Hierarchical Model MCMC

Jarell Cheong Tze Wen
December 14, 2022

Abstract. We study a three-level state-space hierarchical model with observation y, the hidden variables (z, I), and parameter $\theta = (T, \gamma, \sigma^2)$, where T is the transition matrix for the Markov chain corresponding to I, $\gamma \in (-1, 1)$, and σ^2 is a vector of variances. Given this model, two approaches are used to infer θ. We make a Gibbs sampler to draw values from the posterior for (z, I, θ), and we also use Monte Carlo EM to estimate the maximum likelihood estimate for θ. Diagnostic plots show that both MCMC algorithms converge.

Contents

1 Problem Statement

Let I_t, $t = 0, 1, 2, \ldots$, be a 3-state Markov chain with states $\{1, 2, 3\}$ and transition matrix $T = (\tau_{ij})$. Let $\sigma^2 = (\sigma_1^2, \ldots, \sigma_4^2)$ be a vector of variances, and let $\gamma \in (-1, 1)$. Starting from $I_0 = 1$ and $z_0 = 0$, consider the following state-space model:

$$\tau_{ij} = \mathbb{P}(I_t = j \mid I_{t-1} = i), \tag{1}$$

$$z_t = \gamma z_{t-1} + \sigma_{I_t} \epsilon_t, \tag{2}$$

$$y_t = z_t + \sigma_4 \eta_t. \tag{3}$$

Here, ϵ_t and η_t are t-independent $\mathcal{N}(0, 1)$ random noise. With the notation

$$y = (y_1, \ldots, y_n), \quad z = (z_0, \ldots, z_n), \quad I = (I_0, \ldots, I_n),$$

we would like to infer the model parameters $\theta = (T, \gamma, \sigma^2)$ when given only y (for this paper, suppose $n = 500$). The dataset of the observed y is in data/obs.csv.

To do this, we'll consider two complementary approaches using Markov chain Monte Carlo. First, we sample from the posterior distribution of (z, I, θ) utilizing a Gibbs sampler for which we have specified the full conditional distributions for. This lets us obtain useful quantities such as the posterior mean and variances of each element of θ. Second, we implement a Monte Carlo EM algorithm to roughly approximate the maximum likelihood estimate for θ (a vanilla EM algorithm or variational EM algorithm seems infeasible to us so far and is a fruitful direction for further work on this model). Along the way, we grapple with the mathematical details of the model, finding the loglikelihood and suggesting reasonable priors. For reference, the true value of θ used to generate the data y is determined by

$$T = \begin{pmatrix} 0.64 & 0.27 & 0.09 \\ 0.10 & 0.85 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{pmatrix}, \quad \gamma = 0.93, \quad \sigma \approx (1, 3, 5, 1).$$

2 Finding the Loglikelihood

To begin exploring this proposed model, we first write down the complete-data loglikelihood function. By the Markov property and the probability chain rule,

$$L(\theta) = p(y, z, I \mid \theta) = \prod_{t=1}^{n} p(y_t, z_t, I_t \mid z_{t-1}, I_{t-1}, \theta)$$

$$= \prod_{t=1}^{n} p(I_t \mid I_{t-1}, \theta) p(z_t \mid z_{t-1}, I_t, \theta) p(y_t \mid z_t, \theta)$$
$$= \prod_{t=1}^{n} \tau_{I_{t-1} I_t} \varphi\left(\frac{z_t - \gamma z_{t-1}}{\sigma_{I_t}}\right) \frac{1}{\sigma_{I_t}} \varphi\left(\frac{y_t - z_t}{\sigma_4}\right) \frac{1}{\sigma_4},$$

where φ is the $\mathcal{N}(0,1)$ density. The last line above follows from (2) and (3), i.e.

$$p(z_t \mid z_{t-1}, I_t, \theta) = \mathcal{N}(\gamma z_{t-1}, \sigma_{I_t}^2), \quad p(y_t \mid z_t, \theta) = \mathcal{N}(z_t, \sigma_4^2).$$

Thus, if we use $\propto_+$ to denote additive proportionality, we find the loglikelihood

$$\ell(\theta) = \sum_{t=1}^{n} \left(\log \tau_{I_{t-1} I_t} + \log \varphi\left(\frac{z_t - \gamma z_{t-1}}{\sigma_{I_t}}\right) - \log \sigma_{I_t} + \log \varphi\left(\frac{y_t - z_t}{\sigma_4}\right) - \log \sigma_4 \right)$$
$$\propto_+ -n \log \sigma_4 + \sum_{t=1}^{n} \left(\log \tau_{I_{t-1} I_t} - \frac{(z_t - \gamma z_{t-1})^2}{2\sigma_{I_t}^2} - \log \sigma_{I_t} - \frac{(y_t - z_t)^2}{2\sigma_4^2} \right).$$

3 The Bayesian Approach

We proceed with a Bayesian approach, setting priors on (T, γ, σ^2) before finding the posterior distribution of (z, I, θ) given the observed data y. Our priors are

$$p(\tau_{i1}, \tau_{i2}, \tau_{i3}) = \mathrm{Dir}(1,1,1), \quad p(\gamma) = \mathrm{Unif}(-1,1), \quad p(\sigma_j^2) = \mathrm{InvGamma}(2,8).$$

These are conveniently chosen conjugate priors which will result in closed-form posteriors, easing further computation (other priors can certainly be explored as well as an avenue for further work on this model). With this, we readily see that

$$p(\gamma) \propto_\times 1, \quad p(T) \propto_\times \prod_{i=1}^{3} \prod_{j=1}^{3} \tau_{ij}^{1-1} = 1,$$

so $p(\theta) \propto_\times p(\sigma^2)$. Therefore, the logposterior distribution for (z, I, θ) is given by

$$\log p(z, I, \theta \mid y) \propto_+ \log p(y, z, I \mid \theta) p(\theta) \propto_+ \ell(\theta) + \sum_{j=1}^{4} \log p(\sigma_j^2).$$

Using the fact that $\log p(\sigma_j^2) \propto_+ -8/\sigma_j^2 + 3 \log \sigma_j^2$, for all (z, I, θ) in the support,

$$\log p(z, I, \theta \mid y) \propto_+ \ell(\theta) + \sum_{j=1}^{4} \left(-\frac{8}{\sigma_j^2} - 3 \log \sigma_j^2 \right).$$

4 Joint Posterior Sampling

We are now ready to design a Gibbs sampler to draw from the posterior of (z, I, θ).

Updating I

To update each component of I, we begin by computing

$$\begin{aligned} p(I_t \mid I_{-t}, y, z, \theta) &= \frac{p(I_t, I_{-t}, y, z, \theta)}{p(I_{-t}, y, z, \theta)} \\ &\propto_\times p(I_t \mid I_{t-1}, T) p(I_{t+1} \mid I_t, T) p(z_t \mid z_{t-1}, I_t, \gamma) \\ &\propto_\times \tau_{I_{t-1} I_t} \tau_{I_t I_{t+1}} \varphi\left(\frac{z_t - \gamma z_{t-1}}{\sigma_{I_t}}\right) \frac{1}{\sigma_{I_t}}. \end{aligned}$$

Now, we can find a probability for each possible value of I_t, yielding a Multinomial distribution on I_t. In other words, for $j = 1, 2, 3$, define p_{tj} piecewise via

$$p_{tj} = \mathbb{1}(1 \le t < n) \tau_{I_{t-1} j} \tau_{j I_{t+1}} \varphi\left(\frac{z_t - \gamma z_{t-1}}{\sigma_j}\right) \frac{1}{\sigma_j} + \mathbb{1}(t = n) \tau_{I_{t-1} j} \varphi\left(\frac{z_t - \gamma z_{t-1}}{\sigma_j}\right) \frac{1}{\sigma_j}.$$

