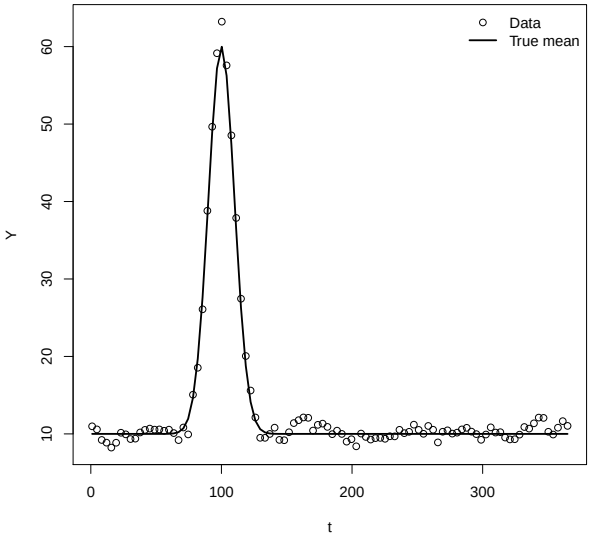
ST440 – Final exam – Due 5/10

THIS IS AN EXAM - DO NOT DISCUSS THE PROBLEM WITH ANYONE (INCLUDING OTHER STUDENTS OR THE TA)! If you have questions, please email me.

In this problem you will conduct a simulation study to investigate the performance of Bayesian non-linear regression. Since this is a simulation study, there is no real data, but the analysis is motivated by the following problem. Let Y_t be the measured pollen count on day $t \in [1, 365]$. Here is a simulated dataset (simulation code below) of daily pollen counts across one year.



Pollen counts follow a seasonal pattern, and so we consider the following non-linear regression model: $Y_t = \mu_t + \varepsilon_t$ where

$$\mu_t = b_1 + b_2 \exp\left\{-\frac{(t-b_3)^2}{2b_4^2}\right\}$$

and the errors ε_t are normal with mean zero, variance σ^2 and correlation $\text{Cor}(\varepsilon_t, \varepsilon_{t+h}) = \rho^h$. The mean curve is determined by four parameters: b_1 is the baseline mean, b_2 is the increase in the mean at the peak of pollen season, $b_3 \in (0, 365)$ is the day of the year with the highest mean and $b_4 > 0$ controls the width of the peak. The objective of the analysis is to estimate $b_1, ..., b_4$.

Conduct a simulation study to test the performance of a Bayesian analysis of this model. Select several (2-4) values of the sample size n and several values of the true correlation parameter ρ, and for each combination of these parameters (let's call a combination, say $n = 100$ and $\rho = 0.9$, a "setting"), generate 100 datasets following the code below (use the same values for all parameters other than n and ρ). For each dataset, fit the model described above except with $\rho = 0$ so the errors are assumed to me independent across time. For each setting, report the bias, mean squared error and coverage of 95% credible intervals for each of the four parameters $b_1, ..., b_4$ (separately so there will be 4*(# settings) biases, etc.). Also, report the Monte Carlo standard errors of the bias, mean square error and coverage. For example, if the posterior mean of b_1 for dataset $s \in \{1, ..., 100\}$ is $\hat{b}_1^s$ and the true value is b_1^*, then the estimated bias and its standard error are

$$Bias = \sum_{s=1}^{100} (\hat{b}_1^s - b_1^*)/100 \quad \text{and} \quad SE = s_1/\sqrt{100}$$

were s_1 is the standard deviation of $\{\hat{b}_1^1, ..., \hat{b}_1^{100}\}$.

Summarize your analysis in a 3-5 page report (double spaced, 11pt, one-inch margins). Papers longer than five pages will be penalized. To avoid penalty, your report MUST have the following sections and contents:

1. Introduction: Briefly describe the problem and your objectives

2. Model to be fit to simulated data: Describe the Bayesian model including priors for all parameters that are uninformative over the parameters' support

3. Computation: Write an `R` function that analyzes one dataset. The function should take as input the vectors `t` and `Y` and return the posterior mean and 95% credible interval for each of the four b_j. So,

   ```
   pollen_fit <- function(t,Y){
      ... MCMC and other code ...
      output <- list(post_mean=...,cred_interval=...)
   return(output)}
   ```

   Give the details of the Bayesian computational algorithms you use, including how you select initial values (which is often important for non-linear models) and how you verify the algorithms were successful

4. Data-generation: Describe how you simulate the data from the non-linear model and the settings you will explore

5. Metrics: Detail (with formulas) the metrics you will use to summarize the performance of the Bayesian analysis

6. Results: Present your results in tables and/or figures and interpret your findings

7. Discussion: Summarize your results and discuss how this simulation study might impact the way you would analyze real pollen count data

Your paper should be written as a professional document with full paragraphs, clearly labeled and numbered figures and/or tables, and few spelling/grammar errors. You should include sufficient detail that another student in class could reproduce your results. Summarize your analysis in a PDF document submitted via email to the instructor. Append your code to the end of this document and submit a single PDF document (code does not count towards the page limit). You do not need to turn in all of your code, but please turn in the `R` function described in Part 3.

Have fun!

```
# Code to generate a dataset from the non-linear regression model

library(mvtnorm)
n        <- 100
b_true   <- c(10,50,100,10)
sig_true <- 1
rho_true <- 0.9

t        <- seq(1,365,length=n)
temp     <- ((t-b_true[3])/b_true[4])^2
mu_true  <-  b_true[1] + b_true[2]*exp(-0.5*temp)
dist     <- as.matrix(dist(t))
S        <- rho_true^dist
Y        <- rmvnorm(1,mu_true,S)
Y        <- as.vector(Y)

plot(t,Y)
lines(t,mu_true,lwd=2)
legend("topright",c("Data","True mean"),pch=c(1,NA),lwd=c(NA,2),bty="n")
```