Detecting Circadian Gene Expressions via Bayesian Analysis: an Application to the *Arabidopsis Thaliana* Dataset

Silvia Montagna[1], Amir Khorrami Chokami[2], and Surya T. Tokdar[3]

[1] Dipartimento di Economia "Marco Biagi", Università degli Studi di Modena e Reggio Emilia, via Berengario 51, Modena 41121,
silvia.montagna@unimore.it
[2] Dipartimento ESOMAS, Università degli Studi di Torino, C.so Unione Sovietica 218/bis, Torino 10134
[3] Department of Statistical Science, Duke University, Durham, NC USA 27708

Abstract. In genomic applications, there is often interest in identifying genes whose time-course expression trajectories exhibit periodic oscillations with a period of approximately 24 hours (circadian genes). While it is natural to expect that the expression of gene i at time j might depend to some degree on the expression of the other genes measured at the same time, widely-used rhythmicity detection techniques do not accommodate for the potential dependence across genes. We develop a Bayesian approach for periodicity identification that explicitly takes into account the complex dependence structure across time-course trajectories in gene expressions. The methodology is applied to a plant gene expression dataset.

Keywords: Bayesian Fourier analysis, high-dimensional data, circadian rhythms, gene expressions

1 Introduction

Circadian rhythms are cycles of biological activity based on a 24-hour period which allow organisms to anticipate and adapt to predictable daily oscillations in the environment [4]. Circadian rhythms are present in almost all plants and animals. In plants, circadian rhythms play a role in the regulation of plant metabolic pathways, such as photosynthesis and carbon metabolism, in the regulation of developmental processes and signalling pathways, such as defence responses. In humans, blood pressure, hormone production, metabolism and other biological cycles are clock-regulated. Disruptions to the circadian rhythms have been linked to a variety of pathologies, for example cancer, psychiatric disorders and neurodegenerative diseases in humans [4].

Circadian rhythms are controlled by the circadian clock, namely, a network of mutually interacting genes controlling the timing of many physiological processes. The interest is in identifying such genes through examination of their

expression levels or "transcripts". Several authors have proposed methods for periodicity identification in biomedical research over the last couple of decades. [5] provide a comprehensive review of the main existing techniques commonly used for circadian rhythm detection, and evaluate their accuracy and reproducibility on various empirical datasets. As a separate line of research, several model-based clustering algorithms have been proposed in both the classical and Bayesian framework (e.g., [3]).

A key assumption in the approaches above is that of independence across genes. Although practical from a computational perspective, the independence assumption is often too strong to be realistic in many applications. In this paper, we propose a Bayesian approach that identifies periodic signals in gene expression profiles while accounting for dependence in the functional data. Specifically, the true underlying signal for each transcript is decomposed into a series expansion of sine and cosine (Fourier) waves. Conditional dependence across genes at each time point is accommodated via a latent factor framework. Dimensionality reduction and sparsity are induced through careful modelling of the latent factors as well as the Fourier basis coefficients.

The rest of the paper is organised as follows. Section 2 outlines the methodology. Section 3 discusses priors and posterior inference for detecting circadian genes. The proposed approach is tested on the *Arabidopsis thaliana* [3] dataset (Section 4). Conclusions are drawn in Section 5.

2 Methods

We consider data in the form of a $p \times T$ matrix $\mathbf{Y} = \{y_{ij}\}$, where entry y_{ij} denotes the expression level for gene i at time t_j, for $i = 1, \dots, p$, and with p denoting the total number of genes. In circadian microarray studies, data are typically collected over two complete circadian cycles and the sampling rate is usually either two ($T = 24$) or four hours ($T = 12$). We assume that the y_{ij}'s are error-prone measurements of an underlying smooth true trajectory:

$$y_{ij} = f_i(t_j) + \nu_{ij}. \tag{1}$$

Suppose that the de-trended and centred true signal for gene i at time t_j, $f_i(t_j)$, can be represented as:

$$f_i(t_j) = \sum_{m=1}^{q} \left[\theta_{i,2m-1} b_{2m-1}(t_j) + \theta_{i,2m} b_{2m}(t_j)\right] = \sum_{m=1}^{q} \boldsymbol{\theta}_{i,m} \mathbf{b}_m(t_j) = \mathbf{b}_j^\top \boldsymbol{\theta}_i,$$