Then, we can set $p'_{tj} = p_{tj} / \sum_{j=1}^3 p_{tj}$ and draw I_t from the Multinomial distribution

$$\text{Mult}_d(1, (p'_{t1}, p'_{t2}, p'_{t3})). \tag{4}$$

Updating z

To update each component of z, we begin by computing

$$\begin{aligned} p(z_t \mid z_{-t}, y, I, \theta) &= \frac{p(z_t, z_{-t}, y, I, \theta)}{p(z_{-t}, y, I, \theta)} \\ &\propto_\times p(z_t \mid z_{t-1}, I_t, \gamma) p(z_{t+1} \mid z_t, I_{t+1}, \gamma) p(y_t \mid z_t, \sigma_4) \\ &= \varphi\left(\frac{z_t - \gamma z_{t-1}}{\sigma_{I_t}}\right) \frac{1}{\sigma_{I_t}} \varphi\left(\frac{z_{t+1} - \gamma z_t}{\sigma_{I_{t+1}}}\right) \frac{1}{\sigma_{I_{t+1}}} \varphi\left(\frac{y_t - z_t}{\sigma_4}\right) \frac{1}{\sigma_4} \\ &\propto_\times \exp\left(-\frac{(z_t - \gamma z_{t-1})^2}{2\sigma_{I_t}^2} - \frac{(z_{t+1} - \gamma z_t)^2}{2\sigma_{I_{t+1}}^2} - \frac{(y_t - z_t)^2}{2\sigma_4^2}\right) \\ &\propto_\times \exp\left(-\frac{(z_t - u_t)^2}{2 s_t^2}\right), \end{aligned}$$

where we can determine u_t and s_t^2 by comparing coefficients of z_t^2 and z_t to get

$$s_t^2 = \frac{1}{\frac{1}{\sigma_{I_t}^2} + \mathbb{1}(1 \le t < n)\frac{\gamma^2}{\sigma_{I_{t+1}}^2} + \frac{1}{\sigma_4^2}}, \quad u_t = \left(\frac{\gamma z_{t-1}}{\sigma_{I_t}^2} + \mathbb{1}(1 \le t < n)\frac{\gamma z_{t+1}}{\sigma_{I_{t+1}}^2} + \frac{y_t}{\sigma_4^2}\right) s_t^2.$$

Therefore, we may readily compute u_t and s_t^2, then draw z_t from the distribution

$$\mathcal{N}(u_t, s_t^2). \tag{5}$$

Updating γ

Subsequently, we can update γ by first computing

$$\begin{aligned} p(\gamma \mid I, y, z, (T, \sigma^2)) &\propto_\times p(z \mid I, y, \theta) \propto_\times \prod_{t=1}^{n} \varphi\left(\frac{z_t - \gamma z_{t-1}}{\sigma_{I_t}}\right)\frac{1}{\sigma_{I_t}} \\ &\propto_\times \prod_{t=1}^{n} \exp\left(-\frac{(z_t - \gamma z_{t-1})^2}{2\sigma_{I_t}^2}\right) = \exp\left(\sum_{t=1}^{n} -\frac{(z_t - \gamma z_{t-1})^2}{2\sigma_{I_t}^2}\right) \\ &\propto_\times \exp\left(-\frac{(\gamma - \ell)^2}{2o^2}\right), \end{aligned}$$

where we can determine ℓ and o^2 by comparing coefficients of γ^2 and γ to get

$$o^2 = \left(\sum_{t=1}^{n} \frac{z_{t-1}^2}{\sigma_{I_t}^2}\right)^{-1}, \quad \ell = \left(\sum_{t=1}^{n} \frac{z_{t-1} z_t}{\sigma_{I_t}^2}\right) o^2.$$

Since our prior on γ has density only on $(-1, 1)$, we can draw γ from a truncated Normal distribution on the support $(-1, 1)$ with ℓ and o^2 found as above, i.e.

$$\mathcal{N}_{(-1,1)}(\ell, o^2). \tag{6}$$

Updating T

Letting $\tau_{i\cdot}$ denote $(\tau_{i1}, \tau_{i2}, \tau_{i3})$, we can update T by computing

$$p(\tau_{i\cdot} \mid I, y, z, (\gamma, \sigma^2)) \propto_\times p(I \mid T) p(\tau_{i\cdot}) \propto_\times \prod_{j=1}^{3} \tau_{ij}^{c_{ij}},$$

where $c_{ij} = \sum_{t=1}^{n} \mathbb{1}(I_{t-1} = i, I_t = j) = \sum_{t=1}^{n} \mathbb{1}(I_{t-1} = i)\mathbb{1}(I_t = j)$. We draw T from

$$\mathrm{Dir}(c_{i1} + 1, c_{i2} + 1, c_{i3} + 1). \tag{7}$$

Updating σ^2

Finally, to update σ^2, notice that for $j = 1, 2, 3$, we have

$$p(\sigma_j^2 \mid I, y, z, (\sigma_{-j}^2, \gamma, T)) \propto_{\times} p(z \mid I, \theta) p(\sigma_j^2),$$

so the logposterior $\log p(\sigma_j^2 \mid I, y, z, (\sigma_{-j}^2, \gamma, T))$ is additively proportional to

$$-\frac{8}{\sigma_j^2} - 3 \log \sigma_j^2 + \sum_{t=1}^{n} \mathbb{1}(I_t = j) \left(-\frac{1}{2} \log \sigma_j^2 - \frac{(z_t - \gamma z_{t-1})^2}{2\sigma_j^2} \right).$$

Pattern-matching to the inverse Gamma logdensity, we therefore sample σ_j^2 from

$$\text{InvGamma}\left(\frac{4 + \sum_{t=1}^{n} \mathbb{1}(I_t = j)}{2}, \frac{16 + \sum_{t=1}^{n} \mathbb{1}(I_t = j)(z_t - \gamma z_{t-1})^2}{2} \right) \tag{8}$$

for $j = 1, 2, 3$. By an analogous argument, we sample σ_4^2 from

$$\text{InvGamma}\left(\frac{4 + n}{2}, \frac{16 + \sum_{t=1}^{n} (y_t - z_t)^2}{2} \right). \tag{9}$$

Identification and Initial Values

With (4), (5), (6), (7), (8), and (9), we have specified the conditional distributions required to implement a Gibbs sampler. We initialize our Gibbs sampler at

$$T = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}, \quad \gamma = 1, \quad \sigma^2 = (8, 8, 8, 8),$$

which turn out to work well empirically. Before implementation, note that we do have an identification issue with the σ_j^2 in the model: we solve this by imposing

$$\sigma_1^2 \leq \sigma_2^2 \leq \sigma_3^2.$$

In other words, when we sample for σ^2, we additionally perform rejection sampling until the σ^2 satisfies the constraint above. Find our implementation at:

