Computational analysis of rhythmic data using RDA

Arthur Vestu[1], **Lily-Jade Roldao**[1], **Miha Moškon**[2]

[1]*ESIGELEC Graduate School of Engineering, Rouen, France*
[2]*Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia*
E-mail: miha.moskon@fri.uni-lj.si

Abstract

Rhythmic processes can be found in different contexts that range from biological to socio-technical systems. Several computational methods have been introduced to study such processes. However, these have mostly been adapted to work well with specific data and need to be manually adapted for a wider usage. We describe a software framework dedicated to a comprehensive analysis of rhythmic datasets. It integrates different state-of-the-art methods dedicated to the identification and characterisation of rhythmic processes. It allows its users to straightforwardly apply different methods to a selected dataset, and to identify the method yielding the results with the largest relevance in a given context. We demonstrate the application of the proposed framework on two examples. Firstly, we report the results of a benchmarking experiment, which also indicate the classification performance of each of the implemented methods (i.e., classifying measurements among rhythmic and non-rhythmic groups). Secondly, we present the application of the proposed framework on the assessment of rhythmic trends in traffic data reflecting circadian rhythmicity. We believe that the proposed package will find a vast scope of applications in different scientific domains.

1 Introduction

Rhythmic processes, i.e., processes that reflect periodic response, are pervasive in our environment. For example, circadian clocks present biological clocks that display daily (approximately 24-hour period) oscillations and regulate up to half of all genes in an organism [1]. Since a disruption of these rhythms can lead to the development of different diseases, their analysis has gained significant importance in recent years [2]. In this context researchers are combining different experimental techniques with computational approaches [3]. Several computational methods for the identification and characterisation of rhythmic trends have been proposed in recent years [2]. These methods have been mostly adapted to work well with specific biological data, i.e. transcriptomic circadian time-series datasets. However, detection and analysis of rhythmic patterns have become an important aspect also in fields of research outside biology and medicine, such as urban planning. On the other hand, the majority of the state-of-the-art methods for rhythmicity detection and analysis cannot be straightforwardly applied to such datasets.

In this work we describe a computational framework for domain-agnostic analysis of rhythmic datasets. The framework combines different methods for rhythmicity detection and analysis. It accepts input data in standardised formats, and is able to produce publication-ready figures and results in a tabular format for straightforward analysis. Moreover, the framework implements a set of functionalities which can be used for benchmarking experiments. We demonstrate the proposed framework and the methods it incorporates on two case studies. Firstly, we describe a benchmarking experiment of synthetically generated data. Secondly, we describe the application of the framework on the real data presenting the average car-travel paces throughout the day on a selected road segment in the city of Ljubljana. We use the obtained results to discuss the pros and cons of each of the selected methods.

2 Methods

2.1 Identification and characterisation of rhythmic data

Several computational approaches to identify and characterise rhythmic data have been introduced in recent years [2]. In this section we describe some of the most popular methods, which have also been incorporated into the proposed computational framework.

2.1.1 Lomb-Scargle

Lomb-Scargle periodogram (LS) is one of the first algorithms devoted to the identification of rhythmicity in data. It presents a parametric model that identifies oscillations by comparing the data to sinusoidal curves [4]. LS can handle data quality issues like replicates, missing values or uneven sampling.

2.1.2 Cosinor

Cosinor is a trigonometric regression model similarly to LS. Cosinor is a parametric model which allow us to precisely identify oscillatory datasets as well as amplitudes and acrophases of oscillations. Cosinor can produce relevant results even when data quality is poor (irregular intervals, unbalanced data full of outliers) [5].

2.1.3 ARSER

ARSER is similar to Cosinor as it is also a parametric method using harmonic regression [6]. ARSER additionally uses autoregressive spectral estimation to estimate the period of the data. Nevertheless, it is sensitive to data quality issues which can lead to inaccuracies. Moreover, it does not work when there are replicates of data.

