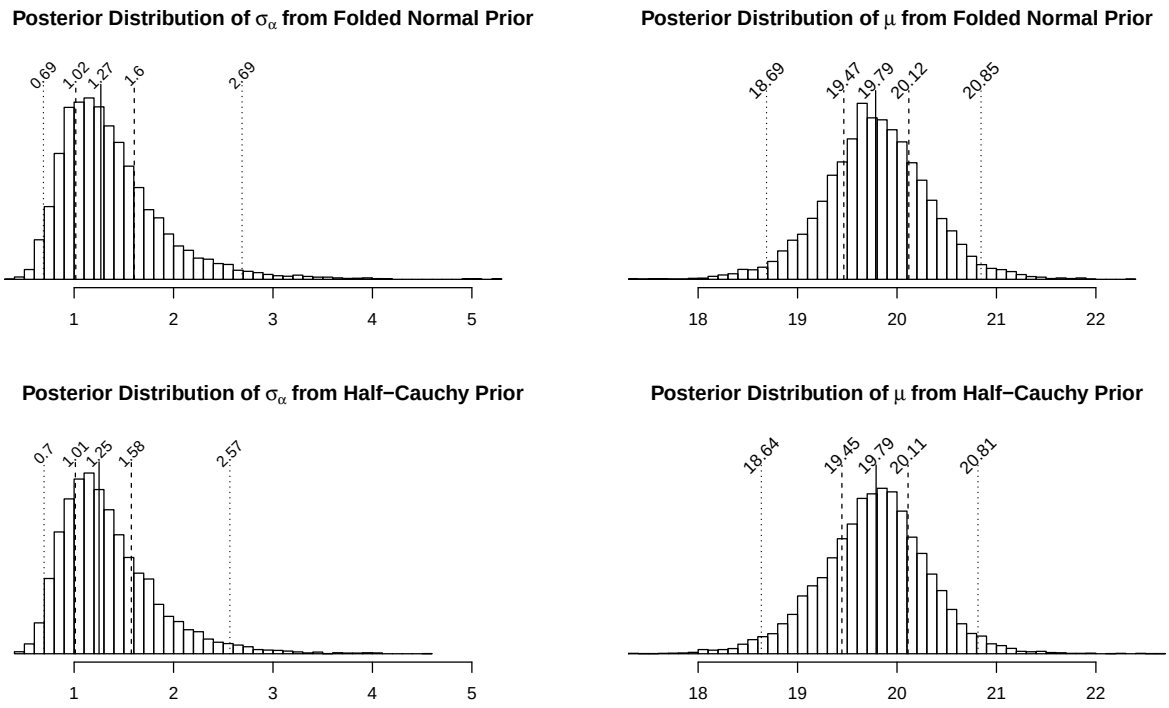


Figure 11: Comparison of posterior distributions of σ_α and μ from the redundant parameter model with two different priors, for the data with large between-group variation. Vertical lines mark the 0.025, 0.25, 0.5, 0.75, and 0.975 quantiles. Top: Folded Normal, scale 4. Bottom: Half-Cauchy, scale 3.

constrained. I chose a scale of $\sigma_0 = 4$ for the folded Normal and a scale of $\sigma_0 = 3$ for the half-Cauchy because these have most of their mass below 6 but do not entirely rule out large values. For μ and σ_y, I used $p(\mu, \sigma_y) \propto \frac{1}{\sigma_y}$.

I fit this model in Stan, again running four chains for a 1,000 iteration warmup period and then for another 2,500 simulations. Stan selected random initial values.

The posterior distributions for the redundant parameter model appear in Figure 11. Both priors resulted in posterior distributions for μ that match the posteriors from (c). The posterior distributions of σ_α have slightly shorter right tails compared to the results form the Uniform priors. However, these posteriors were less tightly constrained than the posterior from the inverse-Gamma prior. These results are what I expected; the Uniform priors allow large values to happen often. The folded-Normal and half-Cauchy distributions restrain the σ_α values without concentrating a large mass at zero like the inverse-Gamma does.