https://github.com/Jarell-Cheong/hierarchical-mcmc

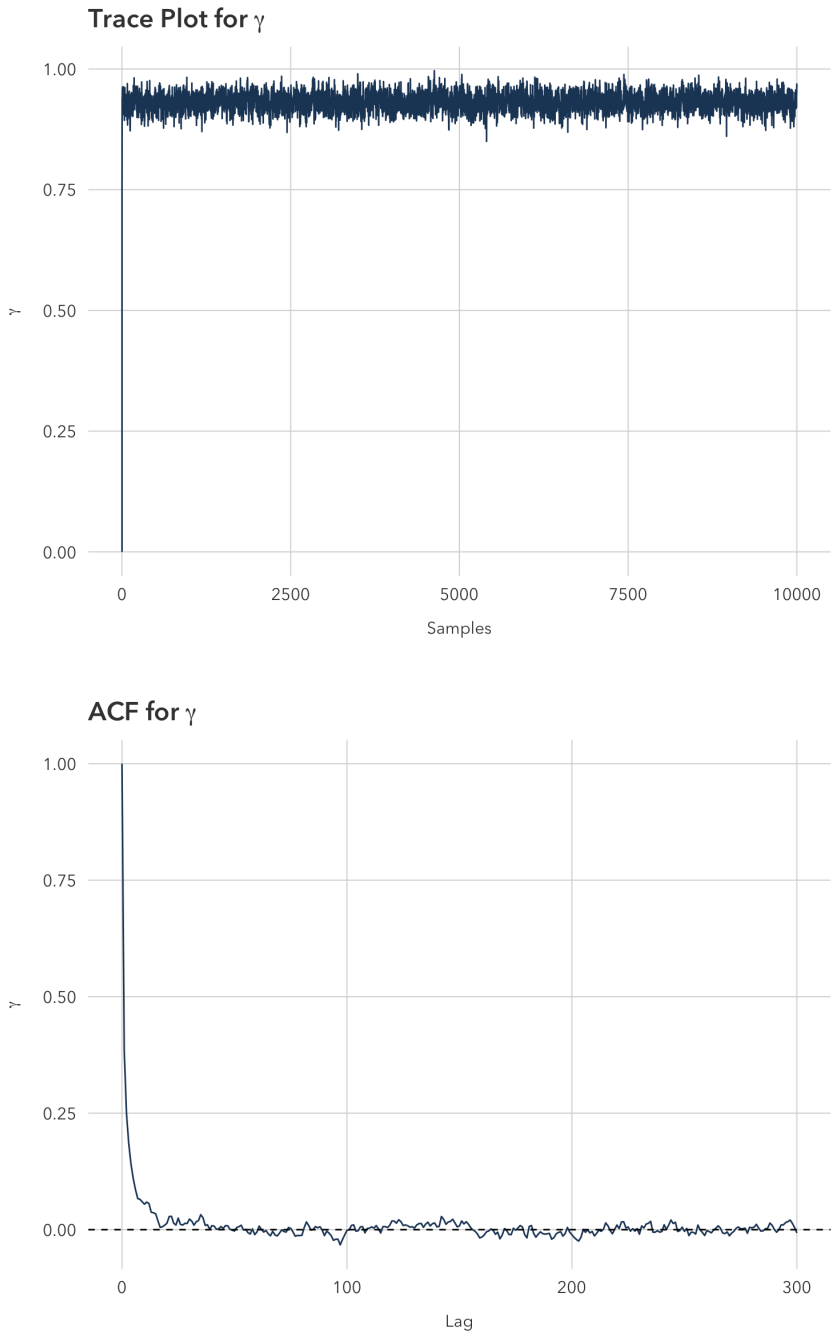

Figure 1: Trace and ACF plots for γ.

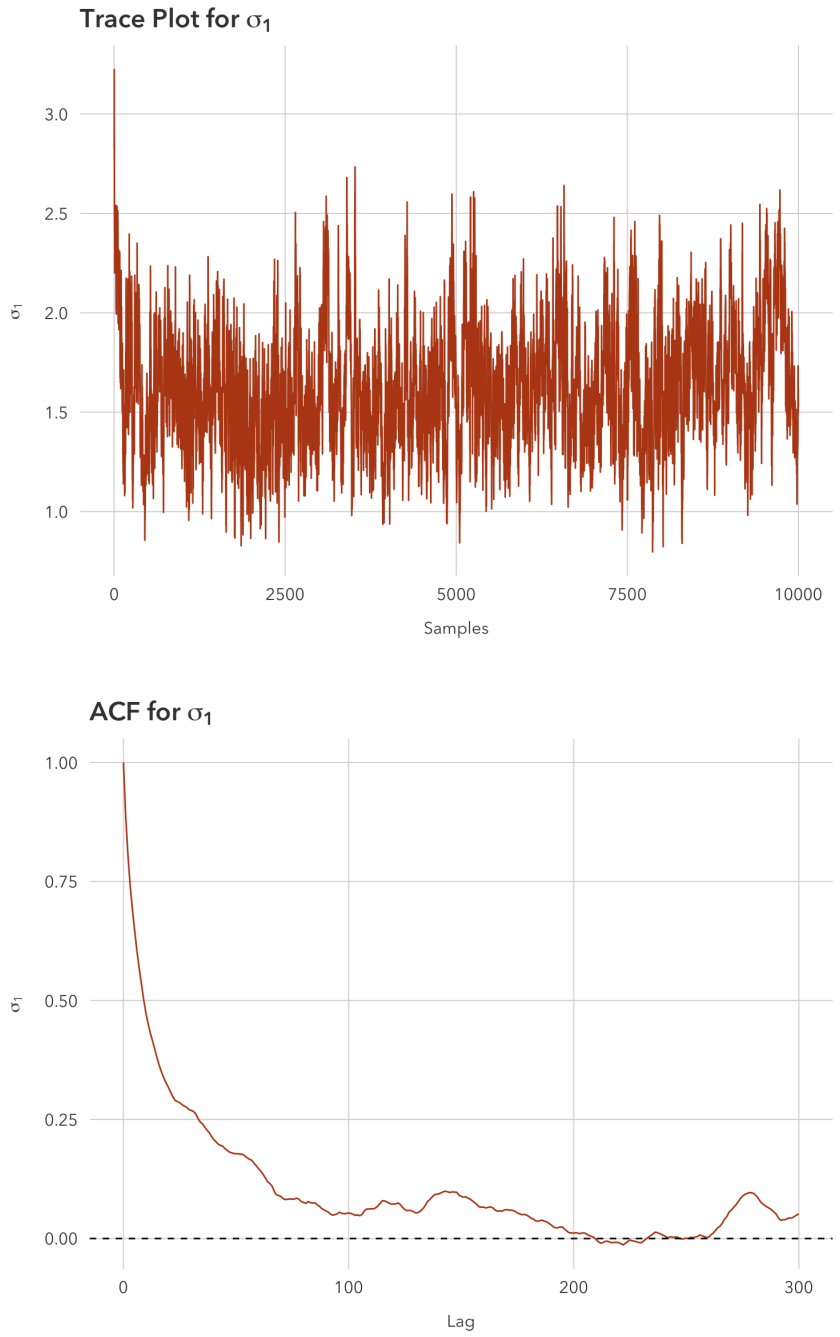

Figure 2: Trace and ACF plots for σ_1.

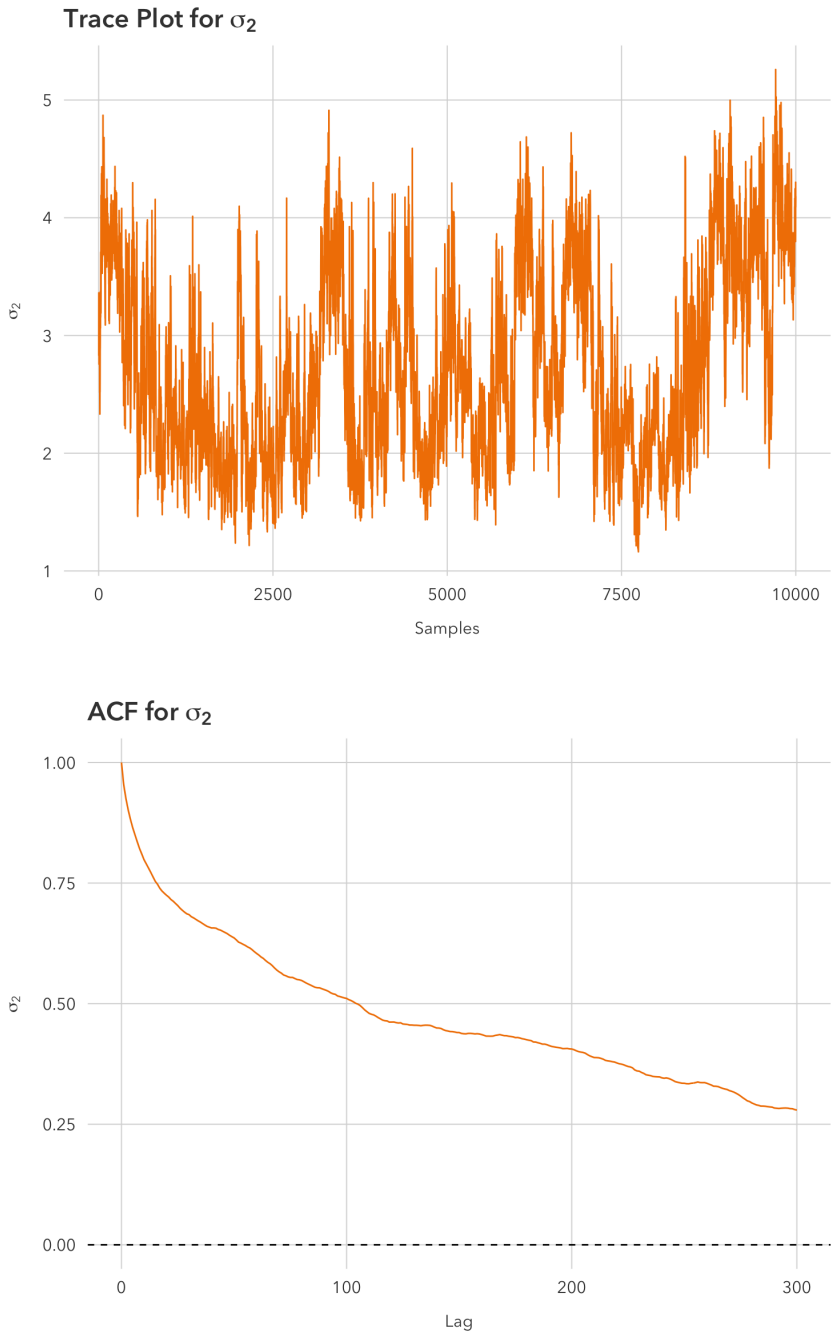

Figure 3: Trace and ACF plots for σ_2.