2.1.4 JTK_CYCLE

JTK_CYCLE is a non-parametric method that detects oscillations by comparing the ranks of the measured values to a set of specified symmetric reference curves [7]. JTK_CYCLE works well with replicates and missing values. However, uneven sampling can lead to inaccurate acrophase estimations.

2.1.5 RAIN

RAIN (Rhythmicity Analysis Incorporating Nonparametric methods) is an upgraded version of JTK_CYCLE using asymmetric waveforms [8]. RAIN examines the increasing and decreasing portions of the curve separately. It is more tolerant to data quality issues than JTK_CYCLE.

2.1.6 meta2d

Metacycle presents a platform that can be used to merge the results of different methods. It applies the function meta2d [9], which is based on Fisher's combined probability test to combine the results of ARSER, JTK_CYCLE and/or LS.

2.2 Benchmarking of methods for rhythmicity analysis

One of the problems of benchmarking of methods on real data is that the ground truth, e.g., does a specific dataset reflect oscillatory response or not, is usually not known. This problem can be solved with the application of synthetic data. Synthetic data presenting rhythmic as well as arrhythmic time-series can be created with samples from signals with user-defined parameters (e.g., periods and amplitudes), and with the addition of Gaussian noise to recreate the natural as well as technical variance [3]. However, there are certain guidelines that should be followed when collecting/generating data for experiments involving rhythmicity analysis [3]. For example, the data should be collected/generated for at least two periods to reduce the sensitivity of computational methods to outliers (number of false negatives).

2.3 A computational framework for domain-agnostic analysis of rhythmic data

We present a Python package RDA (Rhythmic Data Analysis) with the implementation of different functions that allow the user to perform rhythmic data analysis using LS, ARSER, JTK_CYCLE, Cosinor, RAIN, and meta2d. The package can accept the data in generalised input formats suitable for an arbitrary scientific domain, and can produce different types of visualisation of the obtained results, namely p-value distributions, Venn diagrams, and classification performance measures (when the target labels are known). The RDA implementation, documentation and examples are available at `https://github.com/VESTUArthur/RDA`.

3 Results

3.1 Case study 1: benchmarking of methods on synthetic data

We generated the synthetic data using the functionalities of the CosinorPy package [10] with different noise levels relative to the oscillation amplitudes (0.3, 0.6, 0.9) and different number of cosinor components (1, 2 or 3). The data were generated in a 48 hour interval with a 2 hour sampling resolution. We did not use replicates because the ARSER does not work with replicated data. For each configuration we generated 5,000 rhythmic and 5,000 non-rhythmic time-series data, and each configuration was repeated 6 times. The period of rhythmic data was set to 24 hours. We tested these datasets using different methods and we evaluated their performance through various metrics, such as Matthew's correlation coefficient (MCC) [11], area under curve (AUC), precision, recall, f1-score and accuracy (see Figure 1). Figure 2 presents the results of MCC assessment for all the experiments.

