Statistical Analysis on Circadian Gene Expression

Jingjie Zhang, Elina Sergeeva

STAT 689 Course Project

April 23, 2018

Overview

Introduction

A circadian clock is a biochemical oscillator that cycles with a stable phase and is synchronized with solar time. Such a clock's period is almost exactly 24 hours. We will be looking for patterns in cycles of different genes expressions in following data sets:

- Circadian Gene Data Set from Bell-Pederson's Lab at TAMU
- Circadian Gene Transcription in Mammals from Hughes M. E., et al. (2009).

Part 1

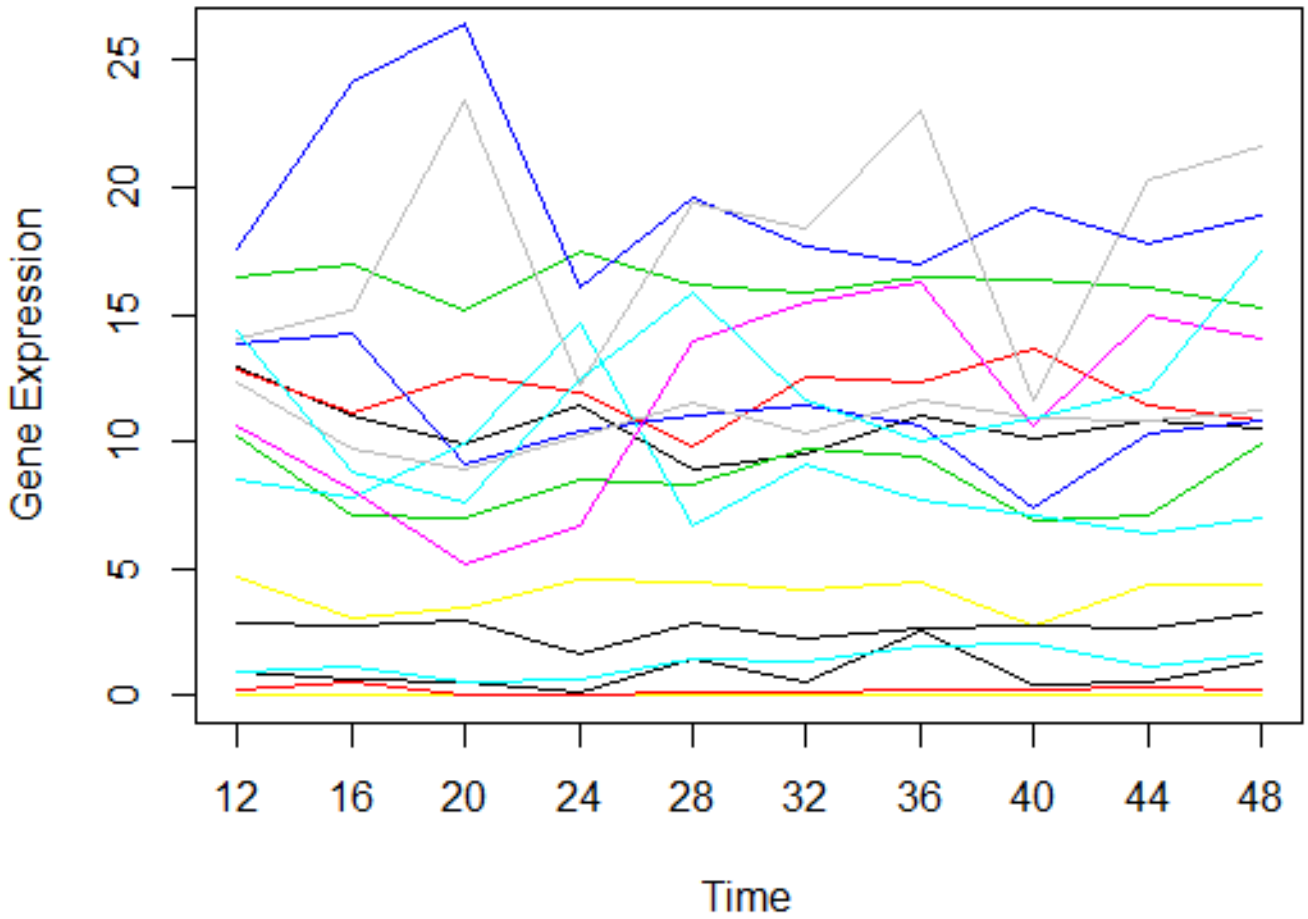
Clustering

Data from Bell-Pederson's Lab

- Each gene expression was measured at times 12 through 48 in time increments 4 hours.
- Three replicates at each time.
- For this report we are averaging two closest expression values at each time for each gene.

Data from Bell-Pederson's Lab

Genes NCU00003 - NCU00023: great variety of curves.



Data from Bell-Pederson's Lab

Complications:

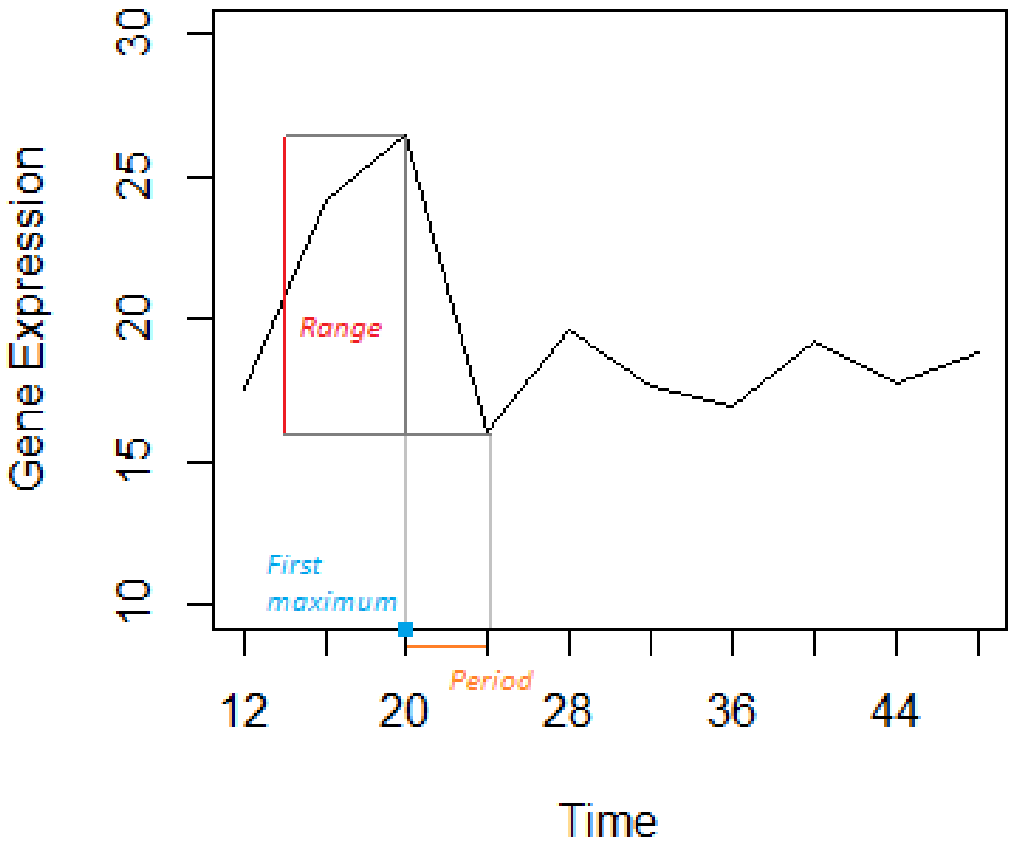
- Limited number of time points for identifying global features.
- There are no known groups.
- Patterns are hard to be discovered with distance based clustering applied to set directly.