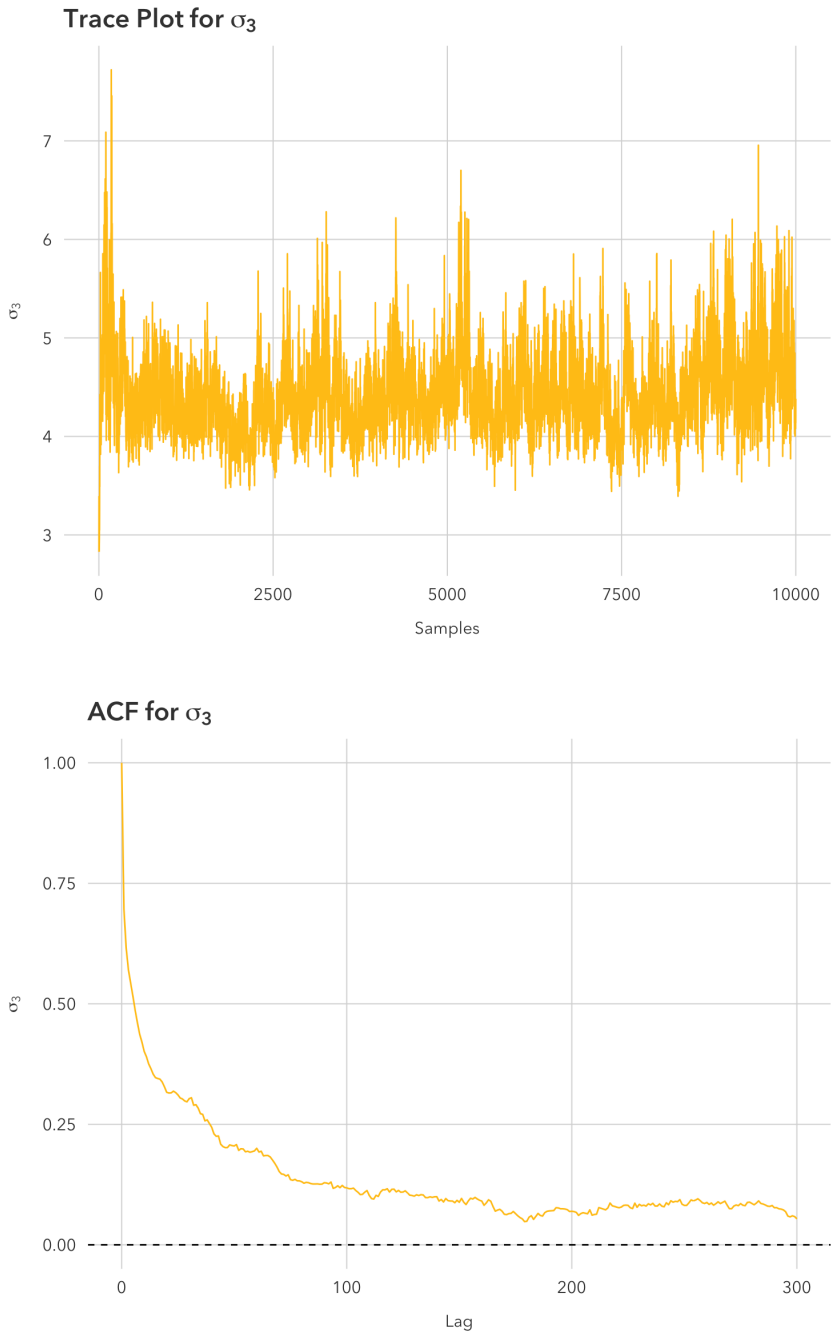

Figure 4: Trace and ACF plots for σ_3.

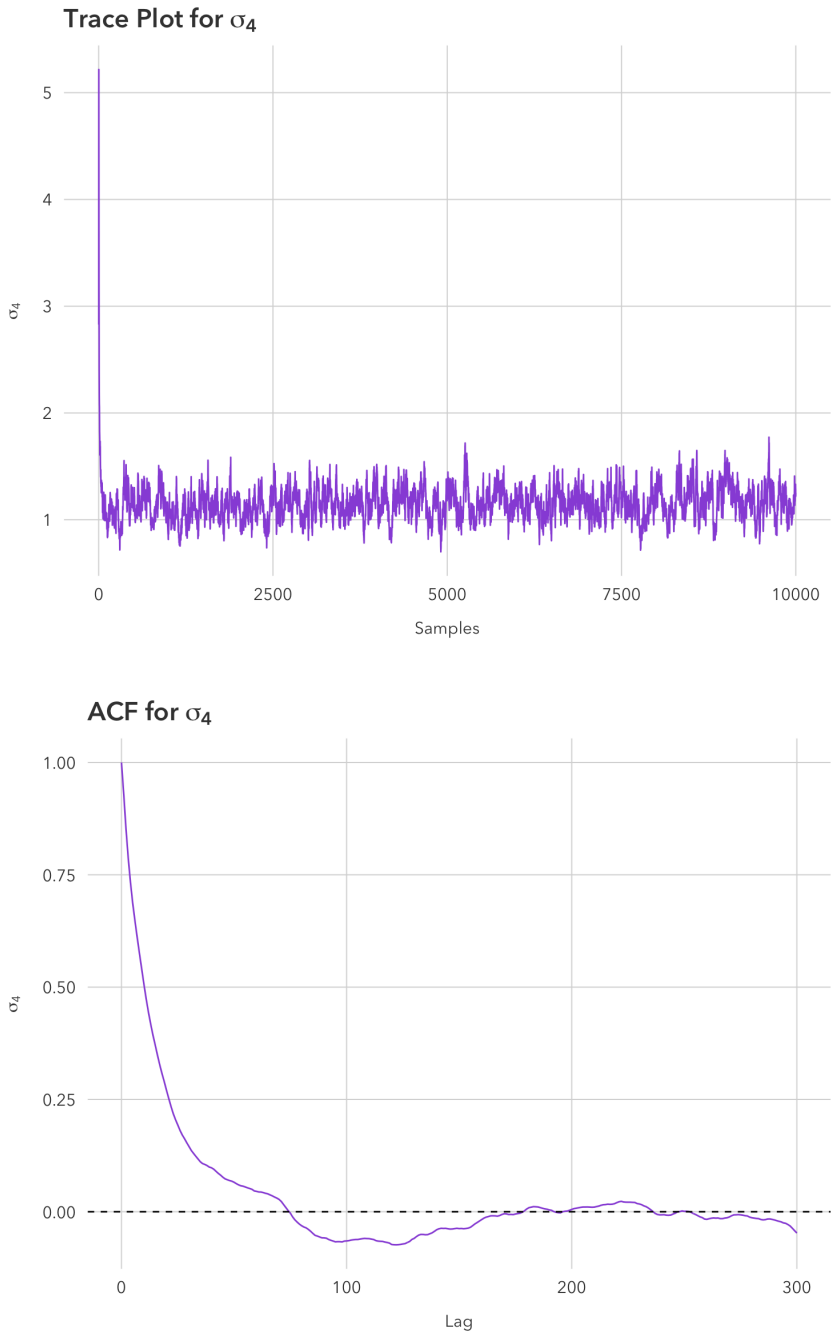

Figure 5: Trace and ACF plots for σ_4.