4. (a) I changed σ_α^2 to 0.01 and re-ran the same code that generated the original data. The new data appear in Figure 12. Ignoring an outlier in group 2, these data have a range of about 9, so there is almost as much overall variation as there was in the first set of simulated data. There is noticeably less variation in where the groups are centered. Since my previous priors were more dispersed than necessary, and I want to continue to be very weakly informative, I used the same priors as before.

Once again, the models were fit in Stan with four chains, 1,000 warmup simulations, 2,500 simulations sampled, and Stan-selected random initial values. The posterior distributions from the basic hierarchical model appear in Figure 13 and the posteriors from the expanded model are shown in Figure 14.

The posterior distributions for σ_α match each other closely, except that the posterior resulting from the inverse-Gamma(0.5, 0.5) prior is shifted right compared to the others and centered at 0.5. This makes no sense, so I also tried using inverse-Gamma(0.01, 0.01), inverse-Gamma(0.1, 0.1), and inverse-Gamma(1, 1). The inverse-Gamma(0.01, 0.01) resulted in a posterior that was tightly stacked against 0; the other two were very similar to the inverse-Gamma(0.5, 0.5) result. This was a very clear illustration that members of the inverse-Gamma(ϵ, ϵ) family are not uninformative in this situation.

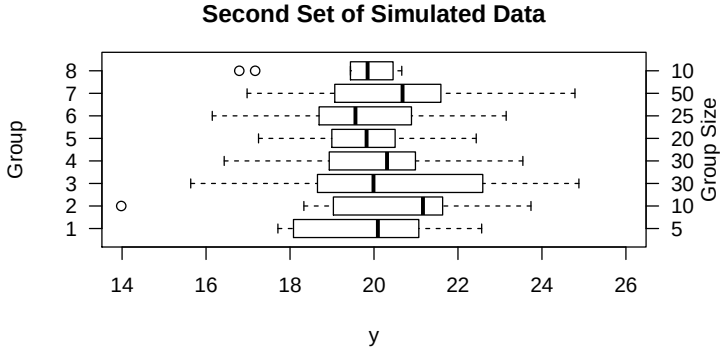


Figure 12: Fake data with $\mu = 20$, $\sigma_y^2 = 4$, and $\sigma_\alpha^2 = 0.01$.

The posterior distributions for μ are all essentially the same, with the exception of slightly longer tails coming from the inverse-Gamma prior.

Compared to the posterior distributions given the previous data, these posteriors for σ_α have much smaller spreads. Except for the result from the inverse-Gamma prior, these posteriors allow the group-level variance to be very close to 0. None of the posteriors in problem 3 allowed that.

Interestingly, the posterior distributions for μ are also narrower than the posteriors in problem 3.

