STAT 656: Bayesian Data Analysis
Fall 2024
Homework 1

Note: For each question, no credit will be given unless work is shown.

Synthetic data

40 points

The *autoregressive model* is frequently used to analyze time series data. The simplest autoregressive model has order 1, and is abbreviated as AR(1). This model assumes that an observation y_i at time point i $(i = 1, \ldots, n)$ is generated according to

$$y_i = \rho y_{i-1} + \epsilon_i,$$

where $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$ independently, and ρ and σ are unknown parameters. For simplicity, we shall assume that y_0 is a fixed constant. We will also assume $|\rho| < 1$.

1. (5 points) Write the log-likelihood function $\log L(\rho, \sigma^2 | y_0, y_1, \ldots, y_n)$ for $(\rho, \sigma^2)^\mathsf{T}$ for the AR(1) model.

   For rest of this problem, we shall take $\rho, \log \sigma$ to be our estimands of interest, and consider data in the file `computation_data_hw_1.csv`, generated from this type of process with $y_0 = 0$.

2. (10 points) Write an R function that computes the log of the likelihood functions for $(\rho, \log(\sigma))^\mathsf{T}$ for this data. Specifically, this function should take the form:

```
ar_loglik <- function(rho,log_sig) {
# Do stuff with data loaded from computation_data_hw_1.csv
# You can make the data also an input to the function,
#   or treat it as a global variable

   ....

# Returns the log-likelihood of the data for the input (rho,log_sig) pair
   return(log_lik)
}
```

   Provide a visualization of this log-likelihood as a contour plot. Hint: The `outer` and `contour` function in `R` can be useful for creating the visualization, see also the code of lectures 2 and 3.

3. (10 points) For the purposes of this problem, suppose we specify $\rho \sim \mathrm{Uniform}(-1, 1)$, $\log(\sigma) \sim \mathrm{N}(0, 10^2)$ independently *a priori* (note that this may not be an appropriate prior for the parameters of an AR(1) model in general). Write an R function that computes the log of the posterior density (upto a constant) for $(\rho, \log(\sigma))^\mathsf{T}$ under this prior. Provide a visualization of this function as above. How does this function compare to the log-likelihood function? Would you say that this prior specification is overly informative? Why or why not?

4. (10 points) Draw 1000 values of $(\rho, \log(\sigma))^\mathsf{T}$ from a discrete grid approximation to the posterior. Be sure to describe your choice of discrete grid. Hint: The previous step can be helpful in this regard, together with the R function `sample`. Again, look at the code associated with the lectures.

5. (5 points) Use these draws to calculate the following summaries for each of ρ and $\log(\sigma)$: 0.025, 0.25, 0.5, 0.75, 0.975 quantiles, mean, standard deviation, skewness, and kurtosis. You can use the library `moments` if you want.

6. (10 points) Write an R function that takes parameters $(\rho, \log(\sigma))^\mathsf{T}$ and simulates a new dataset y^{rep} according to the AR process. Recall that we can simulate from the posterior predictive distribution of new datasets y^{rep} given todays dataset as follows: first simulate a parameter set from the posterior distribution, and use this to simulate a new dataset. Use your two earlier R functions to generate 1000 such posterior predictive samples. Summarize your draws.

7. (10 points) Compare the observed data to these posterior predictive summaries. Also create a plot where the observed data is suporposed on these posterior predictive trajectories. What can you say about the model fit? Does the model appear appropriate?

Real data

60 points

Here we will use your functions from the previous question to determine the utility of the AR(1) model for predicting COVID-19 cases and deaths. Specifically, you will analyze two sets of data that contain the cases and deaths from COVID-19 in the United States overall, and for individual states, respectively. The data on cases and deaths in the entire United States are in `covid_us.txt`, and the data on cases and deaths for three individual states are in `covid_us-states.txt`. Both datasets were downloaded from the *New York Times' Coronavirus (COVID-19) Data in the United States* Github page, which contains further information and details regarding these data.

1. (5 points) Do you believe that the information given on the Github page is sufficient for analyzing the data? Why or why not? If not, what other information would you have requested that the New York Times provide, or what other information would you have collected if you had the resources to do so?

2. (5 points) Specify an AR(1) model for the data on the entire United States (not any of the individual states). In particular, would you model the raw data as an AR(1) process, our would you transform it first in some way? Explain.

3. (5 points) Specify a prior for all of the model parameters. You can choose any prior you like, but you **must** justify your choice.

4. (10 points) Fit your AR(1) model to data from the first time point until June 30. Calculate and provide a visualization of the posterior distribution of the estimand. Summarize the posterior distribution using moments as before.

5. (10 points) Conduct a posterior predictive check to diagnose the fit of your model for the data until June 30. Does your model provide good predictions? Why or why not? If not, describe how you would consider modifying it to address the observed inadequacies (you don't have to actually do this).

6. (10 points) Simulate posterior predictions for the cases and deaths in the United States from July 1 until August 25. Compare these posterior predictions to the actual values of the two variables, and comment on the quality of the AR(1) model for predicting future cases and deaths in this context.

7. (10 points) Now replicate the analyses you performed in step 4 on the entire United States for the state of Indiana individually. Discuss if you think the results agree with the results of the entire United States.

8. (5 points) Provide a non-technical explanation of your findings for an audience with minimal statistical training. Specifically, describe the AR(1) model in lay terms, explain whether or not this model provides good predictions in this context, and provide intuitive justifications for why the AR(1) model succeeds or fails in this context.

Feedback

Brief comments on each of these points would be greatly appreciated.

(a) Does the instructor present material at an adequate pace during lecture (too slow/too fast)?

(b) Does the instructor adequately address questions raised in class?

(c) What general material would you like the instructor to spend more time on?

(d) Are the homework questions generally representative of material covered in lecture?

(e) Please name one concept that you are struggling with.

(f) Which topics/ideas/concepts in lecture were not well-explained? Brief comments are appreciated.

(g) Any further comments/questions/feedback?