Feature creation

To discover similar patterns in curves, three global features were defined for each gene:

- Range - the difference between expressions global maximum and minimum.
- First Maximum - the time point when the global maximum occurred.
- Period - how many time points passed between First Maximum and next local maximum.

Feature creation

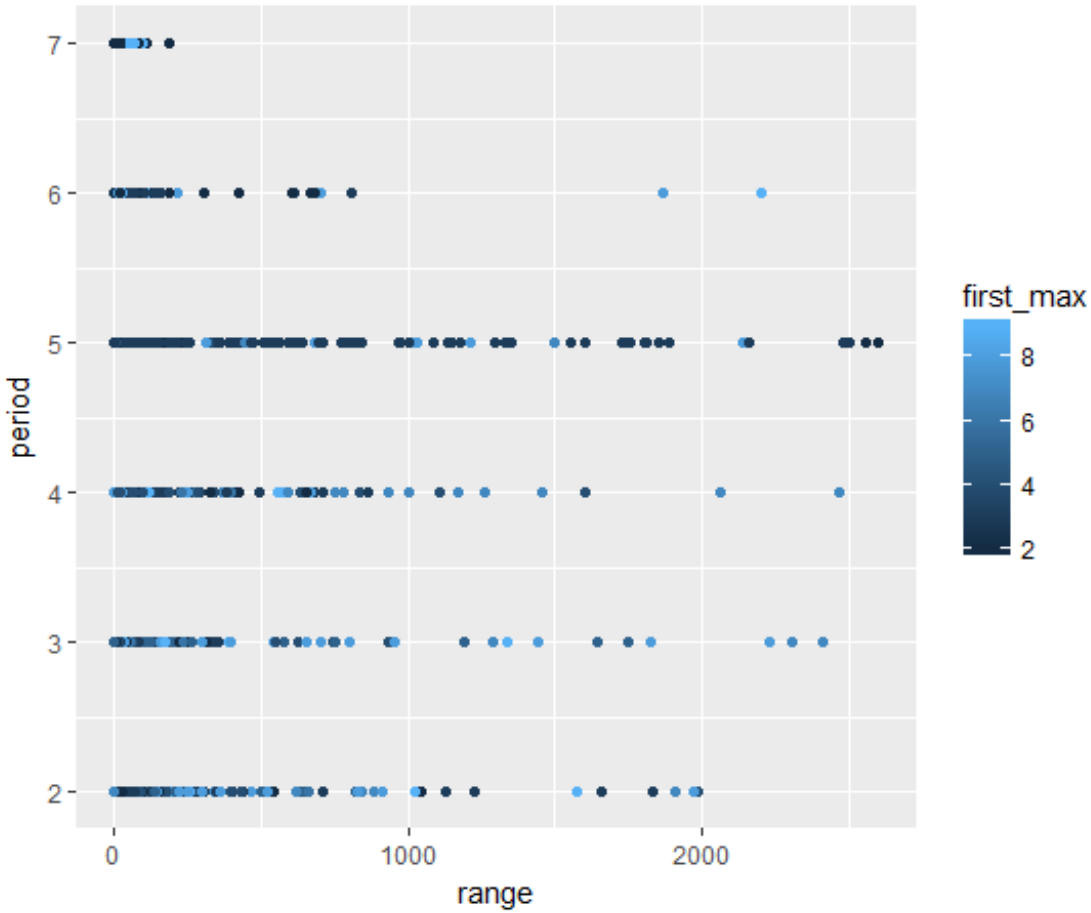


Feature creation

We are not considering observations which have either of the following:

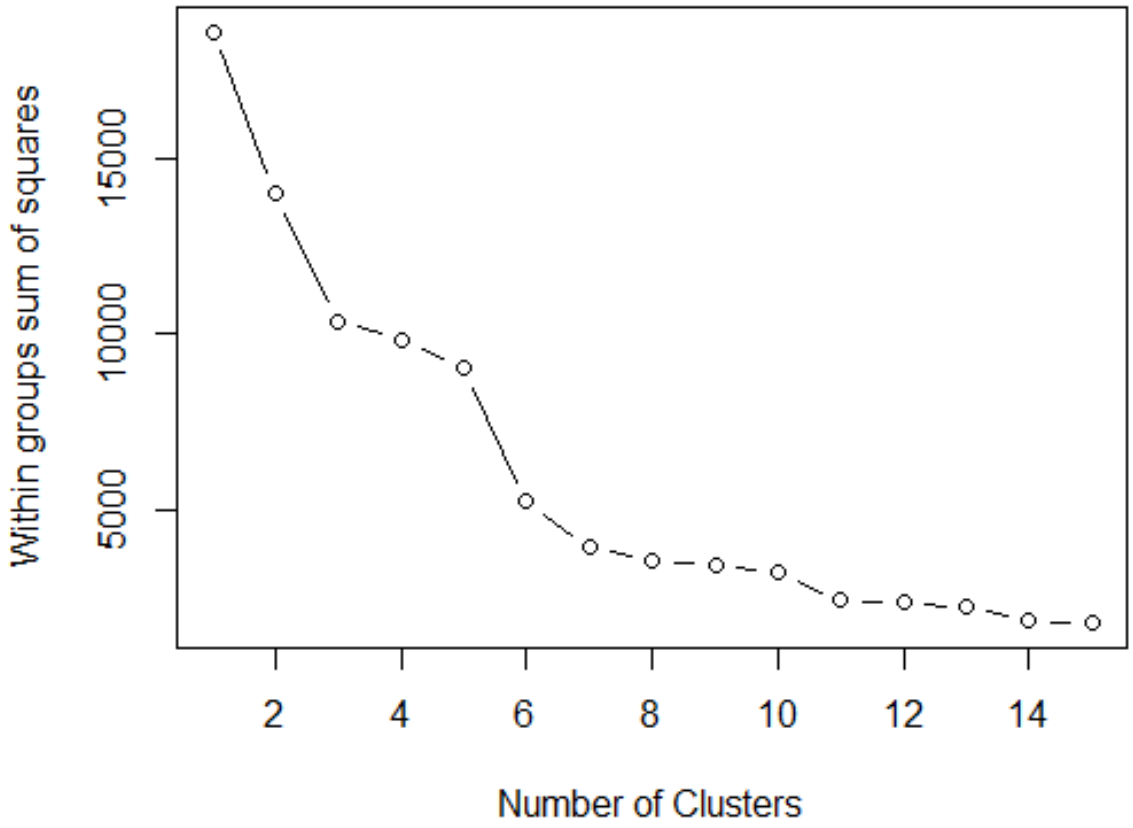
- $Range > 2800$
- First Maximum at time point 12 or 48
- $Period = 0$