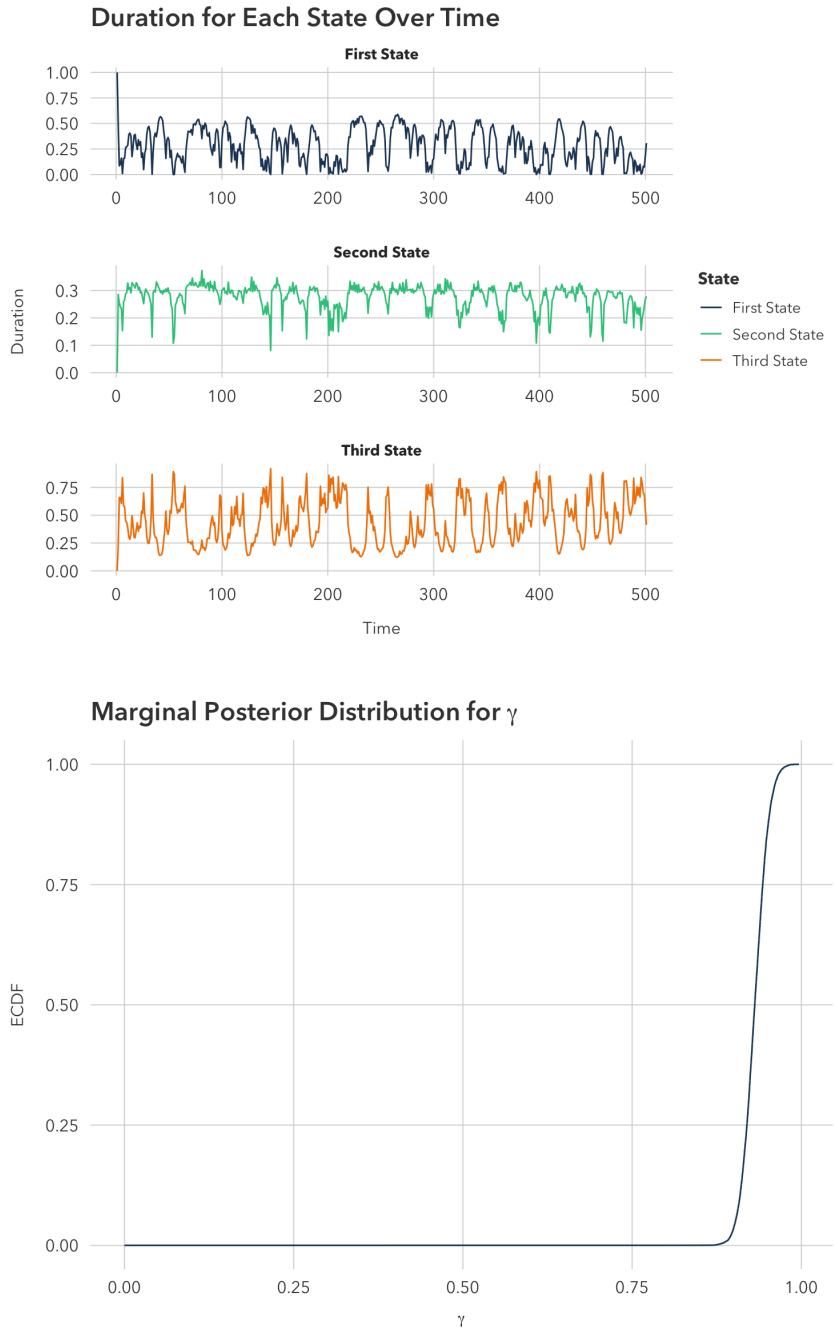

Figure 6: Duration of states and ECDF plot for γ.

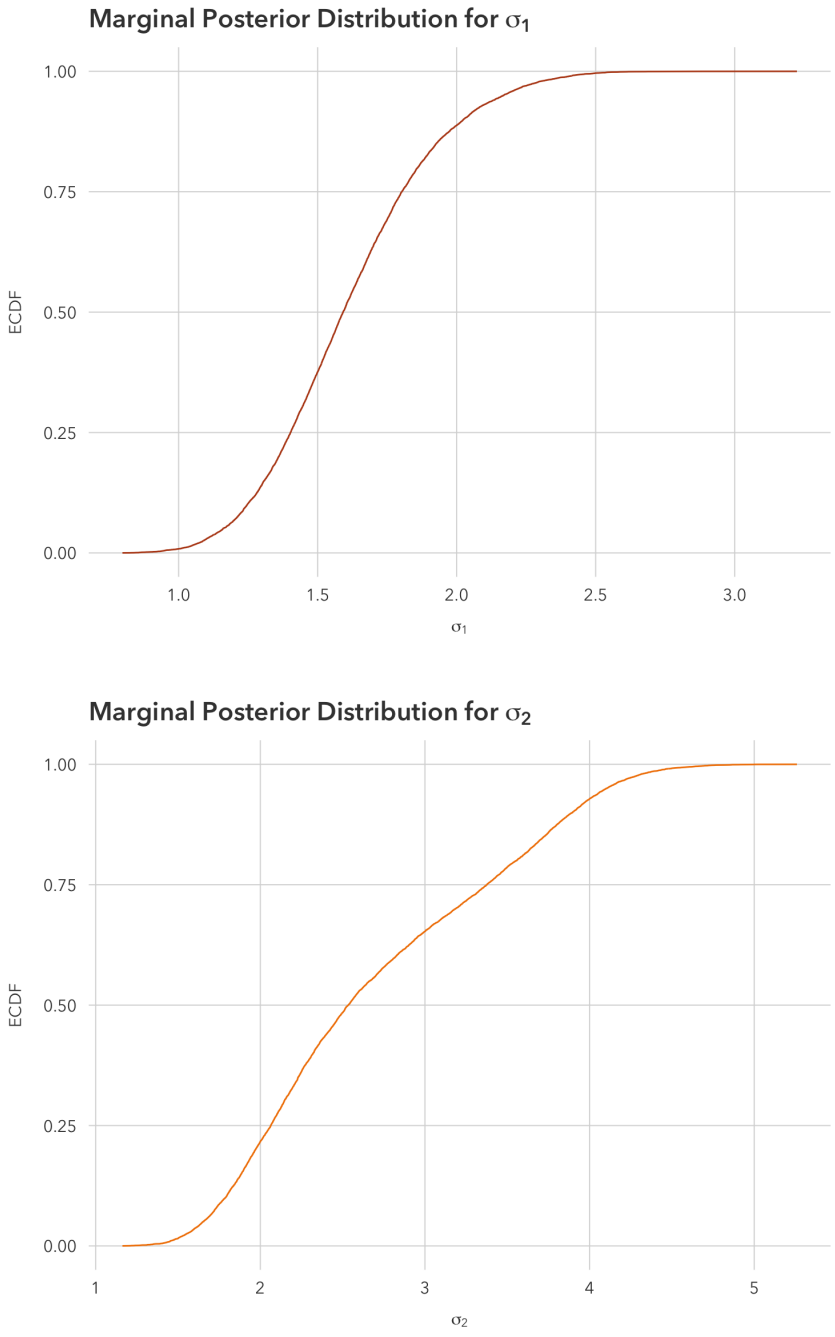

Figure 7: ECDF plots for σ_1 and σ_2.

