Frequency Modulated Möbius Model Accurately Predicts Rhythmic Signals in Biological and Physical Sciences

Cristina Rueda[1], Yolanda Larriba[1], and Shyamal D. Peddada*[2]

[1]Department of Statistics and Operations Research. Universidad de Valladolid, Valladolid, Spain

[1]Department of Biostatistics, Public School of Health. University of Pittsburgh, Pittsburgh, USA

May 20, 2019

Abstract

Motivated by applications in physical and biological sciences, we developed a Frequency Modulated Möbius (FMM) model to describe rhythmic patterns in oscillatory systems. Unlike standard symmetric sinusoidal models, FMM is a flexible parametric model that allows deformations to sinusoidal shape to accommodate commonly seen asymmetries in applications. FMM model parameters are easy to estimate and the model is easy to interpret complex rhythmic data. We illustrate FMM model in three disparate applications, namely, circadian clock gene expression, corticoptropin levels in depressed patients and the temporal light intensity patterns of distant stars. In each case, FMM model is demonstrated to be flexible, scientifically plausible and easy to interpret. Analysis of synthetic data derived from patterns of real data, suggest that FMM model fits the data very well both visually as well as in terms of the goodness of fit measure total mean squared error. An R language based software for implementing FMM model is available.

1 Introduction

Periodic data arise in a variety of contexts, such as the circadian clock, cell-cycle, hormone levels, astrophysics, although the scientific question of interest varies according to

*Corresponding author: sdp47[at]pitt.edu

the application. In the case of gene expression studies involving cell-cycle or circadian clock (chronobiology), researchers are typically interested in identifying genes with rhythmic patterns, as those shown in panels (a) and (b) of Figure 1, and various statistical parameters associated with them, whereas astrophysicists are often interested in classification of stars using temporal patterns of light emitted from them (panels (c) and (d) in Figure 1). Unlike the typical daily stock market or weather patterns data, these data are generally less dense. Secondly, the parameters and questions of interest in time series analysis are generally different from the parameters and questions asked in oscillatory data considered in this paper, such as cell-cycle, circadian clock etc.

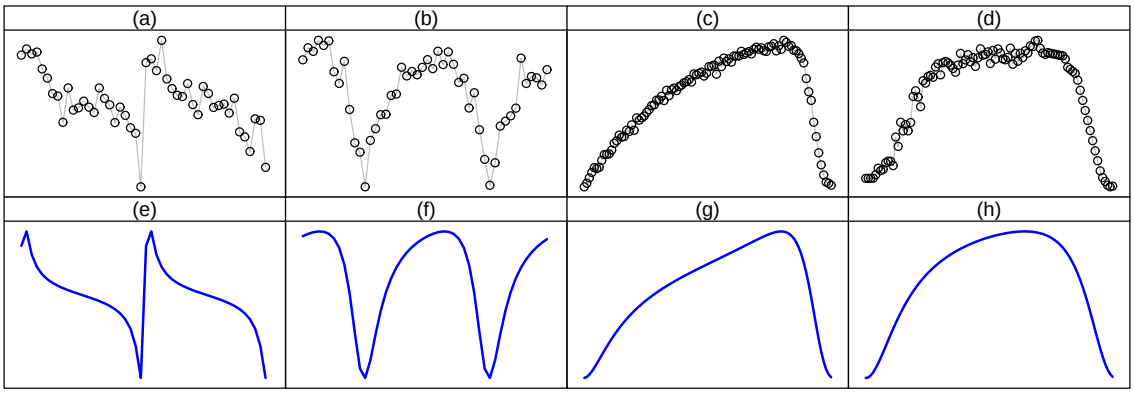


Figure 1: Top: Temporal gene expression patterns along two periods of the circadian genes (a) Iqgap2 and (b) Lonrf3. Temporal patterns of light emitted from (c) Fundamental Cepheides and (d) Mira variable stars. Bottom: FMM fittings for panel (a), (b), (c) and (d).

There are two classes of methods in the literature. One class of methods describes shapes using mathematical inequalities, called order restrictions [1]. A strength of these methods is that they are very flexible because they do not rely on a mathematical function to describe shape. [1] characterized the up-down-up (or down-up-down) patterns in terms of order around a unit circle called the *circular signals*. They demonstrated that these order restrictions-based methods describe rhythmic patterns better than the standard methods used in the literature. However, a weakness of these order restrictions-based methods is that they are not designed to estimates important parameters of rhythmic patterns.

The second class of methods, the focus of this paper, are based on a mathematical function such as variations of cosine function. A commonly used model is the Cosinor model (COS) [2]. Whenever they fit the data well, these methods are useful for describing various characteristics of rhythmic patterns.

The COS model is characterized by a phase and amplitude within a period and is

member of a family of models called *monocomponent* models. Following [3], a rhythmic signal can be described by a sum of *monocomponent* models, each defined by a phase and an amplitude. A wide range of representations of rhythmic signals exist in the literature. Representations vary in the number of monocomponents and whether the amplitude and/or phase are considered to be fixed or variable. In particular, representations with constant amplitude and variable phase are known as Frequency Modulated (FM) representations. For a review on this subject one may refer to [3–6].

One of the popular and widely used representations is the Fourier Decomposition (FD) which is a multicomponent representation where each component has a fixed time amplitude. The COS model is a special form of FD with only one component and it is an appropriate model if the expected functional response is sinusoidal within a period. If a gene displays two peaks (or troughs) within a single period, such as a Quasi-cyclical pattern [7], then the two-component FD model, denoted by FD^2, is potentially a useful model. However, a problem with FD^2 model is that it can potentially introduce two peaks (or troughs) even though scientifically only one peak within a period is justified.