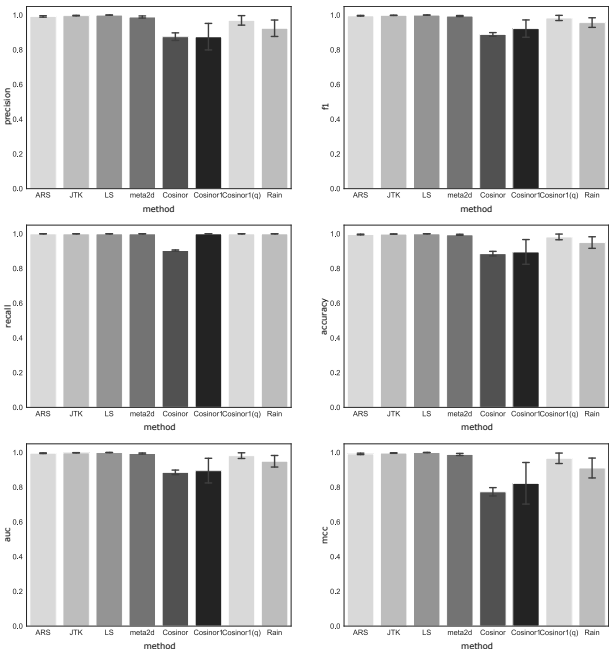

Figure 1: Evaluation of methods on synthetic data generated using a single-component cosinor model and a noise levels of 0.3. The experiment was repeated 6 times and error bars represent standard error of each metric for each model. Abbreviations and symbols: ARS – ARSER, JTK – JTK_CYLCE, LS – Lomb-Scargle, meta2d – Metacycle, Cosinor – zero-amplitude test using a multi-component cosinor, Cosinor1 – zero-amplitude using a single-component cosinor, Cosinor1(q) – model significance test using a single-component cosinor, RAIN – Rhythmicity Analysis Incorporating Nonparametric methods.

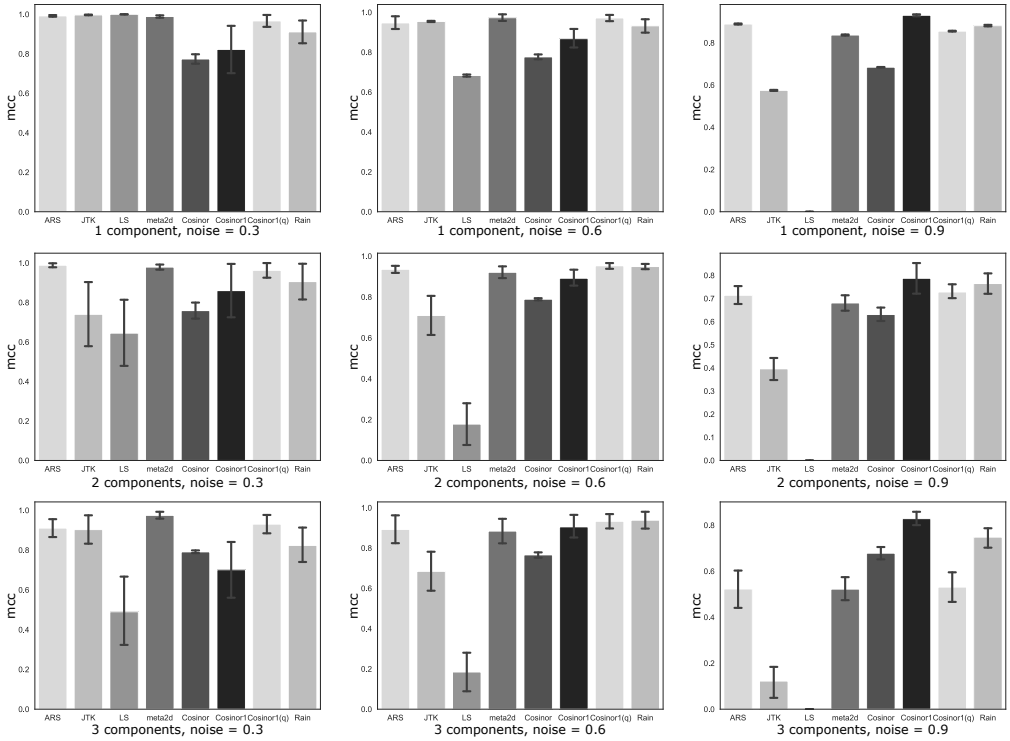

Figure 2: Matthew's Correlation Coefficient (MCC) for each method evaluated on data generated with different number of harmonic component and noise levels. Each experiment was repeated 6 times and error bars represent standard error of each metric for each model. Abbreviations and symbols: ARS – ARSER, JTK – JTK_CYLCE, LS – Lomb-Scargle, meta2d – Metacycle, Cosinor – zero-amplitude test using a multi-component cosinor, Cosinor1 – zero-amplitude using a single-component cosinor, Cosinor1(q) – model significance test using a single-component cosinor, RAIN – Rhythmicity Analysis Incorporating Nonparametric methods.