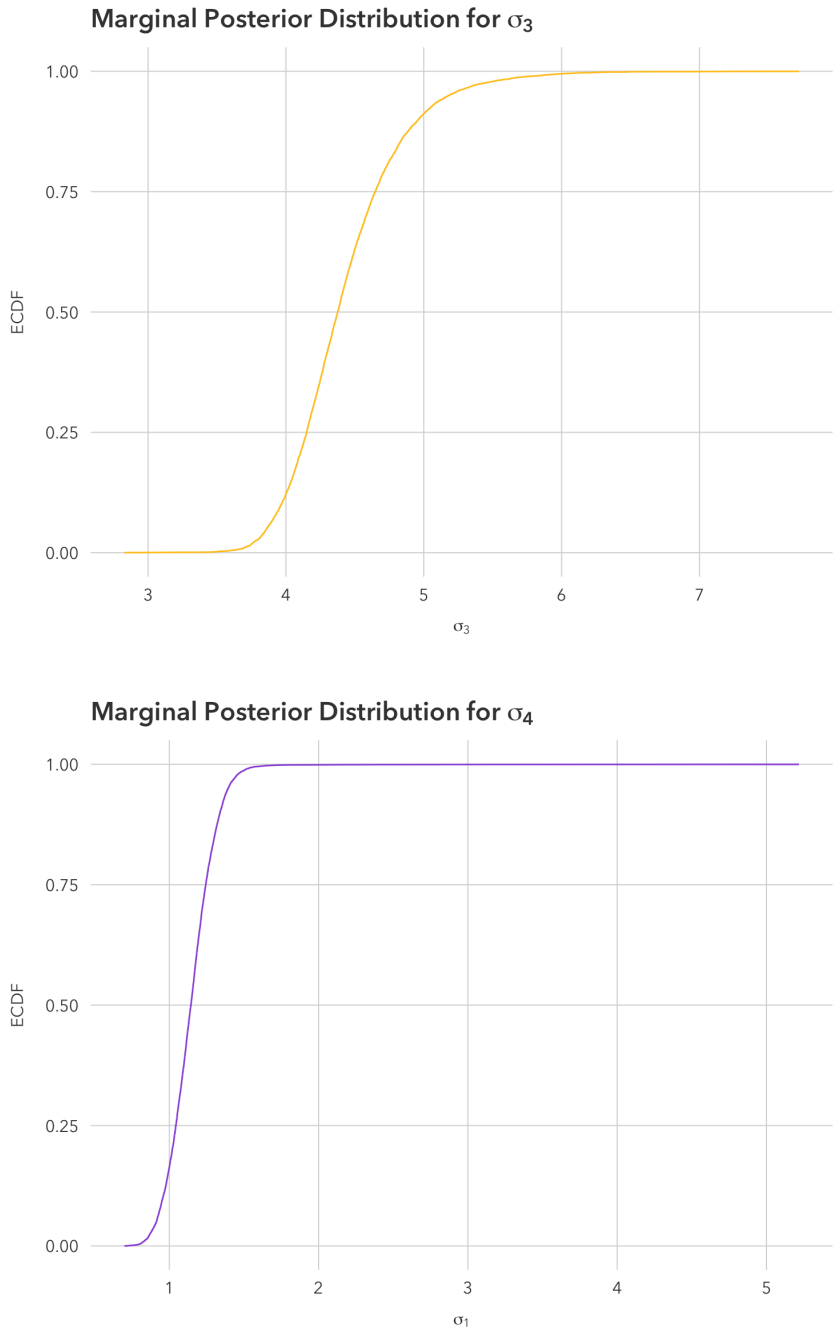

Figure 8: ECDF plots for σ_3 and σ_4.

Diagnostic Plots and θ Inference

Trace and ACF plots are provided for γ and σ in Figures 1, 2, 3, 4, and 5. The trace and ACF plots clearly show the convergence of our MCMC chain, which is much better, in particular, for γ. This is indeed expected from a Gibbs sampler, and our results should be much better than if we used the Metropolis-Hastings algorithm. Posterior means and standard deviations for each parameter in θ are as follows:

Parameter	Posterior Mean	Posterior Standard Deviation
T	$\begin{pmatrix} 0.503 & 0.287 & 0.210 \\ 0.323 & 0.381 & 0.295 \\ 0.116 & 0.232 & 0.653 \end{pmatrix}$	$\begin{pmatrix} 0.244 \\ 0.232 \\ 0.301 \end{pmatrix}$
γ	0.931	0.020
σ	$(1.616, 2.714, 4.431, 1.149)$	$(0.302, 0.797, 0.421, 0.164)$

Observe that the mean of γ is remarkably close to the true value, and that the σ values are also decently close to the true value, but less so because they possess a higher variance. Meanwhile, our predictions for T are the furthest away from the true T value, but this makes sense as T is much more removed from the data y so that we have much weaker information when inferring the value of T. Plus, marginal posterior distributions (ECDF plots) for γ and σ, as well as duration of state plots for the Markov chain I_t, are provided in Figures 6, 7, and 8.

5 Computing the MLE

To find an approximate MLE for θ, we propose a Monte Carlo EM algorithm.

E Step

We will approximate the Q function with m Monte Carlo samples $\{z^{(k)}, I^{(k)}\}_{k=1}^{m}$, each drawn from the posterior $p(z, I \mid y, \theta^{(t)})$. We can easily draw this using the Gibbs sampler we implemented before, where the conditional distribution is the same as what we found in Section 4. With this, our Q function is approximately

$$\frac{1}{m}\sum_{k=1}^{m}\left(-n\log\sigma_4 + \sum_{t=1}^{n}\left(\log\tau_{I_{t-1}^{(k)}I_t^{(k)}} - \frac{(z_t^{(k)} - \gamma z_{t-1}^{(k)})^2}{2\sigma^2_{I_t^{(k)}}} - \log\sigma_{I_t^{(k)}} - \frac{(y_t - z_t^{(k)})^2}{2\sigma_4^2}\right)\right).$$

M Step

It is also not too difficult to find the value of θ which maximizes the Q function above. For τ_{ij}, the MLE is simply the sample mean, so the optimal τ_{ij} is

$$\left(\sum_{j=1}^{3}\sum_{k=1}^{m}\sum_{t=1}^{n}\mathbb{1}(I_{t-1}^{(k)}=i)\mathbb{1}(I_{t}^{(k)}=j)\right)^{-1}\sum_{k=1}^{m}\sum_{t=1}^{n}\mathbb{1}(I_{t-1}^{(k)}=i)\mathbb{1}(I_{t}^{(k)}=j).$$

For γ and σ^2, we perform an iterative optimization. Given γ, the MLE of σ^2 is the MLE of the variance in a linear regression. That is, the optimal σ_j^2 given γ is

$$\left(\sum_{k=1}^{m}\sum_{t=1}^{n}\mathbb{1}(I_{t}^{(k)}=j)\right)^{-1}\sum_{k=1}^{m}\sum_{t=1}^{n}\mathbb{1}(I_{t}^{(k)}=j)(z_{t}^{(k)}-\gamma z_{t-1}^{(k)})^2$$

for $j=1,2,3$ and $(mn)^{-1}\sum_{k=1}^{m}\sum_{t=1}^{n}(y_t-z_t^{(k)})^2$ for $j=4$. Subsequently, given σ^2, the MLE of γ is found by setting partials of Q to zero: the optimal γ given σ^2 is

$$\left(\sum_{k=1}^{m}\sum_{t=1}^{n}\frac{z_{t-1}^{2(k)}}{\sigma^2_{I_t^{(k)}}}\right)^{-1}\sum_{k=1}^{m}\sum_{t=1}^{n}\frac{z_t^{(k)}z_{t-1}^{(k)}}{\sigma^2_{I_t^{(k)}}}.$$

Diagnostic Plots and the MLE of θ

Trace and ACF plots are provided for γ and σ in Figures 9, 10, 11, 12, and 13. The trace and ACF plots here also clearly show the convergence of our MCMC chain. The approximate MLE that we obtained from this algorithm is determined by

$$\hat{T}=\begin{pmatrix}0.490 & 0.301 & 0.209\\ 0.537 & 0.386 & 0.077\\ 0.023 & 0.112 & 0.865\end{pmatrix},\quad \hat{\gamma}=0.936,\quad \hat{\sigma}=(0.904,2.465,4.196,1.441).$$

Here too, γ and σ approach the true value rather well, but T approaches the true value less closely than in our Gibbs sampler. Of course, this last difference is to be expected because our Monte Carlo EM approximates the MLE of θ, while the Gibbs sampler was constructed to find the posterior mean of θ. Once again, our implementation of this Monte Carlo EM algorithm, along with other components of this project (such as all the figures), lives in the following GitHub repository:

https://github.com/Jarell-Cheong/hierarchical-mcmc

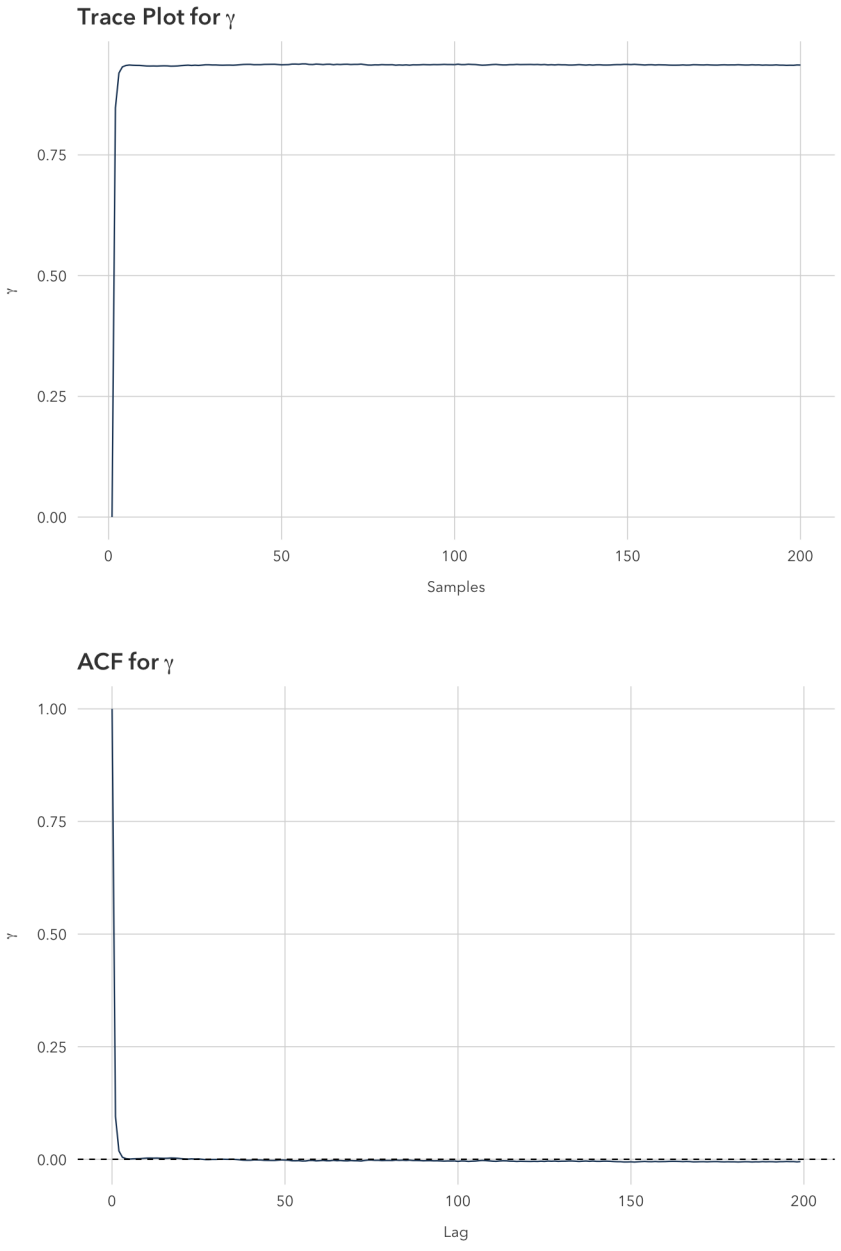

Figure 9: Trace and ACF plots for γ.

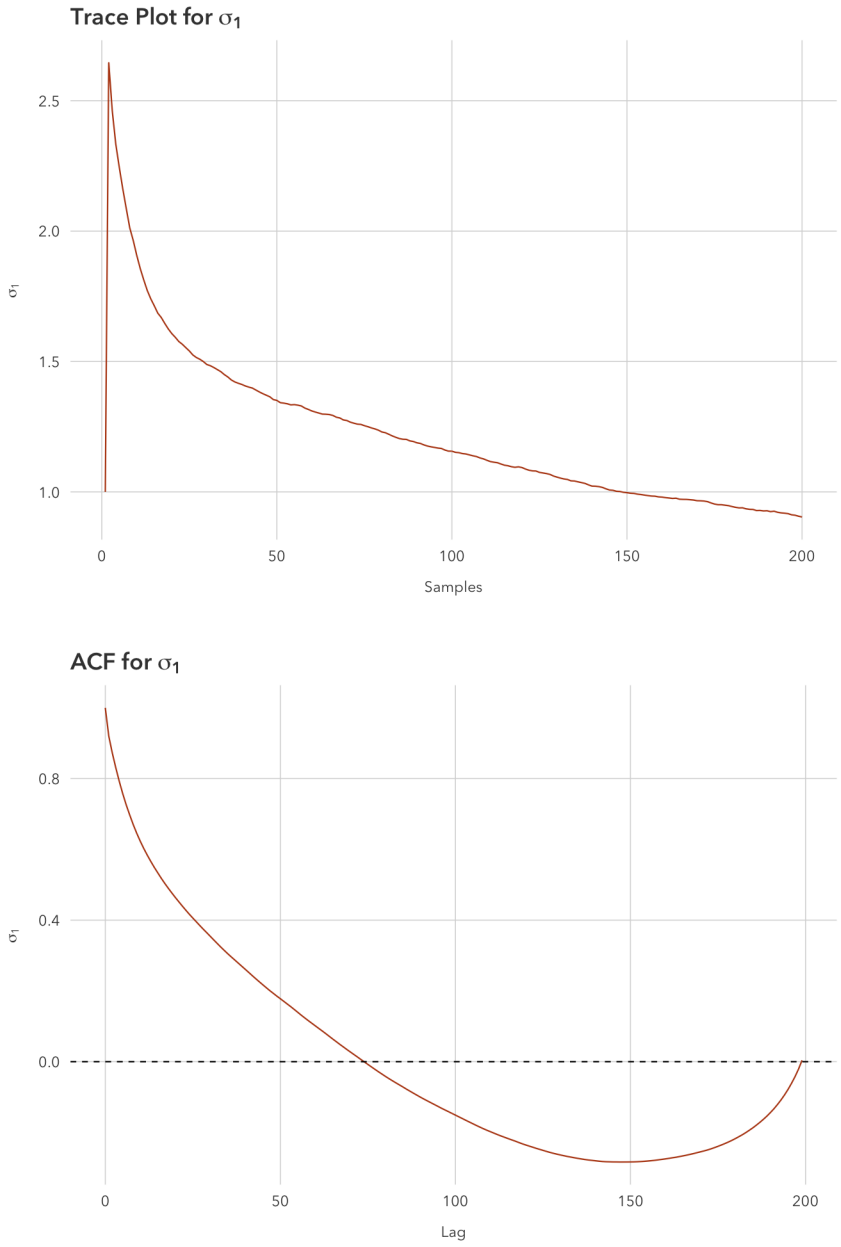

Figure 10: Trace and ACF plots for σ_1.