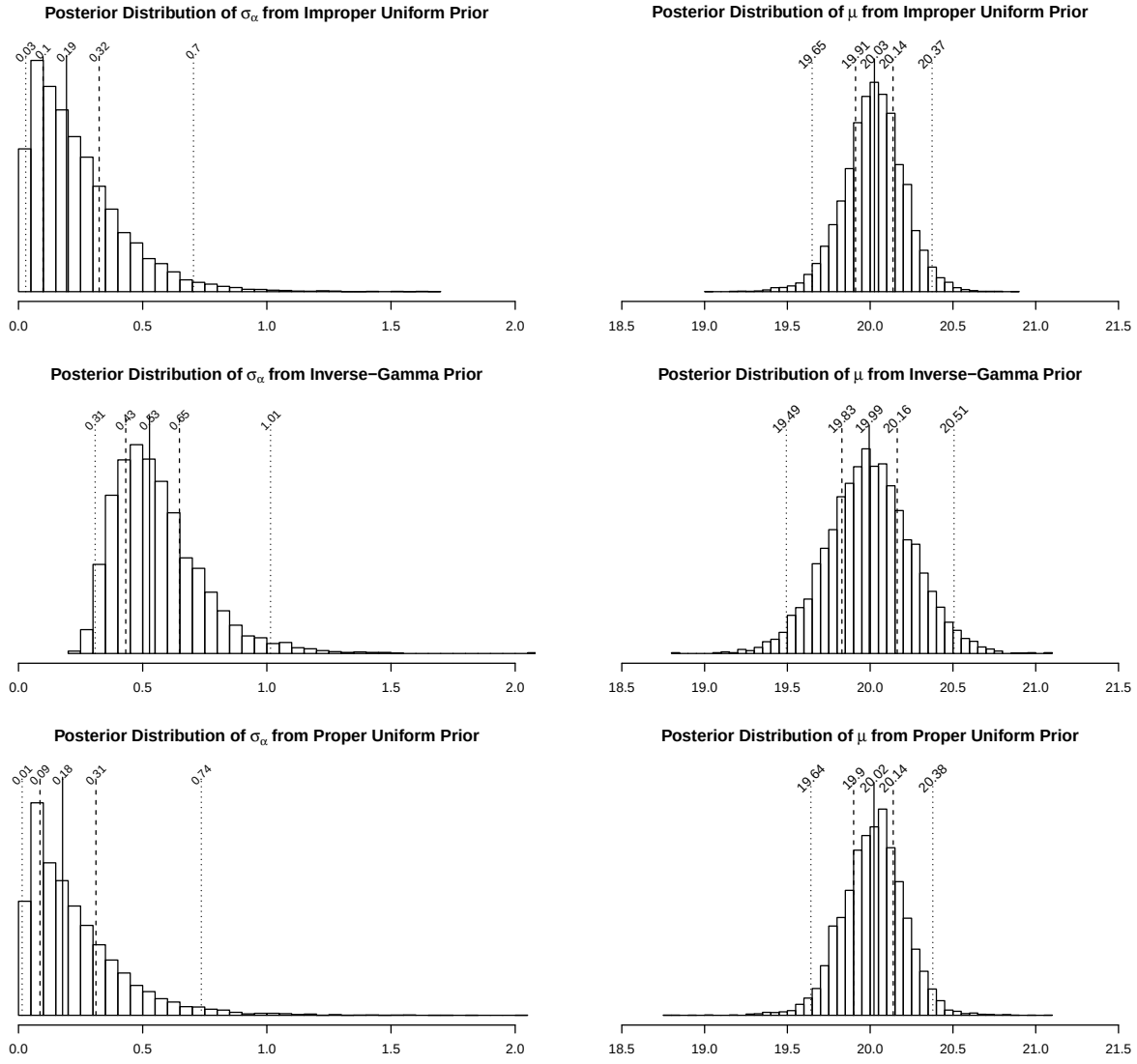


Figure 13: Comparison of posterior distributions of σ_α and μ from three different priors, for the data with small between-group variation. Vertical lines mark the 0.025, 0.25, 0.5, 0.75, and 0.975 quantiles. Top: $\sigma_\alpha \sim \text{Unif}(0, \infty)$. Center: $\sigma_\alpha^2 \sim \text{Inv-Gamma}(0.5, 0.5)$. Bottom: $\sigma_\alpha \sim \text{Unif}(0, 10)$.