Feature creation



K-means clustering

Within groups sum of squares



K-means clustering

Size of clusters on scaled features:

$$737, 1392, 594, 1789, 1608, 84$$

cluster	range	first_max	period
1	-0.10610539	1.15061433	1.1050496
2	-0.12605145	-0.58417277	-1.0064050
3	0.09697282	-0.99034995	1.8979287
4	-0.11996359	-0.68248834	0.1804918
5	-0.12868199	1.10088752	-0.5734642
6	7.35234461	0.04973953	0.6947048

Table: Clusters on scaled features

K-means clustering

Back to original:

cluster	range	first_max	period
1	19.85818	3.920838	2.289070
2	27.06229	2.882210	3.532900
3	1403.95437	5.047619	3.952381
4	24.19197	7.485753	4.423338
5	20.01545	7.375622	2.496891
6	54.13207	2.740614	5.351536

Table: Clusters on original features

K-means clustering

	first_max							
	2	3	4	5	6	7	8	9
1	0	0	0	0	171	204	195	167
2	253	277	575	287	0	0	0	0
3	217	312	65	0	0	0	0	0
4	432	562	396	399	0	0	0	0
5	0	0	0	0	431	405	509	263
6	3	36	6	5	0	16	13	5

Table: Occurrences in clusters [First Max]

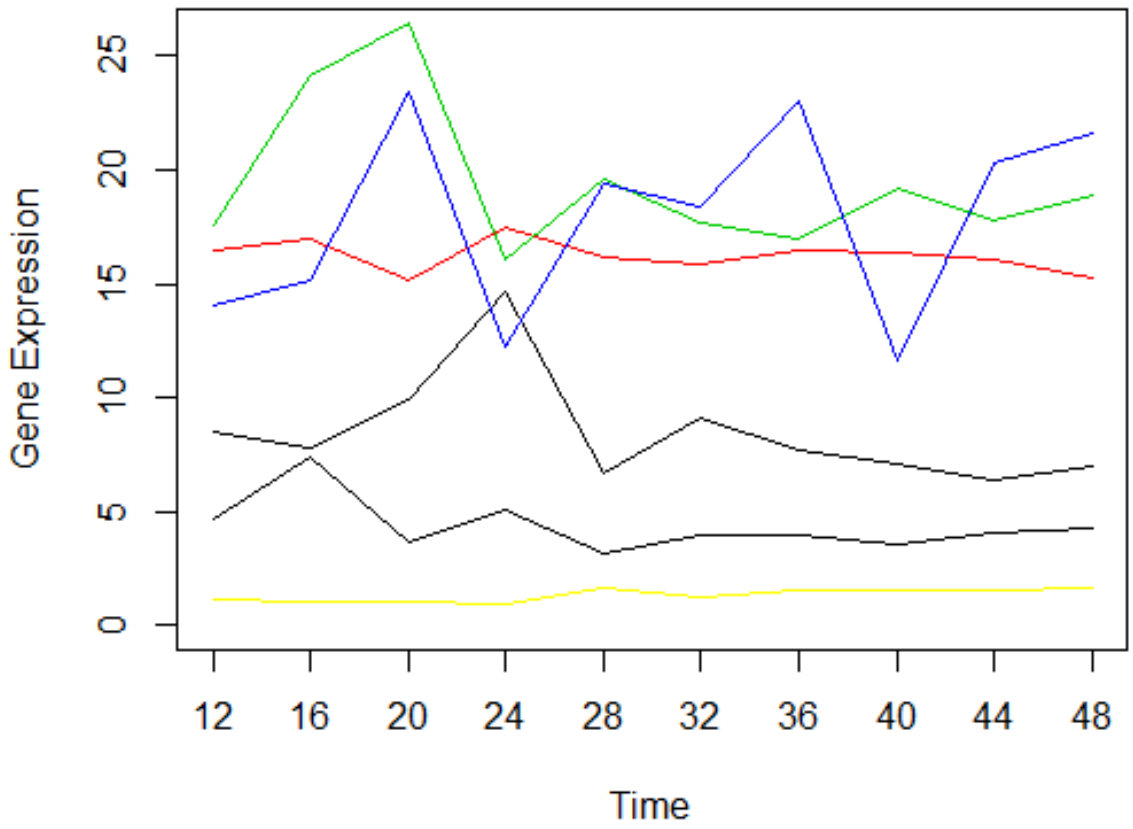
K-means clustering

	period					
	2	3	4	5	6	7
1	0	0	492	187	49	9
2	1392	0	0	0	0	0
3	0	0	8	410	146	30
4	0	1141	648	0	0	0
5	809	799	0	0	0	0
6	16	15	13	37	3	0

Table: Occurrences in clusters [Period]