Astrophysicists conduct temporal studies to investigate the properties of light patterns emitted by distant stars [8] and classify them into groups. In some cases, temporal patterns of the observed light intensities from these stars display asymmetrical shapes (panels (c) and (d) in Figure 1) that cannot be captured by COS or even two component models such as FD^2. Since the family of FD models is rigid, researchers use a large number of components, as many as 10 or even 15 in some cases, to capture the shape of light patterns. Despite using that many components, the asymmetries in the data cannot always be captured by FD models. Furthermore, with increase in the number of components one may lower the bias but increase variability, resulting in over-fitting issues. On the other hand, fewer components may result in over smoothed curves with large bias but low variability. Observed light patterns from each star suggest one or at most two oscillations within a period. Thus, higher order FD models may not be ideal to describe these temporal data. Moreover, other widely used methods to analyse oscillatory signals such as JTK_Cycle [9] and RAIN [10] are nonparametric and do not help to describe the underlying physical phenomenon properly.

Motivated by these limitations of the existing methodologies and urgent need for flexible, scientifically interpretable, parametric models for rhythmic data, in this article we introduce a novel model called Frequency Modulated Möbius (FMM) model.

Suppose $X(t_i)$, $t_1 < t_2 < \ldots < t_n$, are real valued time course observations. We model it using a Möbius phase, as follows.

Definition 1. *FMM model* .

$$X(t_i) = \mu(t_i) + e(t_i) = M + Acos(\phi(t_i)) + e(t_i), i = 1, ..., n;$$

1. $M \in \Re$, $A \in \Re^+$,

2. $\phi(t) = \beta + 2\arctan(\omega \tan(\frac{t-\alpha}{2}))$;$\alpha, \beta, \in [0, 2\pi], \omega \in [0, 1]$

3. $(e(t_1), \ldots, e(t_n))' \sim N_n(0, \sigma^2 I)$.

Methodological details that justify the mathematical formulation of this model are included in Subsection 1.1 of the Supplementary Material. Note that rather than using the linear link function for the phase angle ϕ, as done in COS model, we use the Möbius link proposed in [11, 12] which allows for asymmetric shapes as seen in the examples provided in this paper. In particular, Proposition 1 in Subsection 3.1 demonstrates that, with the above choice of the link function, FMM is suitable for describing rhythmic up-down-up (or down-up-down) patterns. The five parameters of the FMM model characterize various aspects of a rhythmic pattern. M and A are intercept and scale parameters measuring the baseline level and the amplitude of the signal, respectively. α is a phase translation parameter while β and ω are parameters describing the shape. Specifically, an extreme spiked signal corresponds to the case $\omega = 0$ and a sinusoidal curve to $\omega = 1$, thus to the COS model, and in that case $\varphi = \beta - \alpha$ is the well known acrophase. Subsection 1.2 in the Supplementary Material includes figures illustrating the deformations from sinusoidal to spiked shapes in terms of the parameters and a detailed discussion about the parameters, respectively.

Other important parameters that are of practical use are peak and trough times, denoted by t_U and t_L, respectively. They are derived from FMM as follows:

$$t_U = \alpha + 2\arctan(\frac{1}{\omega}\tan(\frac{-\beta}{2}))$$
$$t_L = \alpha + 2\arctan(\frac{1}{\omega}\tan(\frac{\pi-\beta}{2})),$$

and the values of the signal at these points are derived as:

$$Z_U = M + A$$
$$Z_L = M - A$$

It is important to note that FMM is a nonlinear parametric regression model. Asymptotic properties of estimators of parameters of nonlinear models, such as asymptotic unbiasedness and consistency are well-known in the literature [13]. Thus, asymptotic likelihood ratio tests and confidence intervals (CI) for individual parameters can be derived using standard asymptotic statistical methods [13].

2 Results

We illustrate and discuss the performance of FMM model using real and synthetic temporal data. For real data, we used publicly available data on (i) circadian clock gene expression, (ii) corticoptropin hormonal measurements in clinically depressed patients,

and (iii) light intensities from variable stars. Specifically, gene expression data were originally recorded along two periods and then they were averaged. Hormonal and star data were directly given along a unique period but corresponding to averaged values too. In each case, we compared FMM with FD based methodologies. In particular, we focus on COS and FD^2.

To further validate our findings, we generated synthetic data using parameters derived from the above real data. Due to space limitations, results of simulation study are relegated to Subsection 3.3 in the Supplementary Material. Results therein reinforce our findings of this section that FMM is indeed more flexible and a better fitting model than the existing models.

2.1 Circadian Gene expression patterns

Several researchers have studied the two-period circadian clock gene expression data obtained from in-vivo experiments on mouse liver and pituitary gland, and in-vitro experiment data on NIH3T3 and the U2OS human cell lines. All four data are available from the NCBI GEO website (http://www.ncbi.nlm.nih.gov/geo/). These are very comprehensive data which are useful for evaluating the performance of a model fitting strategy.

Here, $Mmse$ denotes the averaged mean squared error (mse) of each data set, details for model performance measures are given in Subsection 3.3. Since FMM is by design a single peak (trough) model with a more flexible shape than COS, we expect FMM to perform the best followed by COS model. As seen in Table 1, in all four data sets, FMM has the smallest $Mmse$ compared to COS and FD^2 models. In some cases the reduction in $Mmse$ of FMM relative to COS was dramatic. We notice a 33% reduction in the case of U2OS cell-line data and a 41% reduction in the case of mouse liver data.

The above dramatic performance of FMM relative to COS function is graphically illustrated in Figure 2 for a sample of rhythmic circadian genes. In each case, not only FMM fits the data better than COS, but more importantly, the times to peak gene expression estimated by the two methods are dramatically different, the difference ranging from 4 to 7 hours approximately, see for instance panels (a) and (c) in Figure 2. In their seminal work [14] noted that phases of circadian clock genes play a key role in drug delivery to patients, and that it is critical to estimate the phases of circadian clock genes as accurately as possible. In view of [14], an error in the range of 4 to 7 hours could potentially have important clinical and pharmacological effects. From the figures displayed, it is clear that FMM provides a better description of these genes. The performance of FMM was equally surprising in the case of quasi-cyclical shaped pattern (patterns with more than one local maximum or minimum within each period), by design, FD^2 is expected to have the smallest $Mmse$. However, surprisingly, FMM was very competitive with FD^2 in terms of $Mmse$. Apart from the mouse liver data, in all other cases FMM had smaller $Mmse$ than FD^2. Again, we provided plots of a subset of genes in Figure 3.

