STAT 6106 - Assignment 5:
Bayesian fitting of cell cycle gene expression data
due on Dec. 20th, email to TA

Cell division cycle is the concerted sequence of processes by which a cell duplicates its DNA and divides into two daughter cells. Many genes are expressed periodically at a specific stage during the cell cycle when they peak and trough over a certain time range. They are termed as "Periodically Expressed genes". With the help of the microarray techniques and various cell phase synchronization methods (synchronizing the progression of cells through the stages of cell cycle), researchers have conducted genome-wide time series expression analyses on synchronized cells for fission yeast (Rustici et al. 2004, Nature Genetics). During the experiment, a culture of cells are grown and synchronized. A set of microarrays are used to measure gene expressions at selected time points. All values were converted to log-ratios with base 2. Let Y_{gt} denote the gene expression log-ratio at time T_t for gene g, where $g = 1, \cdots, G$, $t = 1, \cdots, S$. Here Y_{gt} is the observed data; T_t, the time of the measurement; G, the total number of genes studied; and S, the total number of time points measured in this experiment. We assume the following model for each time series:

$$Y_{gt} = a_g + b_g \cdot T_t + c_g \cdot (\min(T_t - d_g, 0))^2 + A_g \cdot \cos(\mu \cdot T_e + \phi_g) + \varepsilon_{gt} \quad (1)$$

where
$a_g + b_g \cdot T_t + c_g \cdot (\min(T_t - d_g, 0))^2$: trend component
$A_g \cdot \cos(\mu \cdot T_t + \phi_g)$: periodic component
$\varepsilon_{gt} \sim N(0, \sigma_g^2)$: i.i.d. noise

The physical meaning of all parameters:
a_g, b_g: coefficients of the linear trend of a time series
d_g: ending time of block-release effect of a time series
c_g: magnitude of block-release effect of a time series
σ_g^2: noise level of a time series
A_g: amplitude of periodic component of a time series
ϕ_g: gene-specific phase, which decides its peaking time
μ: the cell cycle angular frequency shared by all genes, equal to 2π divided by the period of cell cycle of an experiment, usually in the range of 120-180 minutes.

Data: The data file is Assignment5_data.csv, which contains the time series data for 20 genes. The first row is T_t, the time of the measurement (unit:

minute). The first column is the name of the gene.

Re-parameterize the model: To make our problem easier, let us fix μ at 0.04 and fit the model to each gene separately. We re-parameterize some parameters for computing convenience as follows:

$$Y_{gt} = a_g + b_g^{'} \cdot 0.01 \cdot T_t + c_g \cdot (\min(T_t - d_g, 0))^2 + f_g \cdot \cos(\mu \cdot T_t) - h_g \cdot \sin(\mu \cdot T_t) + \epsilon_{gt}, \tag{2}$$

where $b_g^{'} = 100 \cdot b_g$, $f_g = A_g \cdot \cos\phi_g$, $h_g = A_g \cdot \sin\phi_g$.

Furthermore, Equation (2) can be expressed more concisely via vector as follows:

$$Y_{gt} = X_t \cdot \beta_g + c_g \cdot (\min(T_t - d_g, 0))^2 + \epsilon_{gt} \tag{3}$$

where $X_t = (1,\ 0.01 \cdot T_t,\ \cos(\mu \cdot T_t),\ -\sin(\mu \cdot T_t))$, $\beta_g = (a_g, b_g^{'}, f_g, h_g)^T$.

The model in Equation (1) is equivalent to the model in Equation (2) and Equation (3) in the sense that their parameters have a deterministic one-to-one mapping relationship.

Based on Equation 3, we can write down the **Likelihood** for one gene: $L(\theta_g|Y_g) \propto (\frac{1}{\sigma_g^2})^{T/2} \exp\{-\frac{1}{2\sigma_g^2} \cdot \sum_t [Y_{gt} - X_t \cdot \beta_g - c_g \cdot (\min(T_t - d_g, 0))^2]^2\}$.

To do Bayesian inference, we introduce the following
Prior: $\pi(\theta_g) = \pi_1(\beta_g)\pi_2(c_g)\pi_3(d_g)\pi_4(\sigma_g^2)$,
where

$$\begin{aligned}
&\pi_1(\beta_g) \sim N_4(0, \alpha_1 \cdot I_4),\ \alpha_1 = 4 \\
&\pi_2(c_g) \sim N(0, \alpha_2),\ \alpha_2 = 4 \\
&\pi_3(d_g) \sim I_{[0,500]}, i.e., uniform(0, 500) \\
&\pi_4(\sigma_g^2) \sim IG(a, b),\ a = 10,\ b = 0.01
\end{aligned}$$

I_4 is an Identity matrix of size=4; $I_{[0,500]}$ means a uniform between 0 and 500.

The Code: The MCMC algorithm for estimating the parameters is provided in the file Assignment5-1Gene.r.txt. Detail annotations are also provided with the code as comments after the "#" sign.

To Do:

1. Write down the MCMC algorithm (You should describe the procedure, write down all conditional distributions, proposal distributions and importance ratio/acceptance probabilities. You can get hints from the given R code because the code has implemented this algorithm.)

2. Use the MCMC algorithm in the given code to fit all genes, report the posterior summary (mean, sd) of all parameters, and draw all fitted curves together with the observed data for each gene

3. Now let's change our model in Equation (2) to the following model:

$$Y_{gt} = f_g \cdot \cos(\mu \cdot T_t) - h_g \cdot \sin(\mu \cdot T_t) + \epsilon_{gt}, \tag{4}$$

   Write down the MCMC algorithm, and revise the given code to performance Bayesian inference using this new model. Also draw the fitted curve.