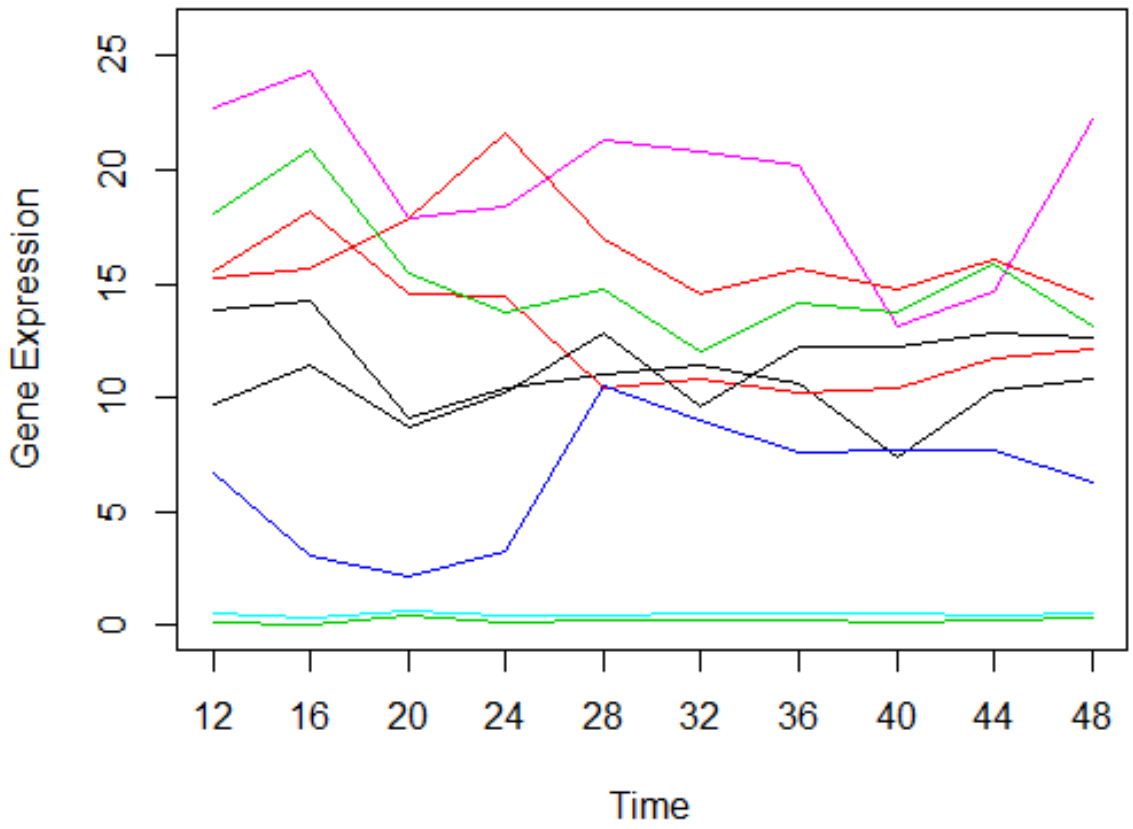
K-means clustering

Cluster 1



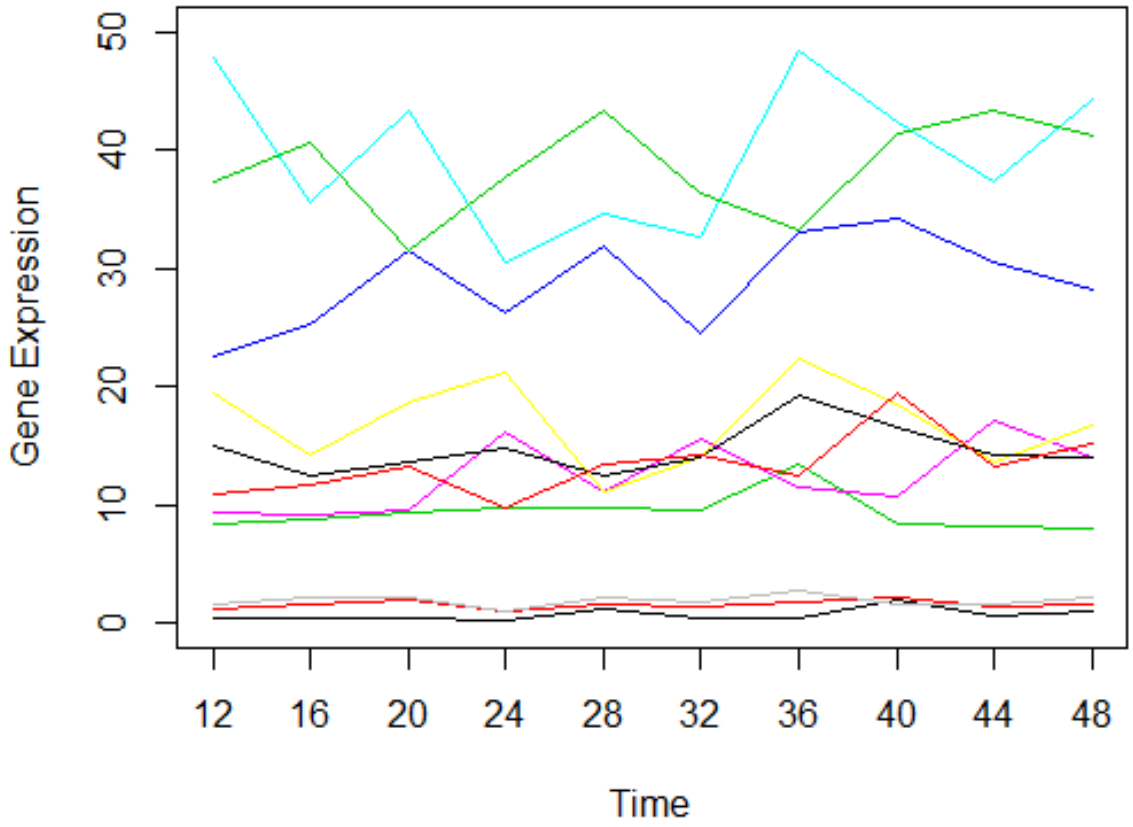
K-means clustering

Cluster 3



K-means clustering

Cluster 4



Bayesian Inference for Sinusoidal Model

Sinusoidal model

Some genes expression data perform like sinusoidal curves over time, one important goal is to estimate the period in the sinusoidal model.

Suppose

$$y_t = \nu cos(\omega t + \varphi) + \epsilon_t \quad \text{for } t = t_1, t_2, \ldots, t_n \tag{1}$$

where ν is the amplitude, ω is the frequency, φ is the phase angle.

- ϵ_t – Noise at time t
- independent and identically distributed, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$

Equivalent form

The Sinusoidal model (1) is equivalent to

$$y_t = \gamma_1 cos(\omega t) + \gamma_2 sin(\omega t) + \epsilon_t \tag{2}$$

where $\nu = \sqrt{\gamma_1^2 + \gamma_2^2}$, $\varphi = -arctan(\frac{\gamma_2}{\gamma_1})$.