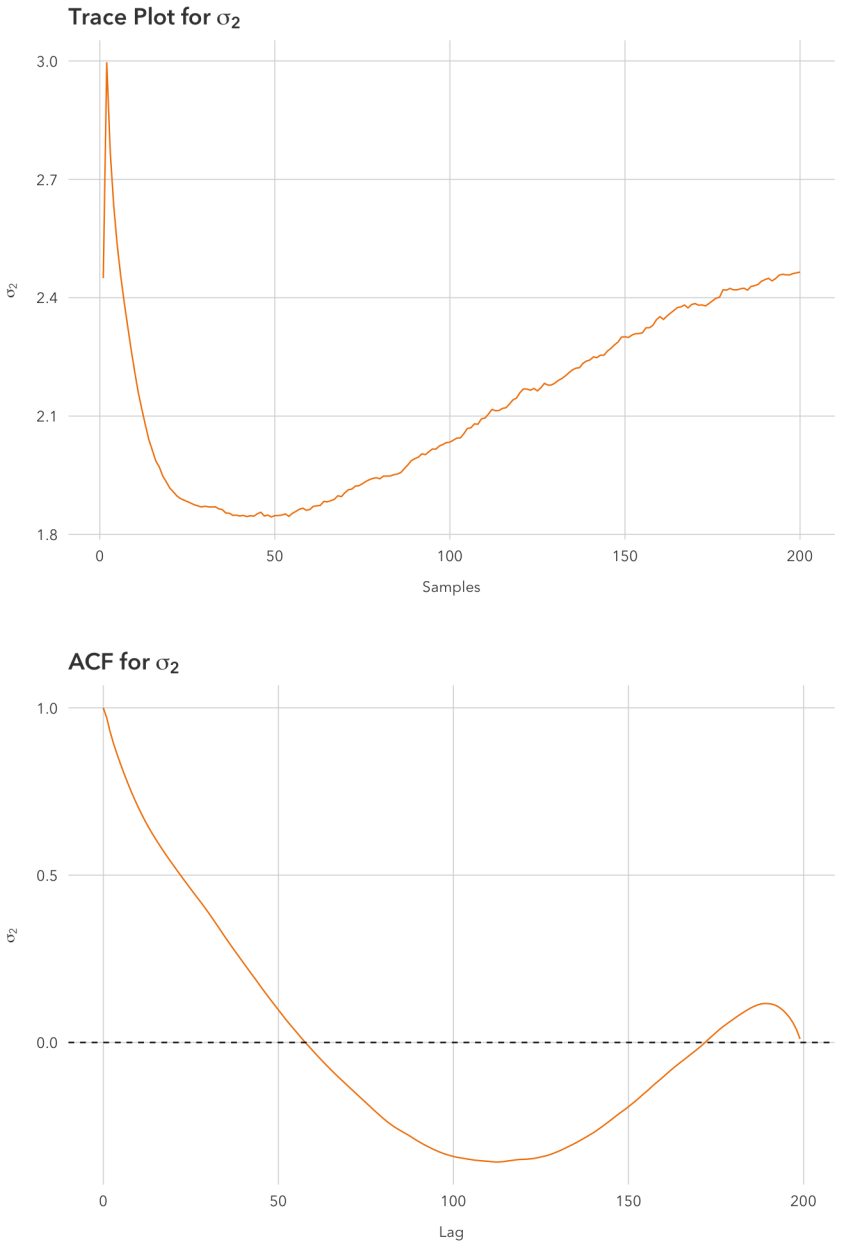

Figure 11: Trace and ACF plots for σ_2.

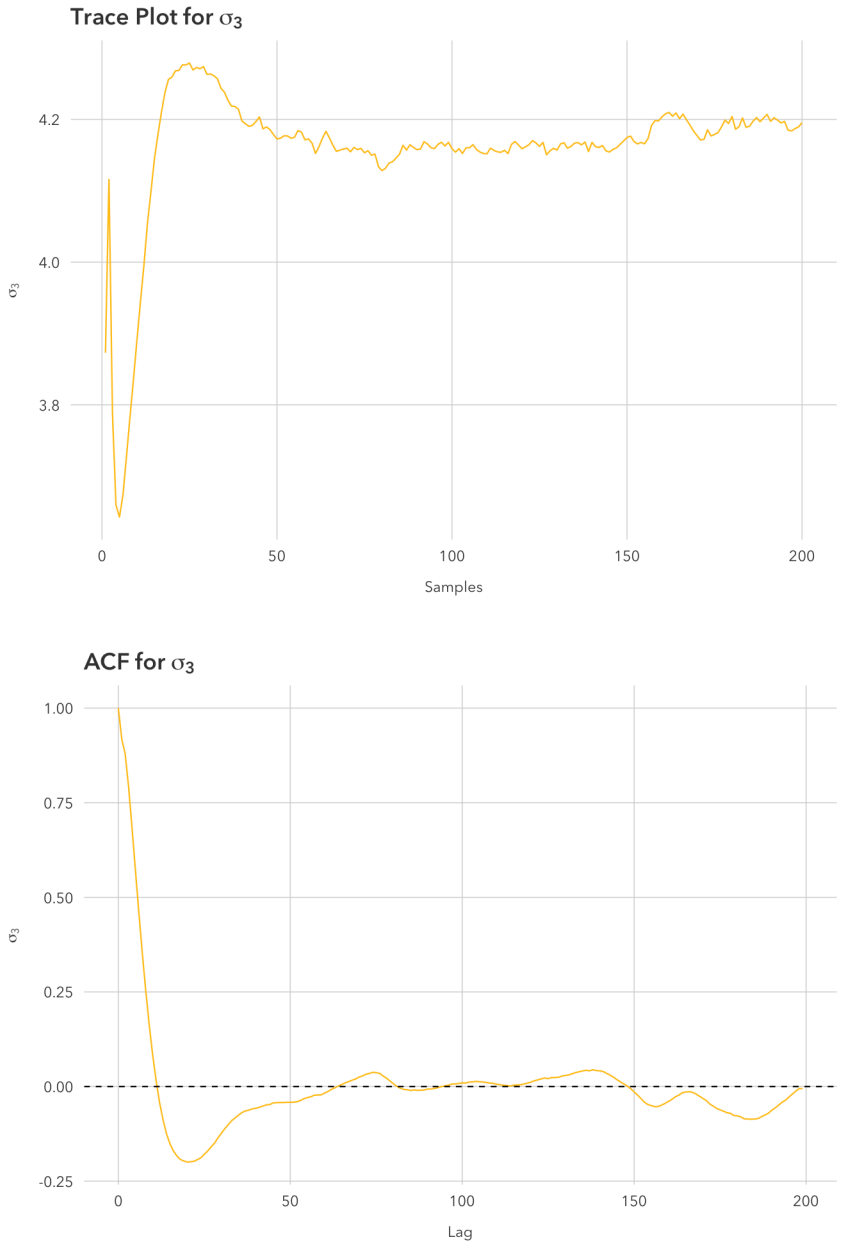

Figure 12: Trace and ACF plots for σ_3.

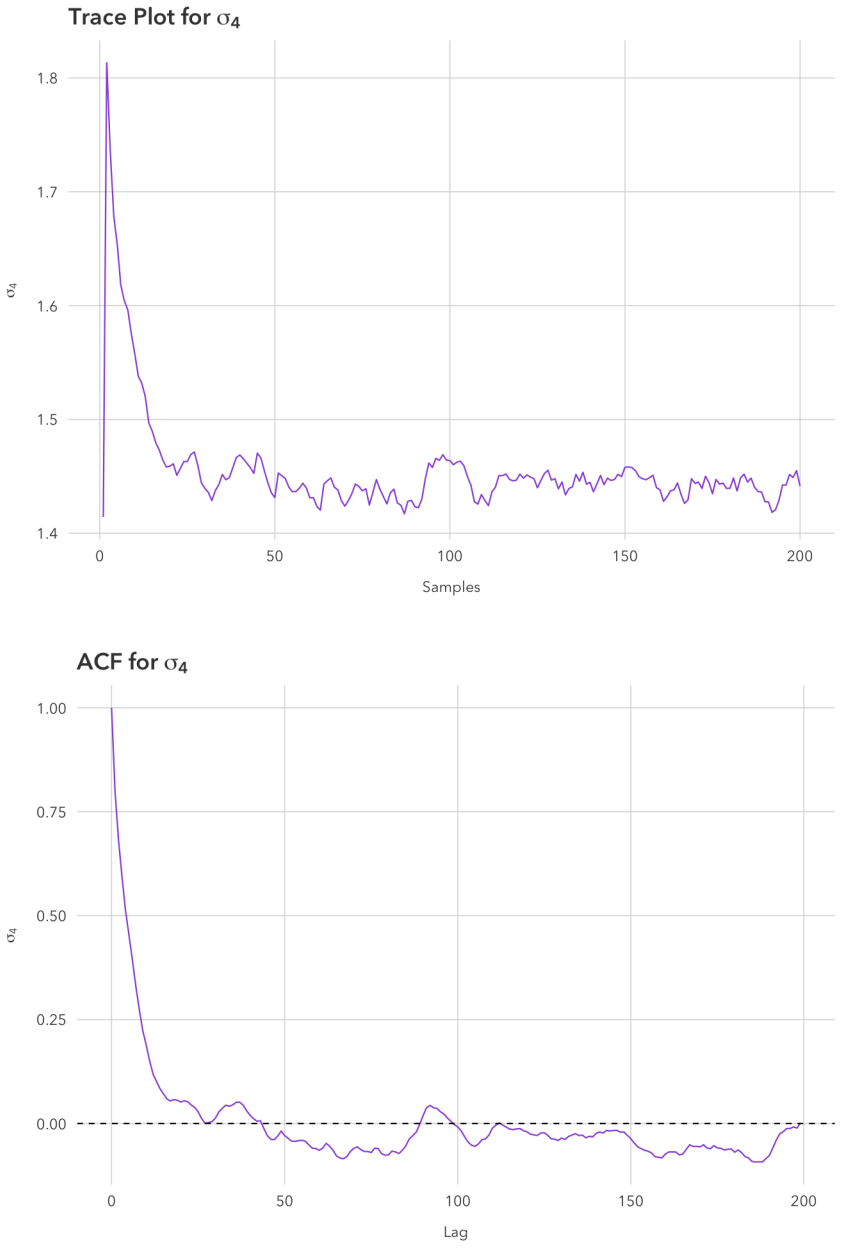

Figure 13: Trace and ACF plots for σ_4.