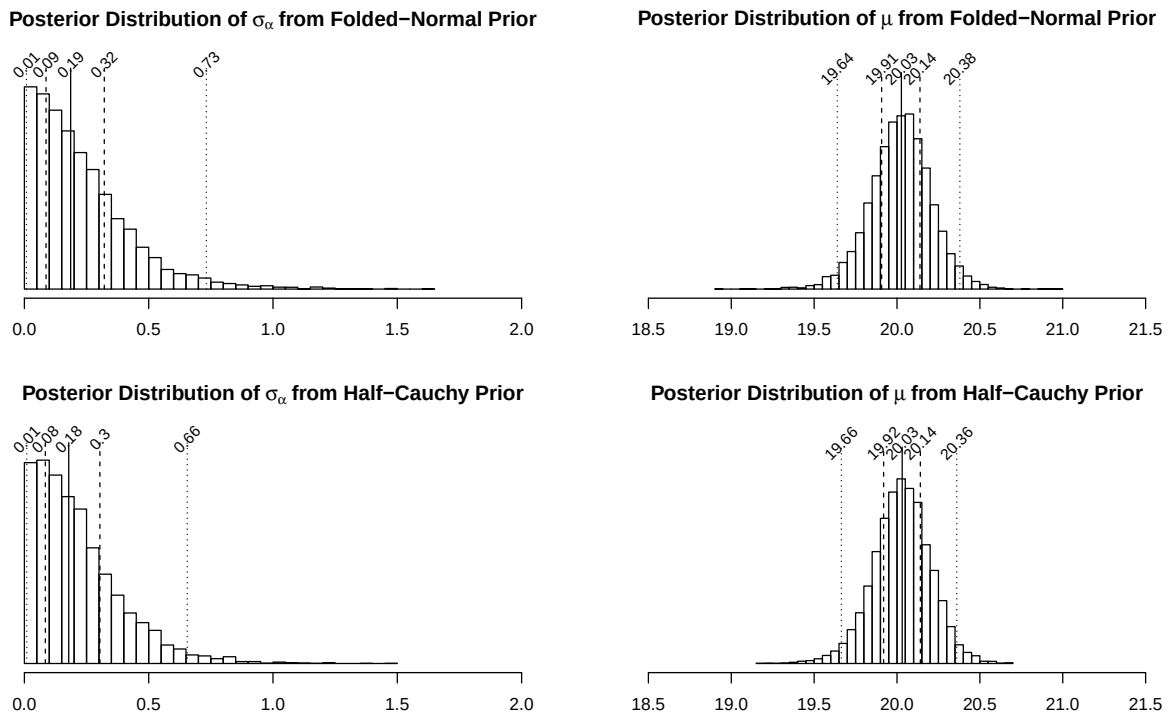


Figure 14: Comparison of posterior distributions of σ_α and μ from the redundant parameter model with two different priors, for the data with small between-group variation. Vertical lines mark the 0.025, 0.25, 0.5, 0.75, and 0.975 quantiles. Top: Folded Normal, scale 4. Bottom: Half-Cauchy, scale 3.

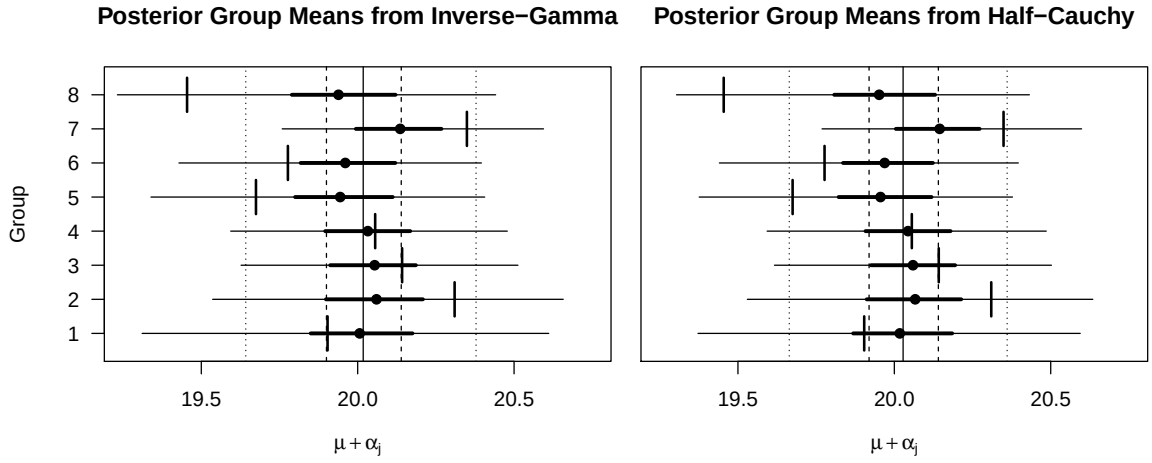


Figure 15: Posterior distributions of the group means, $\mu + \alpha_j$. Vertical segments are the observed sample averages. Vertical lines the posterior mean, 0.025, 0.25, 0.75, and 0.975 quantiles of $\mu + \alpha_j$ over all j.