where $\boldsymbol{\theta}_{i,m} = (\theta_{i,2m-1}, \theta_{i,2m})$ and $\mathbf{b}_m(t_j) = \left[b_{2m-1}(t_j), b_{2m}(t_j)\right]^\top$, for $m = 1, \dots, q$. The vector $\mathbf{b}_j^\top = [b_1(t_j), b_2(t_j), \dots, b_{2q-1}(t_j), b_{2q}(t_j)]$ represents a set of $2q$ fixed basis functions evaluated at time t_j. The natural basis choice for the space of periodic functions is the Fourier basis:

$$\mathbf{b}^\top(t) = \left[\sin\left(\frac{2\pi}{\omega_1} t\right), \cos\left(\frac{2\pi}{\omega_1} t\right), \dots, \sin\left(\frac{2\pi}{\omega_q} t\right), \cos\left(\frac{2\pi}{\omega_q} t\right)\right],$$

where $\{\omega_m\}_{m=1}^q$ denotes the periodicity of the signal and t is time represented by a unit-interval increase. The q period lengths w_m are assumed known and fixed. Since there are 13 time points per transcript in the *Arabidopsis thaliana* dataset (Section 4), we can use up to six sine/cosine pairs of harmonics.

The term ν_{ij} in Equation (1) models the deviation between y_{ij} and the underlying smooth profile. To accommodate for the potential dependence across genes at time j, we adopt a factor model representation:

$$\boldsymbol{\nu}_j = \boldsymbol{\Lambda}\boldsymbol{\eta}_j + \boldsymbol{\epsilon}_j, \tag{2}$$

with $\boldsymbol{\nu}_j = [\nu_{1j}, \ldots, \nu_{pj}]^\top$, $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p]^\top$ is a $p \times k$ factor loading matrix with elements $\{\lambda_{ih}\}_{i=1,\ldots,p;\ h=1,\ldots,k}$, $\boldsymbol{\eta}_j = (\eta_{1j}, \ldots, \eta_{kj})^\top$ is $k \times 1$ vector of latent factors at time j, and $\boldsymbol{\epsilon}_j$ is a residual error. The full model for gene i at time t_j is:

$$y_{ij} = \mathbf{b}_j^\top \boldsymbol{\theta}_i + \boldsymbol{\lambda}_i \boldsymbol{\eta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_i^2), \tag{3}$$

where the first term $\mathbf{b}_j^\top \boldsymbol{\theta}_i$ captures periodic oscillations whereas the second term $\boldsymbol{\lambda}_i \boldsymbol{\eta}_j$ captures across-genes dependence (if present).

Hereafter we follow standard practice and assign a Normal prior to the latent factors at time t_j, $\boldsymbol{\eta}_j \sim N_k(\mathbf{0}, \boldsymbol{I})$. Genes are assumed to be independent given the latent factors, and dependence among genes is induced by marginalising over the distribution of the factors. Therefore, marginally $\mathbf{y}^{(j)} \sim N(\boldsymbol{\Theta}\mathbf{b}_j, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma})$, with $\mathbf{y}^{(j)} = (y_{1j}, \ldots, y_{pj})^\top$, $\boldsymbol{\Theta}$ is the $p \times 2q$ matrix of basis function coefficients, and $\boldsymbol{\Sigma} = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$. In practical applications involving moderate to large p, the number of factors k is typically much smaller than p, thus inducing a sparse characterisation of the unknown covariance matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma}$.

3 Prior elicitation and identification of circadian genes

Inference on parameters $\boldsymbol{\theta}_i$ is the primary interest of our work. Recall that $\boldsymbol{\theta}_{i,m} = (\theta_{i,2m-1}, \theta_{i,2m})$, where $\theta_{i,2m-1}$ is the coefficient of the $(2m-1)$-th sine basis and $\theta_{i,2m}$ is the coefficient of the $2m$-th cosine basis, both harmonics of period w_m. To allow for the correct identification of gene i's periodicity, we need to switch on/off $(\theta_{i,2m-1}, \theta_{i,2m})$ jointly. We assume a latent threshold prior (LTP) [6]:

$$\boldsymbol{\theta}_{i,m} = \tilde{\boldsymbol{\theta}}_{i,m} \mathbb{1}(||\tilde{\boldsymbol{\theta}}_{i,m}|| \geq \varpi_{i,m}), \tag{4}$$

where $\varpi_{i,m}$ is a latent threshold. The idea behind (4) is that the m-th pair of sine/cosine coefficients is shrunk to zero when their (Euclidian) norm falls below a m-th- (and gene-) specific threshold. Further, we model $\tilde{\boldsymbol{\theta}}_i = \{\tilde{\boldsymbol{\theta}}_{i,m}\}_{m=1}^q$ as:

$$\tilde{\boldsymbol{\theta}}_i = \mathbf{W}\boldsymbol{\lambda}_i^\top + \boldsymbol{\delta}_i \qquad \text{and} \qquad \boldsymbol{\delta}_i \sim N_{2q}(\mathbf{0}, \boldsymbol{I}), \tag{5}$$

where $\boldsymbol{\lambda}_i$ is the vector of factor loadings for gene i as in Eq. (3), and $\mathbf{W}$ is a $2q \times k$ matrix such that $\mathbf{W}_j^\top \sim N_k(\mathbf{0}, \boldsymbol{I})$, $j = 1, \ldots, 2q$. Elicitation is completed by placing conditionally conjugate priors on all remaining model parameters, e.g., Gamma priors on precisions, the multiplicative Gamma process shrinkage

prior [1] on the loadings and a uniform prior on the latent thresholds. Posterior inference proceeds via MCMC with conditionally conjugate updating steps.

The LTP eases the identification of circadian genes. Indeed, we estimate the posterior probability of a gene being circadian by counting the proportion of posterior samples for which the coefficients of the 24-hours sine/cosine pair (e.g., $(\theta_{i,2q-1}, \theta_{i,2q})$) are *not* shrunk to zero whilst all the remaning coefficients are switched off:

$$P(\text{Gene } i \text{ is circadian}) = \frac{1}{TS}\sum_{g=1}^{TS} \mathbb{1}\left(\left\{\theta_{i,l}^{(g)}\right\}_{l=1}^{2q-2} \equiv \mathbf{0} \text{ and } \left\{\theta_{i,2q-1}^{(g)}, \theta_{i,2q}^{(g)}\right\} \neq \mathbf{0}\right), \tag{6}$$

where g denotes the iteration number and TS denotes the total number of thinned posterior samples post burn-in. Similarly, the framework above can be used for inference on phase and amplitude of the signals, or for detecting whether a gene exhibits periodicity other than 24-hours.

4 Analysis

We apply our method to the *Arabidopsis Thaliana* dataset [3]. Eight-day-old Columbia seedlings grown under 12-hours-light/12-hours-dark cycles were transferred to constant light at 22°. Plant samples were harvested at 13 time points covering two circadian cycles in 4 hours intervals, starting 26 hours after the last dark-light transition. Here $p = 22810$ genes and $T = 13$ time points.

[2] reports 26 known clock-associated genes in *Arabidopsis*. Among these genes are CCA1 (Circadian Clock Associated 1) and LHY (Late Elongated Hypocotyl), which function synergistically in regulating circadian rhythms of *Arabidopsis*, TOC1 (Timing of Cab Expression 1), which contributes to the plant fitness (carbon fixation, biomass), and ELF4 (Early Flowering 4), which accounts for sustained rhythms in the absence of daily light/dark cycles. We use the 26 well-known circadian genes as a benchmark to evaluate our approach. We compare the proposed approach with its independent version (that is, omitting the latent factor component from Eq. (3)) and JTK_Cycle [4] (chosen for comparison based on performance evaluations presented in [5]). The rankings of the 26 known clock genes are reported in Table 1. All the algorithms were able to identify most of the known clock genes from among their top 25% ranked candidates, and we observe improved performance in terms of placing more genes in the top 1% and 25%. The independent version of our model also performs well. This is likely due to the borrowing of information in modelling the trajectories induced by the Bayesian framework. Estimated trajectories for four clock genes are represented in Figure 1.

We finally remark that we have also evaluated our model on simulation studies, where we could compare its performance to that of competitors in settings of both dependence and independence across synthetic trajectories. These studies show that our construction gives improved performance in identifying rhythmic curves over widely-used rhythmicity detection techniques in both settings. Results are omitted here due to space limitations.

Table 1. Summary of rankings of 26 known clock genes in the *Arabidopsis Thaliana* genome. Genes were ranked by estimated posterior circadian probability for the proposed approach (below Dep.LF) and its independent version (IndepV); by p-value for JTK cycle.