Matrix form

In Matrix notation, we write model (2) as:

$$y = A\gamma + \epsilon \tag{3}$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad A = \begin{pmatrix} cos(\omega t_1) & sin(\omega t_1) \\ \vdots & \vdots \\ cos(\omega t_n) & sin(\omega t_n) \end{pmatrix} \quad \gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- We have $y \sim \mathcal{N}(A\gamma, \sigma^2 \mathcal{I}_n)$

Inference based on Gibbs Sample

The joint posterior distribution is given by

$$\pi(\omega, \gamma, \sigma^2|y) \propto f(y|\omega, \gamma, \sigma^2)\pi(\omega, \gamma, \sigma^2)$$

The prior distribution (suggested by Dou and Hodgson, 1995) is:

$$\pi(\omega, \gamma, \sigma^2) \propto 1 \cdot 1 \cdot \pi(\sigma^2) \propto \frac{1}{\sigma^2}$$

Then the posterior can be written as:

$$\pi(\omega, \gamma, \sigma^2|y) \propto (\sigma^2)^{-\frac{n}{2}-1} exp\left(-\frac{(y - A\gamma)^T(y - A\gamma)}{2\sigma^2}\right)$$

Conditional posterior for Amplitude γ

Recall $\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$, then

$$\begin{aligned}
\pi(\gamma|\omega, \sigma^2, y) &\propto f(y|\omega, \gamma, \sigma^2)\pi(\omega, \gamma, \sigma^2) \\
&\propto exp\left(-\frac{(y - A\gamma)^T(y - A\gamma)}{2\sigma^2}\right) \\
&\propto exp\left(-\frac{\gamma^T A^T A\gamma - 2\gamma^T A^T y}{2\sigma^2}\right) \\
&\sim \mathcal{N}_2\left((A^T A)^{-1}A^T y, \sigma^2(A^T A)^{-1}\right)
\end{aligned}$$

We can sample γ directly from the bivariate Gaussian distribution.

Conditional posterior for Noise variance σ^2

$$\begin{aligned}\pi(\sigma^2|\omega,\gamma,y) &\propto f(y|\omega,\gamma,\sigma^2)\pi(\omega,\gamma,\sigma^2)\\ &\propto (\sigma^2)^{-\frac{n}{2}-1} exp\left(-\frac{(y-A\gamma)^T(y-A\gamma)}{2\sigma^2}\right)\\ &\sim \mathcal{IG}\left(\frac{n}{2},\frac{(y-A\gamma)^T(y-A\gamma)}{2}\right)\end{aligned}$$

We can sample σ^2 directly from the inverse gamma distribution.

Conditional posterior for Frequency ω

Recall $A = \begin{pmatrix} cos(\omega t_1) & sin(\omega t_1) \\ \vdots & \vdots \\ cos(\omega t_n) & sin(\omega t_n) \end{pmatrix}$

then

$$\pi(\omega|\gamma, \sigma^2, y) \propto f(y|\omega, \gamma, \sigma^2)\pi(\omega, \gamma, \sigma^2)$$
$$\propto exp\left(-\frac{(y - A\gamma)^T(y - A\gamma)}{2\sigma^2}\right)$$

we can't sample ω directly!

Conditional posterior for Frequency ω

- Laplace approximation
- MAP estimator

$$\begin{aligned}\hat{\omega}_{MAP} &= arg\,max_{\omega}\left\{exp\left(-\frac{(y-A\gamma)^T(y-A\gamma)}{2\sigma^2}\right)\right\}\\ &= arg\,min_{\omega}\left\{(y-A\gamma)^T(y-A\gamma)\right\}\end{aligned}$$

- Hessian at $\hat{\omega}_{MAP}$

$$\begin{aligned}H &= -\frac{\gamma^T(\frac{\partial A}{\partial \omega})^T(\frac{\partial A}{\partial \omega})\gamma}{\sigma^2}\Big|_{\omega=\hat{\omega}_{MAP}}\\ &= -\frac{1}{\sigma^2}(X_{\omega}^T X_{\omega})\end{aligned}$$

where $X_{\omega} = (\frac{\partial A}{\partial \omega})\gamma|_{\omega=\hat{\omega}_{MAP}}$.

Conditional posterior for Frequency ω

Then approximately, we have

$$\pi(\omega|\gamma, \sigma^2, y) \sim \mathcal{N}(\hat{\omega}_{MAP}, \sigma^2(X_\omega^T X_\omega)^{-1})$$

from which we can easily draw samples.

Gibbs Sampler

Algorithm

1. Initialize $\omega^{[0]}, \sigma^{2[0]}, \gamma^{[0]}$
2. For i=1, 2, ..., N
   - Draw the amplitude
     $$\gamma^{[i]} \sim \mathcal{N}_2\left(\left(A^{[i-1]^T} A^{[i-1]}\right)^{-1} A^{[i-1]^T} y, \sigma^{2[i-1]}\left(A^{[i-1]^T} A^{[i-1]}\right)^{-1}\right)$$
   - Draw the noise variance
     $$\sigma^{2[i]} \sim \mathcal{IG}\left(\frac{n}{2}, \frac{\left(y - A^{[i-1]}\gamma^{[i]}\right)^T \left(y - A^{[i-1]}\gamma^{[i]}\right)}{2}\right)$$
   - Calculate the MAP estimator $\hat{\omega}^{[i]}_{MAP}$
   - Draw the frequency
     $$\omega^{[i]} \sim \mathcal{N}\left(\hat{\omega}^{[i]}_{MAP}, \sigma^{2[i]}(X_\omega^{[i]^T} X_\omega{}^{[i]})^{-1}\right)$$

Simulation

Simulation setting:

$$y_t = \gamma_1 cos(\omega t) + \gamma_2 sin(\omega t) + \epsilon_t$$

with $\gamma_1 = 0.2, \gamma_2 = 0.41, \omega = 0.74$ and $\epsilon_t \sim \mathcal{N}(0, 0.4^2)$.