3.2 Case study 2: application of the framework on traffic data

In our second case study we applied the proposed framework to the evaluation of rhythmic trends in traffic data. We performed the analysis of data obtained on different road segments using Google Directions API [12] as described in [13]. Briefly, travel times on a segment were obtained using a 10 minutes sampling resolution. Route lengths and travel times were then converted to average paces for each route and for each time sampled. In the analysis we employed the same methods as in the first case study. We tested the capability of each method to assess the rhythmicity parameters that can be used for interpretation, namely locations of each peak and MESOR (midline estimating statistic of rhythm) values. Each method returns different parameters, which can be interpreted in a similar way (see Table 1).

We had problems using RAIN due to the large amount of data which caused the memory exceeded error. Moreover, to test the data on ARSER it was necessary to average the replicates on each timepoint, and fill missing values. For each method, we plotted the peaks obtained. Nevertheless, ARSER peak was not plotted since it does not yield the peak occurrence. Figure 3 indicates the consistence between the observed data and the assessed peaks.

We can see that Cosinor yields the most accurate locations of peaks. Moreover, a multi-component cosinor model is the only method, that is able to assess the locations of multiple peaks per one rhythmicity period. Cosinor also yields the most user-friendly output with exact locations of peaks and their heights.

Table 1: Interpretation of rhythmicity parameters as returned by the implementation of each method on the selected datset. Abbreviations and symbols: ARS – ARSER, JTK – JTK_CYLCE, LS – Lomb-Scargle, Cosinor – multi-component cosinor, Cosinor1 – single-component cosinor, N/A – not available, t(peak) – occurrence of peak(s), height(peak) – height of peak(s), MESOR – midline estimating statistic of rhythm.

Method	t(peak)	height(peak)	MESOR
ARS	N/A	amplitude	mean
JTK	LAG	AMP	mean
LS	PeakSPD	PhaseShiftHeight	N/A
Cosinor	peak	heights	mean
Cosinor1	acrophase[h]	amplitude	mean

4 Conclusions

In this paper we proposed a framework for domain-agnostic analysis of rhythmic data. We benchmarked the methods incorporated within the proposed framework using synthetic data. Moreover, we demonstrated the applicability of the framework on the traffic data reflecting circadian trends with two distinct peaks per period.

Each of the methods applied in our analysis exhib-

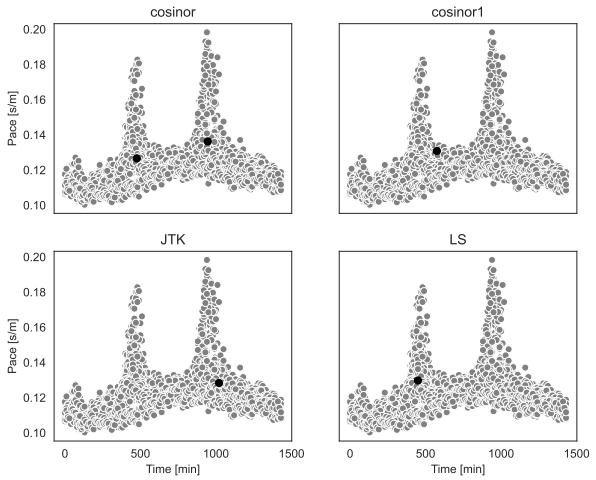

Figure 3: The consistence between the observed data and assessed peaks for each of the selected methods. Black dots represent the locations of assessed peaks. The analysis was performed on route 0 – a more detailed description of the data and the location of this route is available at [13].

ited certain advantages in comparison to other methods. For example, Cosinor can be used to accurately evaluate the rhythmicity parameters, but is in some cases less successful in classification than other (non-parametric) methods (see Figure 2). On the other hand, RAIN and JTK_CYCLE were specifically developed for hypothesis testing but yield inconsistent or biased estimations of rhythmicity parameters [8]. We also tested recently developed PyBOAT method [14], which can be used to accurately assess rhythmicity parameters, but failed on both of our case studies. The period obtained with this method was between 400 and 1000 minutes while the true period of the data was 1440 minutes (i.e., 24 hours) in both cases.