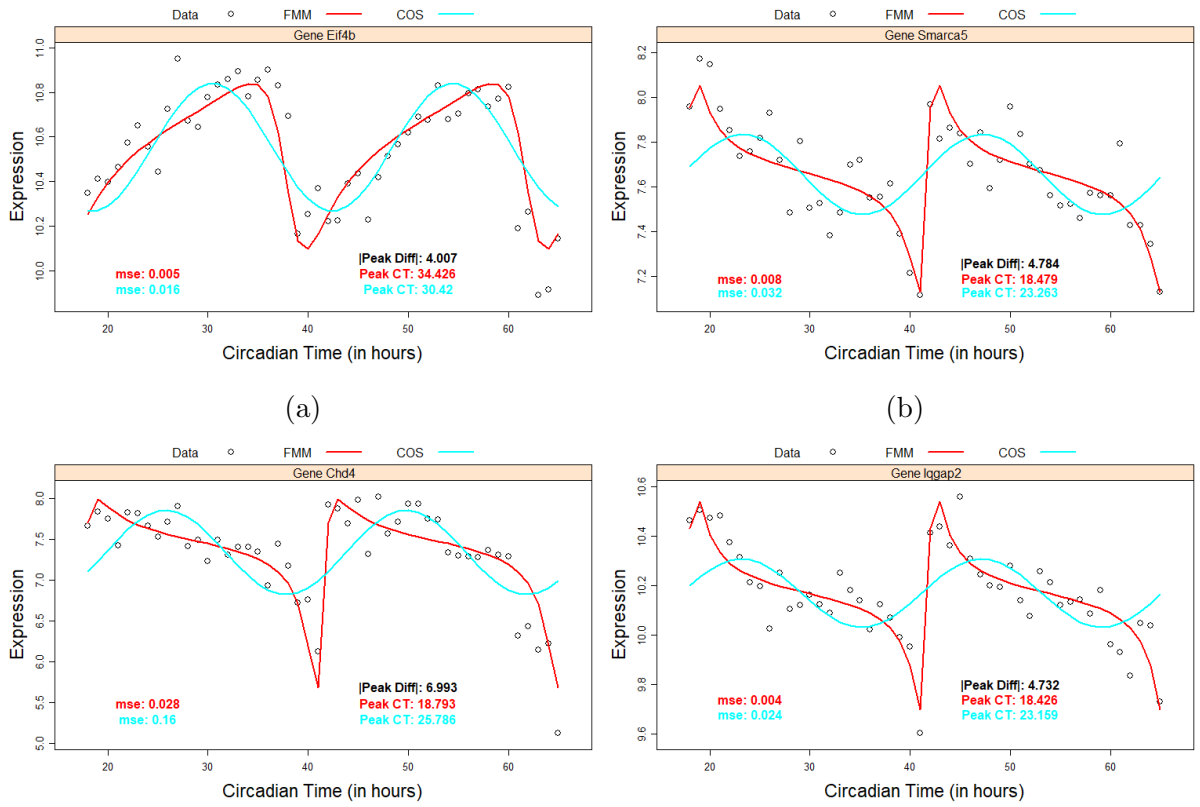


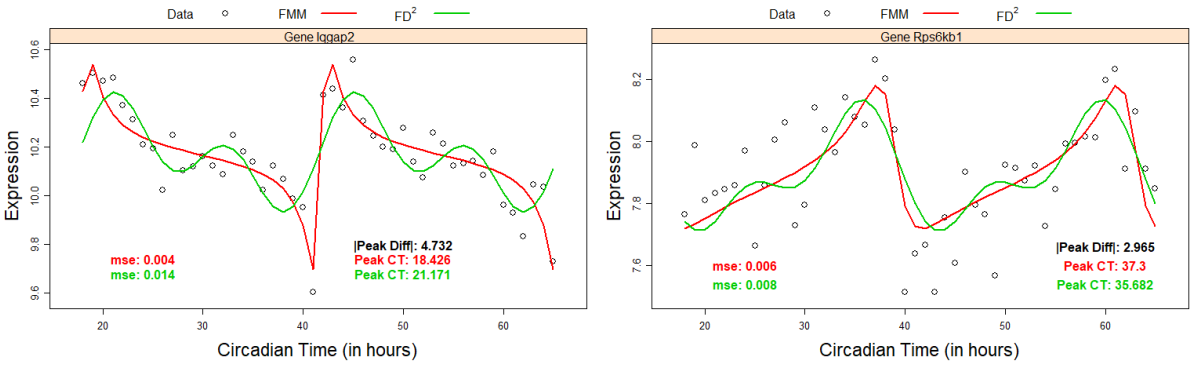
Figure 2: Gene expression (dots) and FMM (red) and COS (light blue) model fittings for the genes from mouse liver: (a) Eif4b, (b) Smarca5, (c) Chd4 and (d) Iqgap2 along two periods of 24 hours. In each panel, mse and circadian time (CT) peak estimates for FMM (red) and COS (light blue) are given as well as the absolute difference between these CT (black).

As we see, FD^2 imposes two peaks by virtue of its functional form when clearly the data does not display two peaks. Secondly, these peaks are not biologically interpretable. On the other hand FMM seems to fit the data better with a single peak.

Thus, these examples exemplify the performance of FMM to describe temporal patterns of circadian clock genes. Subsection 3.1 of the Supplementary Material includes more details about the distribution of the estimated values of α, β and ω.