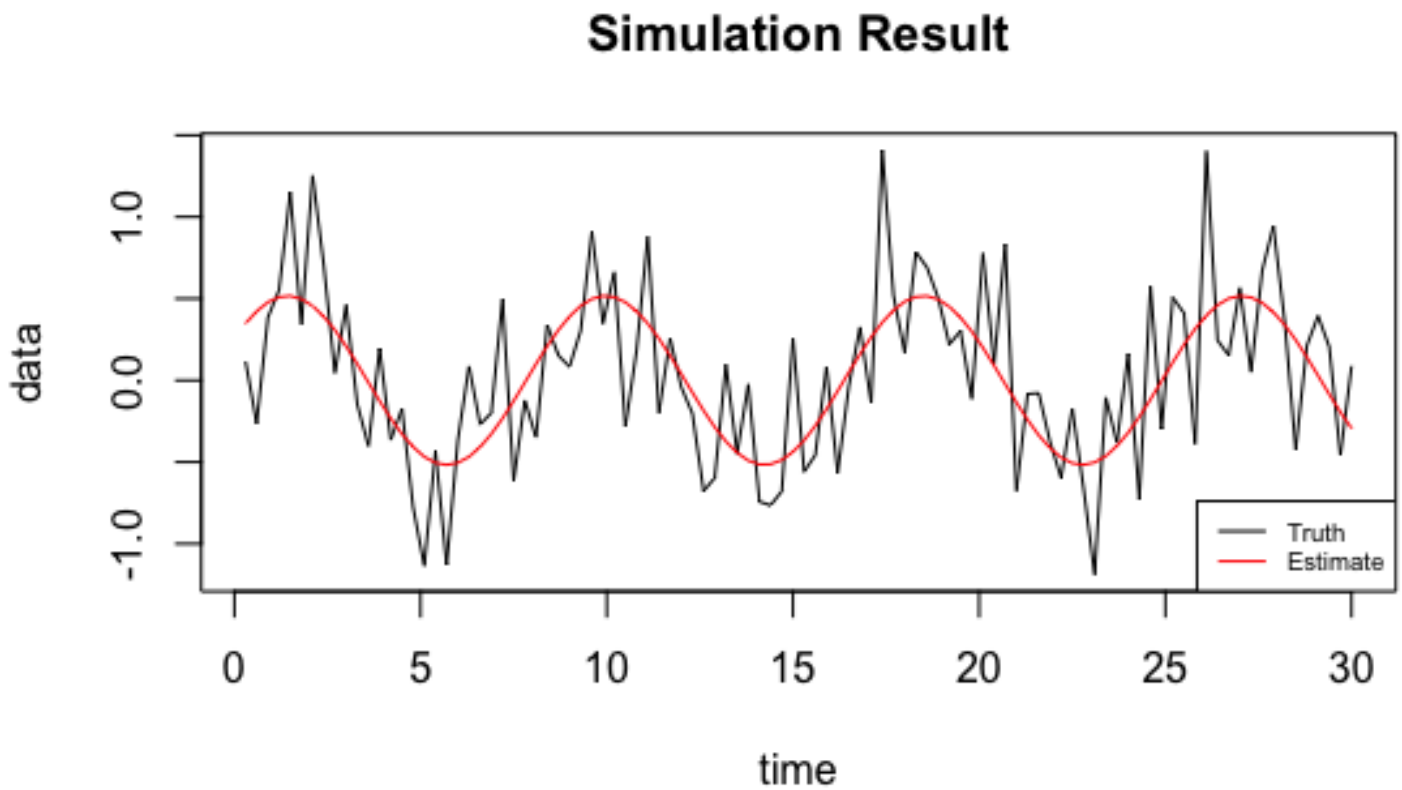
- Sample size $n = 100$
- Time $t = 0.3, 0.6, \ldots, 29.7, 30$

Simulation result

By running Gibbs sampler, we get the estimation

$$\hat{\omega} = 0.7360$$

with 95% credible interval $[0.7328, 0.7420]$.



Application to gene expression data

- **Data Source:** Hughes M. E., et al. (2009). Harmonics of circadian gene transcription in mammals. *PLoS Genet*,5(4), e1000442.

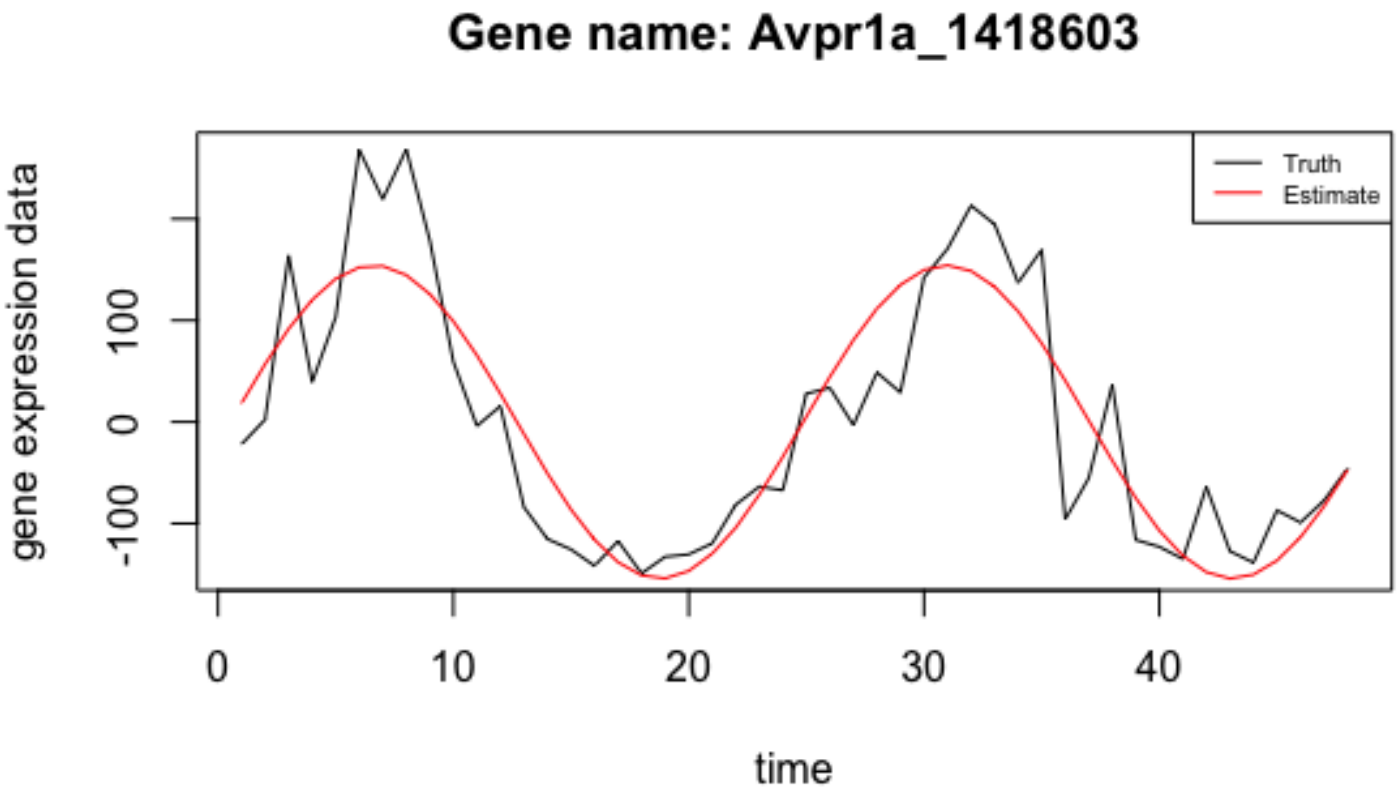
- This data set lists expression profiles of 20 circadian transcripts with 1h-resolution covering two days.