Method	Top 1%	Top 5%	Top 10%	Top 25%	Top 60%
Dep. LF	4	8	14	23	26
IndepV	2	6	13	22	25
JTK cycle	1	8	12	17	22

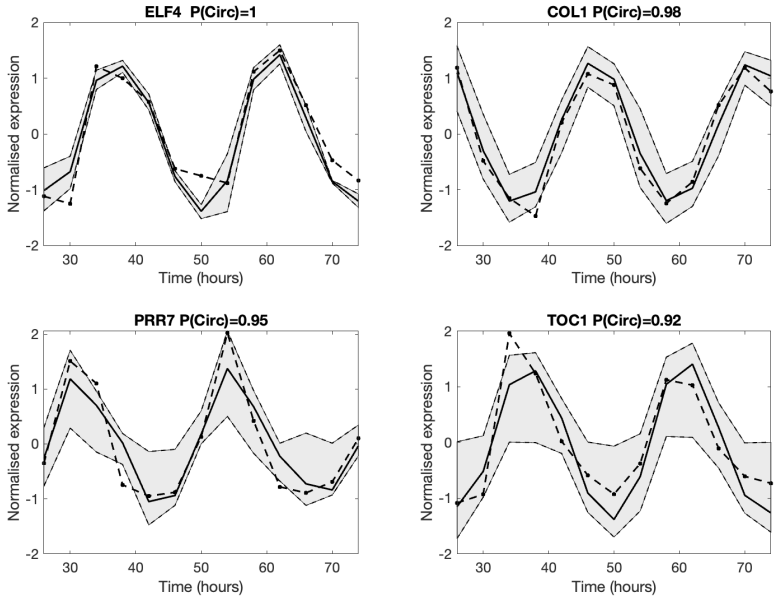

Fig. 1. Four known clock genes in the *Arabidopsis* dataset ranking top by estimated posterior probability of being circadian. The black dashed trajectory connects the true expression levels (dots), the solid black line represents the estimated posterior mean trajectory and the shaded grey area represents point-wise 95% credible intervals around the estimated posterior mean trajectory.

5 Conclusions

In this manuscript, we presented a Bayesian method for periodicity detection. The method employs a Fourier basis expansion coupled with a variable selection prior on the basis coefficients to model gene expression trajectories and identify circadian genes. The core statistical contribution consists in accommodating for

the potential dependence in the trajectories via latent factor modelling. We apply our technique to a well studied gene expression dataset, and its performance is line with (if not better than that) of a widely-used rhythmicity detection technique that does not directly accommodate for dependence across trajectories.

The results presented in this manuscript should be considered as preliminary results, and various extensions could be considered. For example, in animal (e.g., mice) studies, mice could be given a stimulus at the beginning of the experiment. The stimulus may produce local deviations in expression levels, and these deviations may manifest at different times across genes and last for a different amount of time. Deviations could be accommodated for by including an additional local bases decomposition term in Eq. (3). Further, other priors for the basis coefficients or for the latent factor model could also be considered. One drawback of accommodating dependence across curves is the increased computational demand, although the latent factor representation is a convenient way of doing so. Gene-specific updating steps can also be parallelised to speed up posterior computation. Although developed for detecting circadian genes, we finally remark that the approach can be applied to any dataset where the inferential goal is that of detecting periodicity of curves. Indeed, we are currently testing our model to a novel EMG dataset for the identification of periodic bursts in spinae muscles of violin players. Extensions and results will be collected in future research.

References

1. Bhattacharya, A., Dunson, D. B.: Sparse Bayesian infinite factor models. Biometrika. 98, 291–306 (2011).
2. Dodd, A. N., Gardner, M. J., Hotta, C. T., Hubbard, K. E., Dalchau, N., Love, J., Assie, J.-M., Robertson, F. C., Jakobsen, M. K., Goncalves, J., Sanders, D., Webb, A. A. R.: The *Arabidopsis* circadian clock incorporates a cADPR-based feedback loop. Science. 318, 1789–1792 (2007).
3. Edwards, K. D., Anderson, P. E., Hall, A., Salathia, N. S., Locke, J. C. W., Lynn, J. R., Straume, M., Smith, J. Q., Millar, A. J.: Flowering locus C mediates natural variation in the high-temperature response of the Arabidopsis circadian clock. The Plant Cell. 18, 639–650 (2006).
4. Hughes, M. E., Hogenesch, J. B., Kornacker, K.: JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. J. Bio. Rhythms. 25, 372–380 (2010).
5. Mei, W., Jiang, Z., Chen, Y., Chen, L., Sancar, A., Jiang, Y.: Genome-wide circadian rhythm detection methods: systematic evaluations and practical guidelines. Briefings in Bioinf. 22, 1–13 (2021).
6. Nakajima, J., West, M.: Bayesian analysis of latent threshold dynamic models. J. Busin. Econ. Stat. 31, 151–164 (2013).
7. Straume, M..: DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. Meth. in Enzym. 383, 149–166 (2004).