2.2 Temporal patterns of corticoptropin levels in clinically depressed patients

In this section we illustrate the performance of FMM for modeling hourly corticoptropin levels during a day in patients suffering from major clinical depression. We used data from [15] which consisted of 3 groups of subjects where 11 were patients with psychotic major depression (Pmd), 38 were patients with nonpsychotic major depression (Npmd), and 33 were healthy controls. From the fitted curves in Figure 4 it is apparent that COS



(a) (b)

Figure 3: Gene expression (dots) and FMM (red) and FD^2 (green) model fittings for the genes from mouse liver: (a) Iqgap2 and (b) Rps6kb1 along two periods of 24 hours. In each panel, mse and circadian time (CT) peak estimates for FMM (red) and FD^2 (green) are given as well as the absolute difference between these CT (black).

Table 1: $Mmse$ and $SDmse$ for FMM, FD^2 and COS obtained for genes in Mouse Liver, Pituitary gland, NIH3T3 cell lines and U2OS human cells by type of pattern (cyclical and quasi-cyclical) proposed in [7].

		Cyclical			Quasi Cyclical			ALL		
		n	$Mmse$	$SDmse$	n	$Mmse$	$SDmse$	n	$Mmse$	$SDmse$
Liver	FMM	9167	0.0126	0.0172	92	0.0547	0.1056	9259	0.0131	0.0205
	FD^2	9167	0.0138	0.0189	92	0.0357	0.0616	9259	0.0140	0.0199
	COS	9167	0.0204	0.0296	92	0.0868	0.1683	9259	0.0211	0.0345
Pituitary	FMM	3363	0.0142	0.0179	18	0.0193	0.0142	3381	0.0142	0.0179
	FD^2	3363	0.0168	0.0238	18	0.0248	0.0149	3381	0.0168	0.0237
	COS	3363	0.0192	0.0279	18	0.0328	0.0237	3381	0.0193	0.0279
NIH3T3	FMM	1411	0.0164	0.0225	13	0.0257	0.0372	1424	0.0165	0.0227
	FD^2	1411	0.0211	0.0272	13	0.0282	0.0374	1424	0.0211	0.0274
	COS	1411	0.0263	0.0379	13	0.0358	0.0551	1424	0.0264	0.0381
UOS2	FMM	906	0.0166	0.0242	8	0.0209	0.0121	914	0.0167	0.0241
	FD^2	906	0.0209	0.0325	8	0.0251	0.0153	914	0.0209	0.0324
	COS	906	0.0245	0.0367	8	0.0273	0.0167	914	0.0245	0.0366

model does not fit the data as well as FMM. It is also apparent that FD^2 performs nearly as well as FMM in the case of Pmd and control groups but does not fit as well as FMM in the case of Npmd group. Furthermore, among the three models, FMM is the best fitting model because it has the smallest mse in all three patient groups (Table 2). We also estimated two important parameters relevant for this hormonal study, namely, t_U: the peak time and Z_U: the mean hormone level corresponding to the peak time. These estimates are provided in Table 2 and confidence intervals for pairwise differences between groups are in Table 3.

Consistent with the plots in the Figure 4, we notice that t_U values obtained from FMM are smaller than those of the other two methods.

Besides, 90% CI for pairwise differences in Table 3 derived under FMM show significantly different Z_U values between the three groups, which is not detected with the other approaches. This is a clinically relevant finding because it suggests that there are differences in the mean peak hormone levels among the three groups with control group having the smallest peak followed by nonpsychotic major depression group and psychotic major depression groups. The psychotic major depression has the largest peak. Thus FMM model allows us to discover a trend in the peak levels of corticotropin with the disease severity.

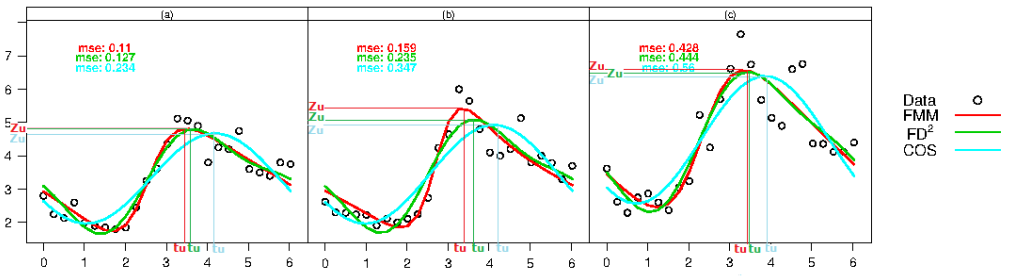


Figure 4: Observed data (dots) and fitted FMM (red), FD^2 (green) and COS (light blue) models by patient group: (a) Control, (b) Npmd and (c) Pmd

Table 2: mse and estimates for σ^2, t_U and Z_U obtained from FMM, FD^2 and COS, for each patient group

	FMM			FD^2			COS		
	mse	$\hat{t}_U$	$\hat{Z}_U$	mse	$\hat{t}_U$	$\hat{Z}_U$	mse	$\hat{t}_U$	$\hat{Z}_U$
Control	0.110	3.489	4.816	0.127	3.659	4.803	0.234	4.186	4.676
Npmd	0.159	3.346	5.431	0.235	3.643	5.096	0.347	4.119	4.932
Pmd	0.428	3.348	6.575	0.444	3.449	6.525	0.560	3.864	6.401

Table 3: Bootstrap 90% CI for pairwise t_U and Z_U, differences obtained from FMM, FD^2 and COS.

t_U	FMM	FD^2	COS
Control vs Npmd	[-0.150,0.424]	[-0.225,0.271]	[-0.139,0.298]
Control vs Pmd	[-0.218,0.631]	[-0.139,0.541]	[0.080,0.559]
Pmd vs Npmd	[-0.353,0.399]	[-0.235,0.521]	[-0.043,0.530]
Z_U	FMM	FD^2	COS
Control vs Npmd	[0.153,1.213]	[-0.101,0.777]	[-0.214,0.634]
Control vs Pmd	[1.164,2.451]	[1.259,2.323]	[1.154,2.241]
Pmd vs Npmd	[0.422,1.865]	[0.854,2.076]	[0.861,1.952]

2.3 Temporal patterns of light emitted by stars

Light intensities of stars from six different star groups [8], namely, RR Lyraes (RRab and RRc), Cepheids [Fundamental, (FU) and Overtone (FO)], Mira, and Eclipsing Binary (EB), are investigated in this section. The data consisted of 17,606 variable stars with 100 time points on each. FMM is more flexible fitting a wide range of patterns seen in the six groups of stars. On the other hand FD based methods, such as COS and FD^2, fit well only when the data are approximately symmetric sinusoidal in shape.