geneName	**CT1**	**CT2**	...	**CT48**
Hist1h1c_1416101	2700.3357	2394.2878	...	2502.8316
Fkbp5_1448231	60.4610	56.3778	...	64.86352
⋮	⋮	⋮	...	⋮
Tef_1424175	794.4965	635.3928	...	690.0365
Nr1d2_1416958	1098.2922	1035.3626	...	862.4657

Table: Gene expression data

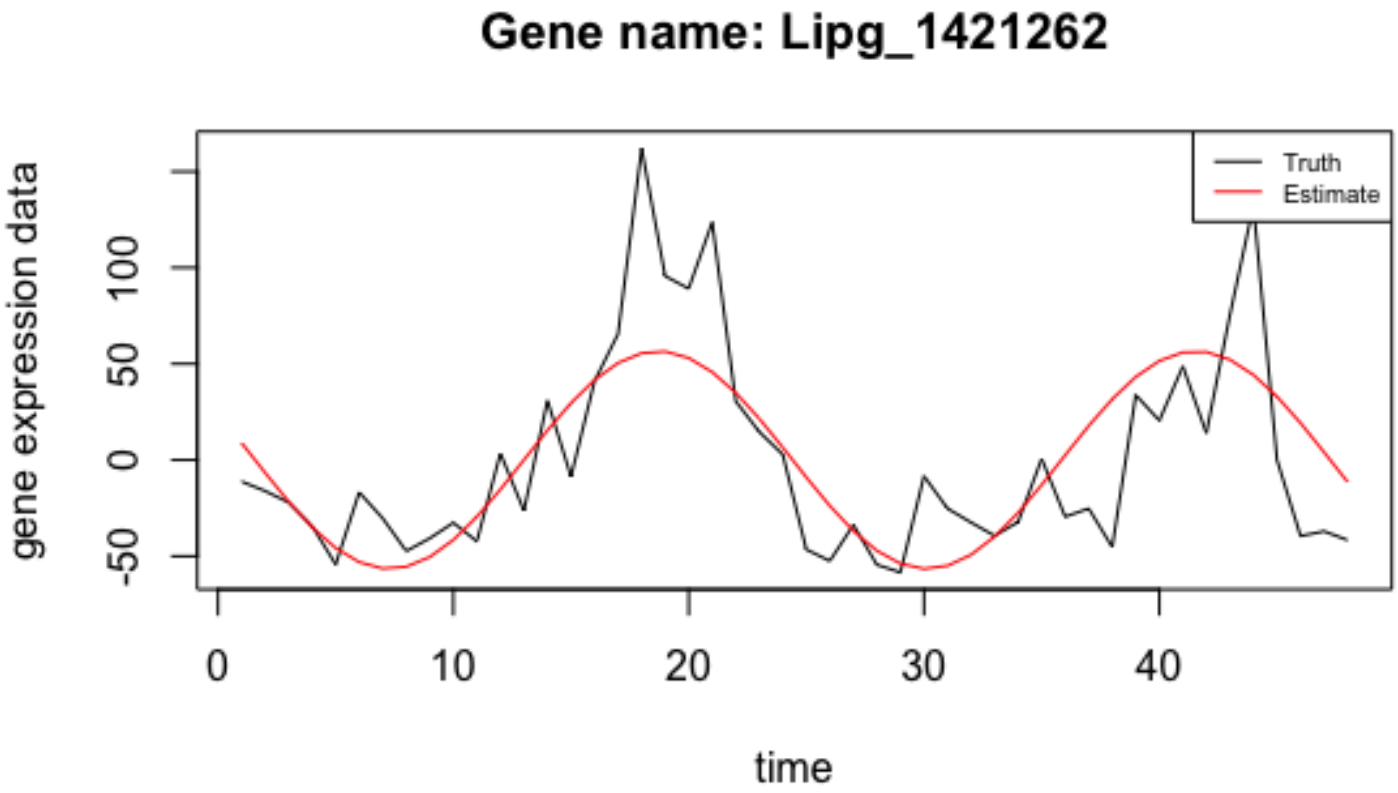
Application to gene expression data

For gene Avpr1a_1418603, we get $\hat{\omega} = 0.25821$.
95% credible interval $[0.25810, 0.25825]$.