(b) Figure 15 displays 50% and 95% posterior intervals for the group means from the models using inverse-Gamma and half-Cauchy priors. Both plots are nearly identical, despite the very different posteriors for σ_α. There is heavy shrinkage towards the overall mean, consistent with the between-group variance being close to 0.

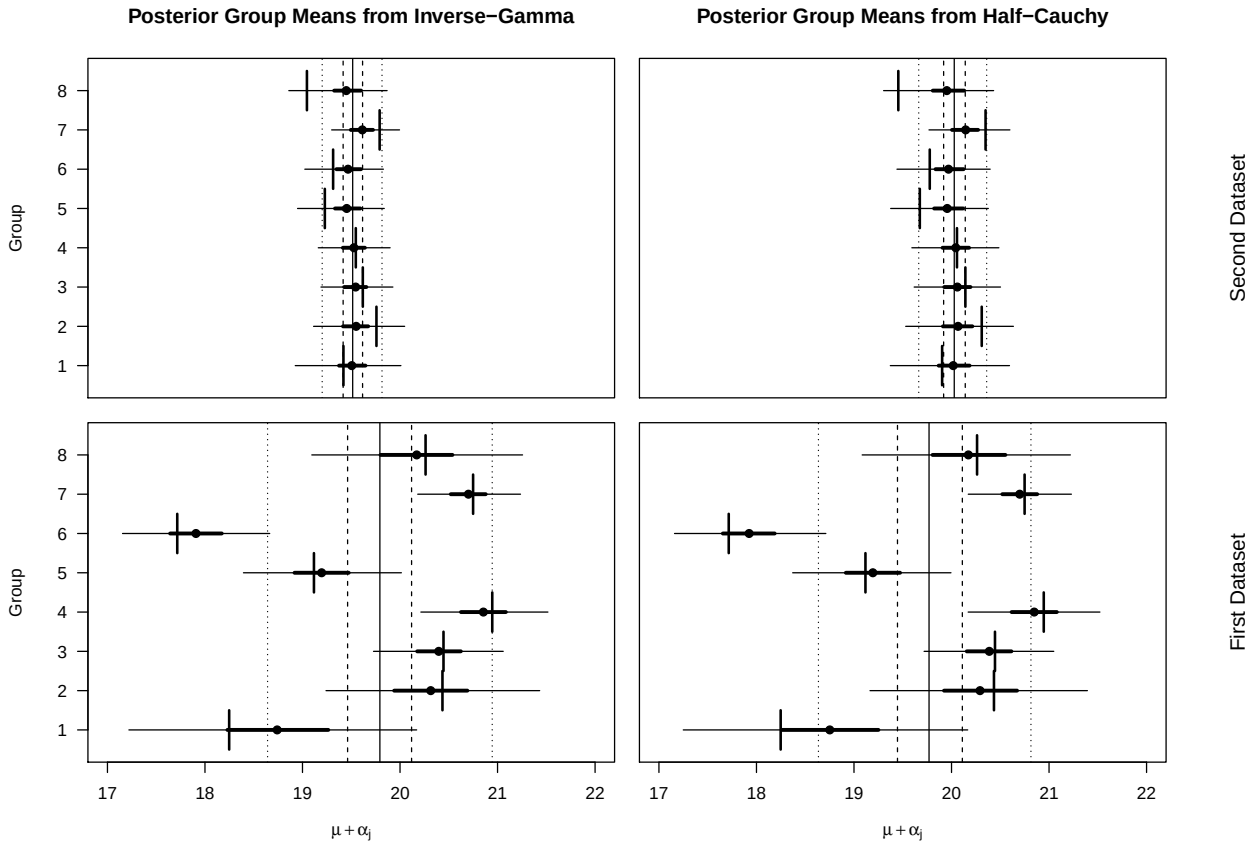


Figure 16: Posterior distributions of the group means for both datasets.

(c) Figure 16 compares the posterior distributions of the group means of both datasets. In both cases, the inverse-Gamma and half-Cauchy priors resulted in nearly identical posteriors. Very little shrinkage is seen for the first dataset. The intervals are centered near the sample averages, away from the overall mean. This is what we expect when the between-group variance is greater than 0.