Temporal plots of a sample of typical curves from each of these groups are provided in Figure 5. A representative from the RRc group is not shown because RRc patterns are similar to those of FO group; two different representative patterns from EB are provided instead. We overlaid on each figure the fitted curves obtained from COS, FD^2 and FMM methodologies along with respective mse values. Except for one of the EB subclasses (panel (f)) where FD^2 performs best, in all other cases, FMM displays great flexibility to fit the data. Moreover, as seen in panels (a), (b), (d) and (e) of Figure 5, the COS and FD^2 perform poorly to fit asymmetric patterns.

The estimated $Mmse$ values for the three models are summarized in Table 4 for each star group. In almost all cases, FMM has the smallest estimated $Mmse$, suggesting that it fits the data best for almost all groups. The only slight exception is the star group EB, but even there, the $Mmse$ for FMM is very slightly larger than that of FD^2, 0.021 versus 0.020, a difference of 0.001. In comparison to FD^2 and COS, the performance of FMM is best in the cases of RRab and FU.

Table 4: $Mmse$ for FMM,FD^2 and COS by star group

	Number of stars	FMM	FD^2	COS
RRab	5835	0.004	0.009	0.019
RRc	1751	0.004	0.004	0.005
FU	1829	0.001	0.005	0.016
FO	1228	0.002	0.002	0.003
Mira	2878	0.005	0.006	0.015
EB	4085	0.021	0.020	0.042

In addition to fitting models, researchers are typically interested in classifying stars

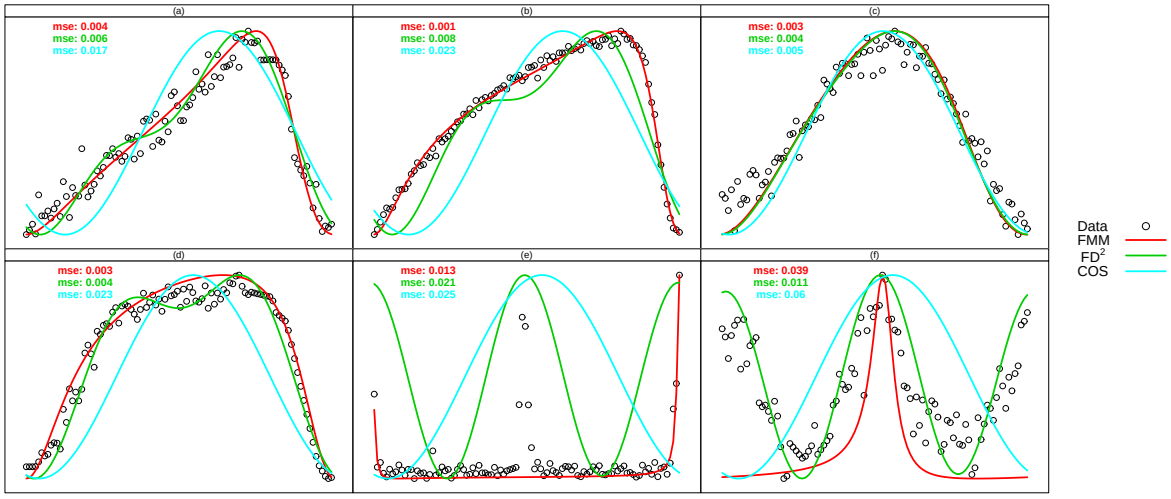


Figure 5: Selected temporal light patterns (dots) emited from: (a) RRab, (b) FU, (c) FO, (d) Mira and (e,f) EB classes of variable star together with FMM (red), COS (light blue) and FD^2 (green) model fittings.

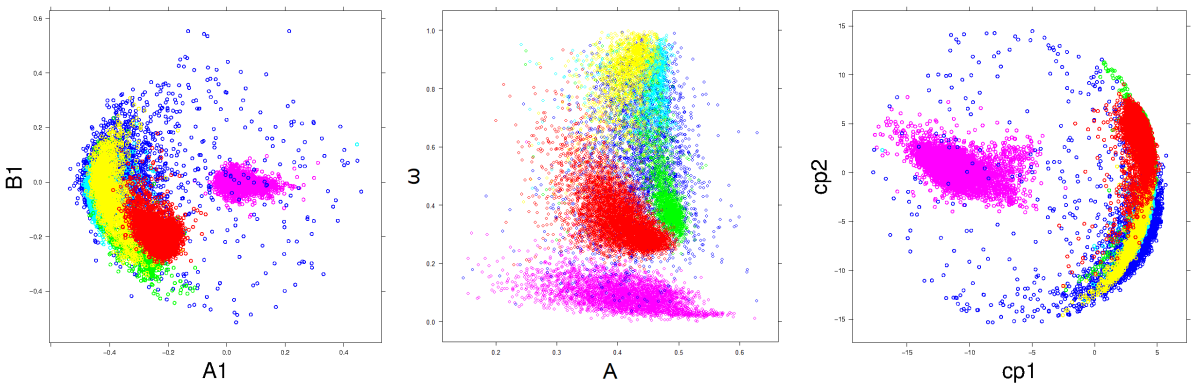
into various groups. PCA (Principal Component Analysis) and FD have been the two most popular approaches until now [8, 16].