When performing the analyses one must identify its goals and select the most suitable method accordingly. Our recommendation is that different methods are used together to take advantages of each method separately and thus obtain the results with higher significance. The proposed framework allows the researcher to do this straightforwardly.

Our future work will be focused to more detailed benchmarking of the applied methods as well as to the application of the proposed framework on other types of rhythmic data. Moreover, our work has been mostly focused to the identification and characterisation of individual time-series data. In the near future we will extend the framework and its benchmarking with comparative analyses of rhythms from different fields of science. Additionally, we will extend the framework with implementations of additional state-of-the-art methods.

Acknowledgements

This work has been partially supported by the scientific research program P2-0359, and by the basic research projects J5-1798, both financed by the Slovenian Research Agency. AV and LJR were supported by the Erasmus+ exchange programme of the European Union. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data nor in writing the manuscript.

References

[1] R. Zhang, N. Lahens, H. Ballance, E. Hughes, and J. Hogenesch, "A circadian gene expression atlas in mammals: implications for biology and medicine.," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. 16219–16224, 2014.

[2] M. Wenwen, J. Zhiwen, C. Yang, C. Li, S. Aziz, and J. Yuchao, "Genome-wide circadian rhythm detection methods: systematic evaluations and practical guidelines.," *Briefings in Bioinformatics*, vol. 22, no. bbaa135, 2021.

[3] M. Hughes, K. Abruzzi, R. Allada, R. Anafi, A. Arpat, G. Asher, P. Baldi, C. de Bekker, D. Bell-Pedersen, J. Blau, S. Brown, M. Ceriani, Z. Chen, J. Chiu, J. Cox, A. Crowell, J. DeBruyne, D. Dijk, L. DiTacchio, and J. Hogenesch, "Guidelines for genome-scale analysis of biological rhythms," *Journal of biological rhythms*, pp. 380–393, 2017.

[4] J. T. VanderPlas, "Understanding the lomb–scargle periodogram," *The Astrophysical Journal Supplement Series*, vol. 236, no. 1, p. 16, 2018.

[5] G. Cornelissen, "Cosinor-based rhythmometry," *Theoretical Biology and Medical Modelling*, vol. 11, no. 1, pp. 1–24, 2014.

[6] R. Yang and Z. Su, "Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation," *Bioinformatics*, vol. 26, no. 12, pp. i168–i174, 2010.

[7] 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*, vol. 25, no. 5, pp. 372–380, 2010.

[8] P. F. Thaben and P. O. Westermark, "Detecting rhythms in time series with rain," *Journal of biological rhythms*, vol. 29, no. 6, pp. 391–400, 2014.

[9] G. Wu, R. C. Anafi, M. E. Hughes, K. Kornacker, and J. B. Hogenesch, "Metacycle: an integrated r package to evaluate periodicity in large scale data.," *Bioinformatics*, vol. 32(21), no. 3351–3353, 2016.

[10] M. Moškon, "CosinorPy: a Python package for cosinor-based rhythmometry," *BMC bioinformatics*, vol. 21, no. 1, pp. 1–12, 2020.

[11] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[12] Google Inc., "The Directions API overview." `https://developers.google.com/maps/documentation/directions/overview`, 2022.

[13] Špela Verovšek, M. Juvančič, S. Petrovčič, T. Zupančič, M. Svetina, M. Janež, Žiga Pušnik, N. Velikajne, and M. Moškon, "An integrative approach to neighbourhood sustainability assessments using publicly available traffic data," *Computers, Environment and Urban Systems*, vol. 95, p. 101805, 2022.

[14] C. Schmal, G. Mönke, and A. E. Granada, "Analysis of complex circadian time series data using wavelets," in *Circadian Regulation*, pp. 35–54, Springer, 2022.