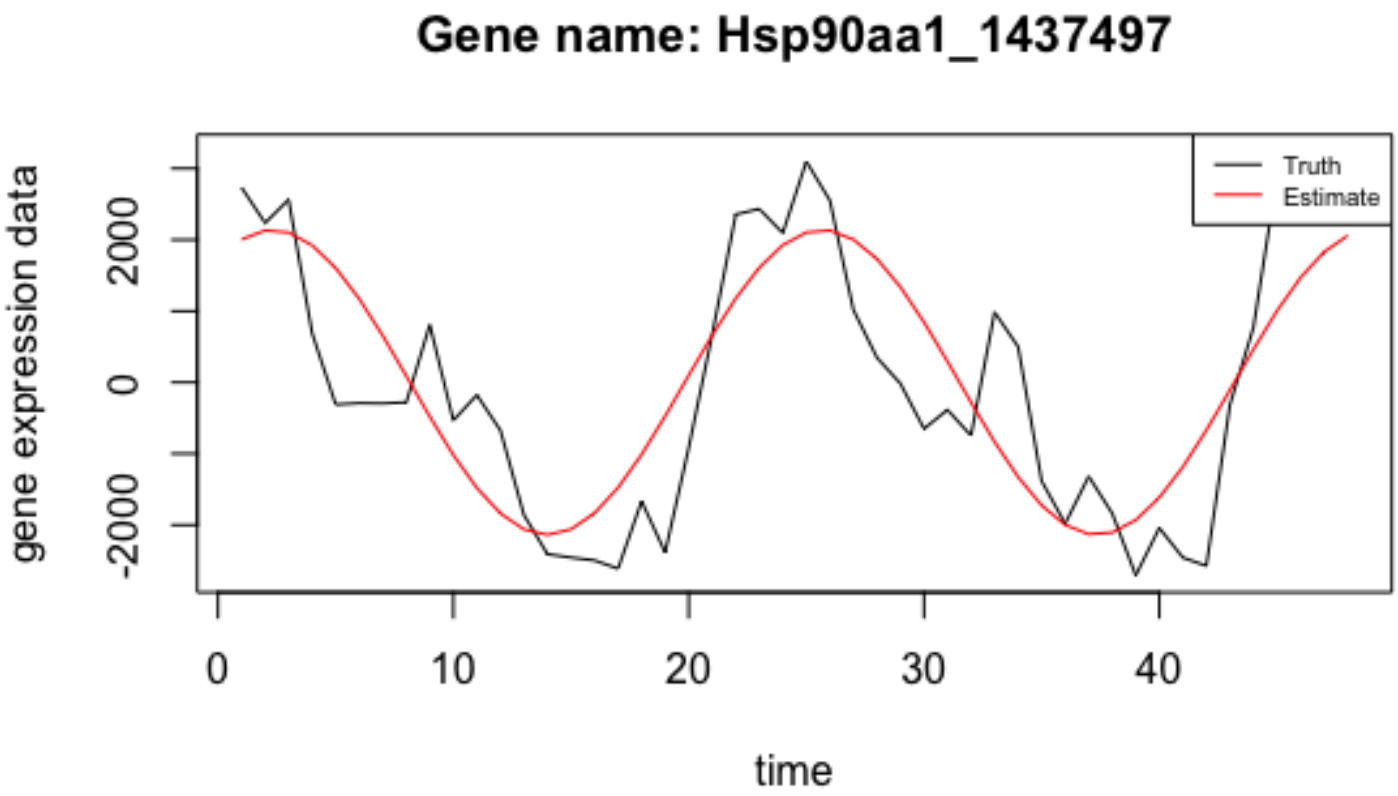
Application to gene expression data

For gene Lipg_1421262, we get $\hat{\omega} = 0.27506$.



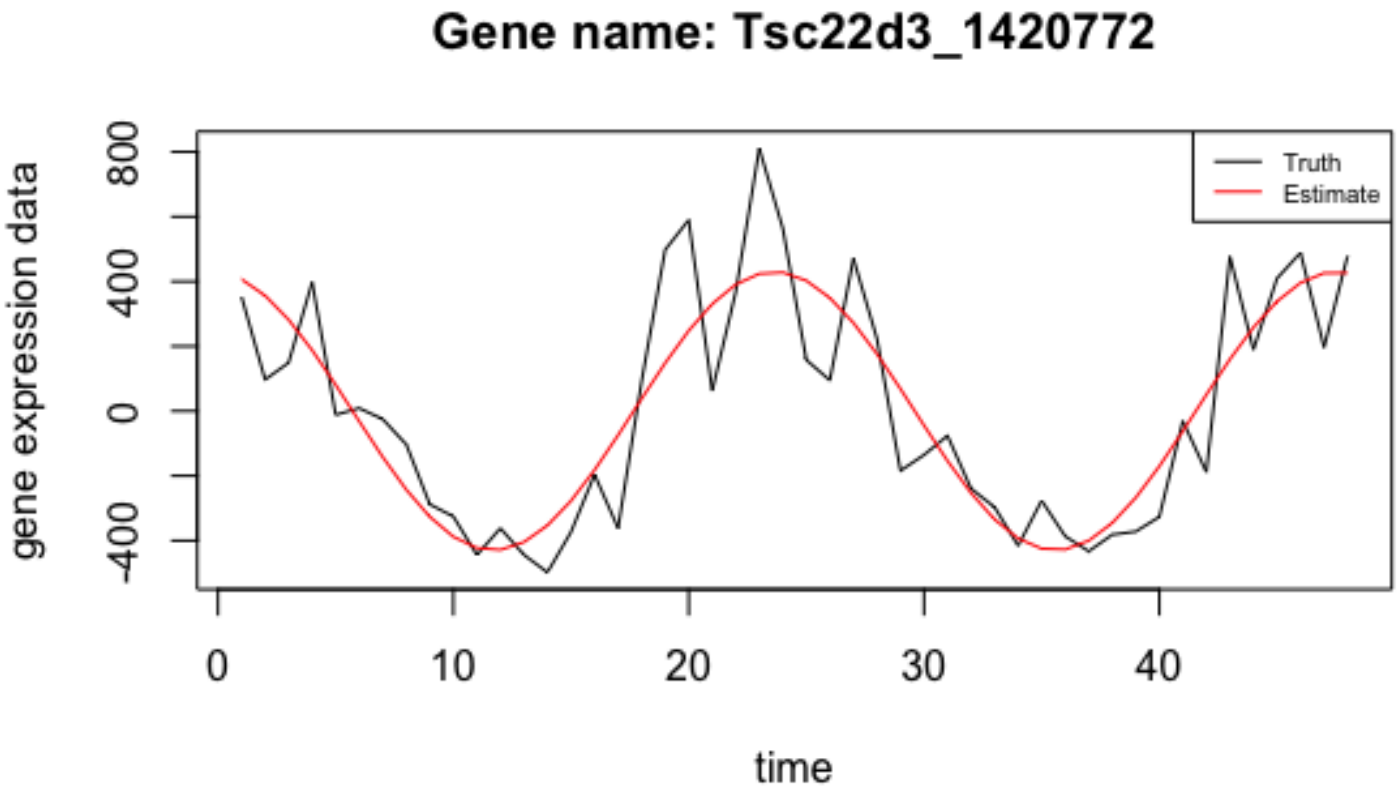
Application to gene expression data

For gene Hsp90aa1_1437497, we get $\hat{\omega} = 0.26904$.



Application to gene expression data

For gene Tsc22d3_1420772, we get $\hat{\omega} = 0.26302$.



Conclusions

The cycle of the circadian gene varies by species which is commonly 24 hours for mammals. As far as data from Professor Bell-Pederson's Lab was collected at only 10 time points during a period of 36 hours and species are not specified it is challenging to capture the real period. Still we consider performed k-means clustering useful as primary filter for further implementation of more advanced methods for identifying genes behavior.

The Bayesian Inference for sinusoidal model on the Hughes dataset with larger time frame has proved that it is possible to capture specific of a circadian gene and the estimation is quite good.

References

- Hughes M. E., et al. (2009). Harmonics of circadian gene transcription in mammals. *PLoS Genet*,5(4), e1000442.
- Dou, L., & Hodgson, R. J. W. (1995). Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation. I. Inverse problems, 11(5), 1069.
- Cohen, A. L., Leise, T. L., & Welsh, D. K. (2012). Bayesian statistical analysis of circadian oscillations in fibroblasts. Journal of theoretical biology, 314, 182-191.

Thanks!