We compared the performance of FMM, FD^2 and PCA in classifying samples using standard canonical discriminant analysis with two variables from each model and classification errors estimated using leave one out cross-validation. The variables used for discrimination were the first two principal components, $PC1$, $PC2$ from PCA; the two parameters with the highest discriminative power, those associated with the first component, denoted by A_1 and B_1, from FD; and ω and A from FMM. The scatterplots for the three pairs of variables are shown in Figure 6 where it is shown that $A1, B1$ and $PC1, PC2$ clearly separate EB from the rest, but they are not very successful in separating the remaining groups. On the other hand, from panel (b) in Figure 6, it is very clear that the FMM model based parameters perform well in separating all groups of stars. In particular, the shape parameter ω plays a critical role in discriminating all groups of stars. In fact, smaller misclassification rates are obtained when FMM variables are used, as it is shown in Table S2 of the Supplementary Material.

Subsection 3.2 of the Supplementary Material provides graphical displays and comments on the distribution of the estimated values of α and β for the star set by groups.

3 Methods

We begin with some definitions and notations. In the following we assume $t \in [0, 2\pi]$; if observed times takes values on a real interval then $t' \in [t_0, T+t_0]$, $t = \frac{(t'-t_0)2\pi}{T}$, $t \in [0, 2\pi]$.



(a) (b) (c)

Figure 6: Scatter plots for pairs of parameters from: (a) FD^2, (b) FMM and (c) PCA . The color identify the group: EB (pink), Mira (dark blue), FO (light blue), FU (green), RRc (yellow) and RRab (red).

3.1 Circular signal and FMM

It is generally accepted that for a given oscillatory phenomenon, there exists an underlying complex valued signal. Even more, [5], among others, argues that a physical phenomenon is not entirely modelled unless the complex signal it is related to, has been defined. In this paper we deal with periodic signals, which are described as complex functions of time, which we denote as $S(t)$, $t \in [0, 2\pi]$.

Definition 2. *A complex-valued signal* $S(t)$

$$S(t) = \mu(t) + i\nu(t) = \rho(t)e^{i\phi(t)}, t \in [0, 2\pi]. \tag{1}$$

From the complex formulation, a model for a real signal is derived as:

$$Re(S(t)) = \mu(t) = \rho(t)cos(\phi(t)), t \in [0, 2\pi].$$

The latter term in equation (1) is known as the *quadrature form* of the signal $S(t)$, where $\rho(t)$ and $\phi(t)$ are the signal's *amplitude* and *phase* respectively. The derivative of $\phi(t)$ is known as Instantaneous Frequency (IF), a parameter that is expected to be non negative in applications, as argue [5].

When $\nu(t)$ is unknown, there are infinite pairs $\rho(t), \phi(t)$ for which $\mu(t)$ may be equivalently described. An important subclass, is that of analytic signals. In particular, one of the elemental is the *Fourier atom* which is defined using the Möbius transform. Besides, analytic signals having a non negative IF and constant amplitude are often used by researchers in applications due to their interpretability and simplicity. Specifically, the real signal corresponding to these latter signals is a *monocomponent*. Definitions of these signals and additional theoretical details are given in Section 1 of the Supplementary

Material.

We now introduce *circular signals* as follows:

Definition 3. *Circular signal* in the Euclidean space

$\mu(t) \in \mathbb{R}$, $t \in [0, 2\pi]$ is circular iff $\exists\ t_U, t_L$ such that

if $t_U \leq t_L$: $\mu(t) \geq \mu(t'), t_U \leq t \leq t' \leq t_L$, and $\mu(t) \leq \mu(t')$, $0 \leq t \leq t' \leq t_U$; $t_L \leq t \leq t' \leq 2\pi$.

or equivalently

if $t_U \geq t_L$: : $\mu(t) \leq \mu(t'), t_L \leq t \leq t' \leq t_U$, and $\mu(t) \geq \mu(t')$, $0 \leq t \leq t' \leq t_L$; $t_U \leq t \leq t' \leq 2\pi$.

Without loss of generality, we assume that $t_U \leq t_L$. In the Euclidean space, such a signal is also called an up-down-up signal (*resp.* down-up-down) [1], as it monotonically increases (*resp.* decreases) to t_U (*resp.* t_L) and then decreases (*resp.* increases) to t_L (*resp.* t_U) before decreasing (*resp.* increasing) again. As illustrated in Subsection 2.2, t_U is an important parameter in applications because it is the time to first peak.

In addition, a *circular signal* on the unit circle is a signal that follows the circular order (see [17] for a definition on circular order),

Definition 4. *Circular signal* on the unit circle

$\phi(t) \in [0, 2\pi]$, $t \in [0, 2\pi]$ is circular iff $\phi(t) \leq \phi(t'), 0 \leq t \leq t' \leq 2\pi$ (*resp.* $\phi(t) \geq \phi(t'), 0 \leq t \leq t' \leq 2\pi$)

The most popular *circular signal*, and also the simplest one, is the sinusoidal signal: $\mu(t) = cos(t + \varphi)$. Its corresponding *circular signal* is $\phi(t) = t + \varphi$.

It is straight forward to derive that, if $e^{i\phi_a(t)}$ is a Fourier atom, $Re(e^{i\phi_a(t)})$ is a *circular signal*.

Next, we provide a useful characterization of FMM to demonstrate the relationship between FMM models and Fourier atoms. In particular, Proposition 1 demonstrates that FMM is restricted to *circular signals* and that the IF is non negative. Thus, the FMM model is appropriate for describing typical periodic up-down-up signals

The FMM phase can be equivalently derived from:

$$e^{i\phi(t)} = e^{i\varphi}\frac{e^{it} + a}{\overline{a}e^{it} + 1},$$

where, $\varphi \in \Re$ and $a = re^{iv} \in \mathbb{C}$. Then, the relationship between this formulation and FMM model (see Definition 1) is given by:

$$v = \alpha, \varphi = \beta - \alpha \text{ and } r = \frac{1 - \omega}{1 + \omega}.$$

The equivalence formulation above is also stated in the seminal papers of circular regression: [11, 12].

Proposition 1. Let $\mu(t) = M + Acos(\phi(t))$, $\phi(t) = \beta + 2\arctan(\omega \tan(\frac{t-\alpha}{2}))$ and $t \in [0, 2\pi]$, then:

1. $\mu(t)$ is a *circular signal* in the Euclidean space.
2. $\phi(t)$ is a *circular signal* in the unit circle.
3. $\phi'(t) = \frac{\omega}{2(1+\omega^2 sin^2(\frac{t-\alpha}{2}))}$

The proofs follow immediately from the definitions.

3.2 Estimation algorithm

A two-step algorithm is developed to estimate FMM parameters. First, initial parameter estimation is given by solving a least square problem along the lines of proposed in [2]. Second, we used Nelder-Mead optimization method [18] to obtain the final FMM parameter estimates, see Section 2 in the Supplementary Material for details. The proposed methodology is not limited by which optimization method is used. Based on our experiences with complicated objective functions involving angular data [19, 20], as well as the data analyzed in this paper, we find Nelder-Mead to provide estimates that fit data well. It tends to successfully avoid local solutions. For example, see figures presented in this paper.

3.3 Model performance measures

To assess the performance of various models, we use the total mean squared error (mse) over all observed times as a criterion. This is a common measure of goodness of fit used by statisticians when assessing the performance of an estimator or a model and routinely discussed and used in textbooks [21]. Smaller the value suggests the better the model fits the data. More precisely, in simulations, mse is a measure of distance from the known signal μ as $mse = \sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2/n$. In practice, the mse is calculated as $mse = \sum_{i=1}^{n}(\hat{\mu}_i - X_i)^2/n$.

In addition, mse estimates for a specific parameter θ are denoted as $mse(\theta)$ in simulations. Finally, when mse values are averaged across different scenarios or individuals, an M is added. Thus, $Mmse$ is the average mse over all measurements available. The standard deviation of mse is denoted by $SDmse$.

3.4 Time to first peak and trough in COS and FD

COS model has well defined maximum and minimum $t_U = -\varphi$ and $t_L = \pi - \varphi$ respectively. However, the computation of extremes is not trivial for FD. In fact, FD^N model has multiplicity of N extrema (see [22] for details). There is no close form expressions for

t_U and t_L and they are numerically derived as the values where $\mu(t)$ reaches its maximum and minimum using an optimization algorithm for the analysis given in Section 2 of this work.

4 Discussion

As seen in this paper, oscillatory systems arise naturally in a wide range of applications including biology, medicine, pharmacology, astronomy and so on. An oscillatory system consists of several components that display rhythmic temporal patterns. The temporal patterns and the associated parameters, such as the amplitude and phase, have critical scientific importance and implications. For example, as demonstrated in [14] the efficacy of a drug in treating a patient may depend upon the time of the day the drug is delivered, and this determination is made based on the phases of some circadian clock genes. Thus, in all such applications it is not only important to determine all components (e.g. genes) that display a temporal rhythmic pattern, but it is critically important to derive an appropriate parametric model and estimate the associated parameters correctly. A poor choice of the model may result in wrong estimates of phase and amplitude that may have important downstream implications. For example, as we saw in the circadian clock genes example discussed in this paper, a poor model may result in a 4 to 7 hours difference in the phase estimate relative to what might be the true phase. In view of [14] findings this might have major clinical and pharmacological impact on when patient receives a drug.

A common parametric model used in almost all applications to fit a temporal rhythmic data is the cosinor model (COS). While it is easy to fit and interpret, it is a very rigid model in the sense that the observed temporal signals are required to have a sinusoidal shape, which is intrinsically symmetric. As we saw in the examples presented in this paper, the temporal patterns of rhythmic signals do not always follow this rigid structure. In fact, as observed in [11], it is common to have a nonlinear relationship or link between an angular parameter and time. Although, Fourier decomposition (FD) was developed in the literature to provide some flexibility from COS, intrinsically it too has a symmetric shape. Secondly, because it is a linear combination of several sinusoidal functions, it may induce multiple peaks (or troughs) within a period. In many applications, especially in the circadian clock or cell-cycle, those multiple peaks are hard to interpret.

The primary contribution of this paper is to derive a flexible parametric model that allows deformations to the sinusoidal shape and contains easy to interpret parameters. As demonstrated in this paper, the model performs extremely well in a very disparate types of applications. The model fits circadian clock data, hormonal data as well as light data from distant stars. The rhythmic patterns are very varied and yet in each case the model seems to outperform the existing models. Extensive simulations seem to confirm these findings. It is important to reiterate that we fill an important gap in the literature to derive a flexible parametric model for describing rhythmic patterns that are deformations

to sinusoidal models.

Once an appropriate nonlinear model is derived, as noted in the paper, given decades of literature, statistical inference regarding the parameters of the nonlinear model is routine problem. In this paper we used bootstrap based methodology.

As frequently quoted by modelers, a quote attributed to George Box, "No model is perfect but some models are more useful", the proposed basic FMM model has limitations. Firstly, we have not discussed here the problem of detecting if a component of an oscillatory system is rhythmic or not rhythmic and if it is rhythmic, then whether it is also a sinusoidal. However, as described in Subsection 1.3 of the Supplementary Material, parametric hypothesis testing problems to test the above hypotheses can be easily addressed using FMM.

Secondly, in many studies researchers are interested in fitting nonlinear models after adjusting for covariates. This is particularly true for modeling hormone data. For example, gender and age would be two important factors to consider when modeling hormonal data. The problem can be even more complex when potential interactions may be suspected. Specifically, [15] used COS model but adjusted for important covariates, such as age and gender as additive effects in a linear model. The endocrine system for males and females is fundamentally different. This leads to differences in biological responses and hence it is reasonable to expect males and females to have curves with different shapes. A similar phenomena may occur with age. Compared to sinusoidal models such as COS, an advantage of using FMM in the above formulations is that it allows for deformations to sinusoidal shape. The current formulation of FMM requires further refinements and modifications to model interactions and covariates.

Finally, other important limitation of FMM is that it does not parametrize the period but takes it as a fixed known quantity. While in many examples the period of a cycle is determined by the experimental design, such as in a circadian clock or cell-cycle experiment, there are also examples, such as the EB star data where the period may be poorly determined. However, the period can be formulated as an unknown parameter in the model, then the methodology can be suitably modified by designing a computational intensive algorithm that considers different period values and then chooses the period that results in a smaller total mse.

References

[1] Y. Larriba, C. Rueda, M.A. Fernández, and S.D. Peddada. Order restricted inference in chronobiology. *Submited*, 2019.

[2] G. Cornelissen. Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling*, 11(1):16, 2014. doi: 10.1186/1742-4682-11-16.

[3] B. Boashash. *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*. Elsevier Science, 2016. ISBN 9780123985255. URL `https://books.google.es/books?id=WbYoRC1-lMkC`.

[4] B. Picinbono. On instantaneous amplitude and phase of signals. *IEEE Transactions on Signal Processing*, 45(3):552–560, 1997. ISSN 1053-587X. doi: 10.1109/78.558469.

[5] S. Sandoval and P. De Leon. Theory of the hilbert spectrum. *arXiv*, 2015.

[6] Pushpendra Singh. Comments on the representations of instantaneous frequency using the hilbert transform, direct quadrature and hilbert quadrature. working paper or preprint, 2017.

[7] Y. Larriba, C. Rueda, M.A. Fernández, and S. D. Peddada. Order restricted inference for oscillatory systems for detecting rhythmic signals. *Nucleic Acids Research*, 44 (22):e163, 2016. doi: 10.1093/nar/gkw771.

[8] S. Deb and H.P. Singh. Light curve analysis of variable stars using fourier decomposition and principal component analysis. *A&A*, 507(3):1729–1737, 2009. doi: 10.1051/0004-6361/200912851.

[9] M.E. Hughes, J.B. Hogenesch, and K. Kornacker. JTK CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of Biological Rhythms*, 25(5):372–380, 2010.

[10] P.F. Thaben and P.O. Westermark. Detecting rhythms in time series with rain. *Journal of Biological Rhythms*, 29(6):391–400, 2014.

[11] T.D. Downs and K.V. Mardia. Circular regression. *Biometrika*, 89(3):683–697, 2002.

[12] S. Kato, K. Shimizu, and G. Shieh. A circular-circular regression model. *Statistica Sinica*, 18:633–645, 2008.

[13] G.A.F. Seber and C.J. Wild. *Nonlinear regression*. John Wiley & Sons, New York, 1989.

[14] R. Zhang, N.F. Lahens, H.I. Ballance, M.E. Hughes, and Hogenesch J.B. A circadian gene expression atlas in mammals: Implications for biology and medicine. *PNAS*, 111(45), 2014. doi: 10.1073/pnas.1408886111.

[15] J.A. Posener, C. DeBattista, Williams G.H., H. Kraemer, B. Kalehzan, and A.F. Schatzberg. 24-hour monitoring of cortisol and corticotropin secretion in psychotic and nonpsychotic major depression. *Archives of General Psychiatry*, 57(8):755–760, 2000. doi: 10.1001/archpsyc.57.8.755.

[16] K.B. Johnston and H.M. H.M. Oluseyi. Generation of a supervised classification algorithm for time-series variable stars with an application to the linear dataset. *New Astronomy*, 52:35 – 47, 2017. ISSN 1384-1076. doi: 10.1016/j.newast.2016.10.004.

[17] N.I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1993.

[18] J.A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965. doi: 10.1093/comjnl/7.4.308.

[19] S.D. Peddada and T.C. Chang. Bootstrap confidence region estimation of the motion of rigid bodies. *J. of Amer. Statist. Assoc.*, 81:231–241, 1996.

[20] D. Liu, D.M. Umbach, S.D. Peddada, L. Li, P.W. Crockett, and C.R. Weinberg. A random periods model for expression of cell-cycle genes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7240–7245, 2004.

[21] D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to Linear Regression Analysis (5th ed.)*. Wiley & Sons, 2012.

[22] J.P. Boyd. Computing the zeros, maxima and inflection points of chebyshev, legendre and fourier series: solving transcendental equations by spectral interpolation and polynomial rootfinding. *Journal of Engineering Mathematics*, 56(3):203–219, 2006. ISSN 1573-2703. doi: 10.1007/s10665-006-9087-5.

Acknowledgments

The authors gratefully acknowledge the financial support received by the Spanish Ministerio de Ciencia e Innovación and European Regional Development Fund; Ministerio de Economía y Competitividad grant [MTM2015-71217-R to CR] and Spanish Ministerio de Educación, Cultura y Deporte [FPU14/04534 to YL].

Author contributions

CR: Conceived aims, theoretical proporsal, conceptual design, data analysis, interpretation of the results, wrote and approved manuscript, proposed future applications; YL: Involved in the discussion about some aspect of the methodology, processed original data, generated simulations, analyzed data, interpreted the results, wrote and approved manuscript; and SP: involved in the discussion about some aspect of the methodology, interpretated the results, wrote and approved manuscript, proposed future applications.

Competing interests

The authors declare no potential conflict of interests.

Materials & Correspondence

Materials and correspondence should be sent to Dr. Peddada. email: sdp47@pitt.edu