



## Article

# Generative Models for Periodicity Detection in Noisy Signals

Ezekiel Barnett <sup>1</sup>, Olga Kaiser <sup>1</sup>, Jonathan Masci <sup>1</sup> , Ernst C. Wit <sup>2</sup> and Stephany Fulda <sup>3,\*</sup>

<sup>1</sup> NNAISENSE, 6900 Lugano, Switzerland; ezekebarnett@gmail.com (E.B.); olga.kaiser@nnaisense.com (O.K.); jonathan.masci@nnaisense.com (J.M.)

<sup>2</sup> Institute of Computing, Università della Svizzera Italiana, 6962 Lugano, Switzerland; wite@usi.ch

<sup>3</sup> Sleep Medicine Unit, Neurocenter of Southern Switzerland, EOC, 6900 Lugano, Switzerland

\* Correspondence: stephany.fulda@gmail.com

**Abstract:** We present the Gaussian Mixture Periodicity Detection Algorithm (GMPDA), a novel method for detecting periodicity in the binary time series of event onsets. The GMPDA addresses the periodicity detection problem by inferring parameters of a generative model. We introduce two models, the Clock Model and the Random Walk Model, which describe distinct periodic phenomena and provide a comprehensive generative framework. The GMPDA demonstrates robust performance in test cases involving single and multiple periodicities, as well as varying noise levels. Additionally, we evaluate the GMPDA on real-world data from recorded leg movements during sleep, where it successfully identifies expected periodicities despite high noise levels. The primary contributions of this paper include the development of two new models for generating periodic event behavior and the GMPDA, which exhibits high accuracy in detecting multiple periodicities even in noisy environments.

**Keywords:** periodicity; periodicity detection; algorithm; generative models; periodic leg movements during sleep



**Citation:** Barnett, E.; Kaiser, O.; Masci, J.; Wit, E.C.; Fulda, S. Generative Models for Periodicity Detection in Noisy Signals. *Clocks&Sleep* **2024**, *6*, 359–388. <https://doi.org/10.3390/clockssleep6030025>

Academic Editor: Arcady Putilov

Received: 29 May 2024

Revised: 6 July 2024

Accepted: 17 July 2024

Published: 23 July 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

From heartbeats to commutes, global climatic oscillations to Facebook log-ons, periodicity—the phenomenon where events occur at regular intervals—is ubiquitous. Detecting periodicity in time series, often referred to as the “periodicity detection problem”, is crucial across various fields. In binary time series, which indicate only the occurrence of events, this problem has traditionally been tackled using algorithms such as Fast Fourier Transform (FFT) and autocorrelation, typically focusing on single, stationary periodicities [1–5]. However, several challenges remain inadequately addressed in the literature, including (i) the development of generative models that accurately describe noise in periodic behavior, considering variance in interval length and false positives and negatives; (ii) the detection of multiple overlapping periods; and (iii) non-stationary periodic signals.

Existing periodicity detection algorithms, primarily based on FFT or the Autocorrelation Function (ACF), often focus on single-period detection. FFT maps a time series to the frequency domain, using the inverse of the frequency with the strongest power as the predicted period. However, FFT is sensitive to sparse data [6], noise [7], and suffers from “spectral leakage” at low frequencies/large periods [8]. The Lomb–Scargle periodogram [7,9], a least-squares method for fitting sinusoids, also shares these problems and may not be suitable for non-stationary signals. Furthermore, in real applications, the hierarchy implied by the FFT may not be appropriate to describe the signal, especially when the periodic signals are random walks with Markov properties and the signal is non-stationary.

ACF-based methods estimate similarity between sub-sequences of event intervals, selecting periods that maximize the ACF. These methods have been used for multiple periodicity detection in character series such as texts [10,11]. However, ACF detects numerous candidate periods, often requiring a self-selected significance threshold to identify

true periodicities, and struggles with smaller data sets. These methods are generally not designed for multiple periodicities in event time series.

Alternative approaches include E-periodicity [12], which focuses on single period detection using the modulus operation for unevenly/under-sampled time series. E-periodicity segments the time series into all possible periodicities within some a priori specified range. It then overlays the segments and selects the true periodicity as the periodicity that “covers” the most events. Methods like partial periodic patterns, chi-squared tests [13], max sub-pattern trees [13], and projection-based techniques [14] also target single periodic patterns in stationary signals and face challenges with low-frequency periodicities and low sampling rates [15].

Multiple period detection in time series event data has received limited attention. Most methods use a hierarchical extraction approach, iteratively removing the frequency with the highest power (FFT) or the most probable periodicity (ACF). While FFT is suitable for complex functions, it is not designed for event data [16]. ACF-based methods, combined with comb filters, identify periods but lack robustness against noise [17].

Generative models, like those in [18], describe discrete signals with Gaussian PDFs for periods and Poisson processes for noise but fail to detect multiple overlapping periods. Other approaches in “periodic pattern-finding” identify time slots for periodic events, optimizing for anomaly detection tasks with low sampling rates [15].

To address some of the challenges of multiple periodicity detection for noisy event time series, we propose the Gaussian Mixture Periodicity Detection Algorithm (GMPDA). This algorithm is based on a novel generative model scheme for periodic event time series, incorporating Gaussian-distributed noise to account for interval variability. We compare the GMPDA to existing algorithms across extensive test cases, demonstrating superior performance in accuracy, sensitivity, and computational efficiency.

The rest of the paper is organized as follows. In Section 2, we introduce the generative models and discuss their inference. Section 3 presents the GMPDA. The performance of the GMPDA framework is tested in Section 4. An application of the GMPDA to real data is detailed in Section 5. Section 6 concludes this paper.

## 2. Generative Models

Consider a uni-variate event time series  $(X_t)_{t=1,\dots,N_T}$ , where  $x_t = 1$  if there is an occurrence of an event at time  $t$ , and  $x_t = 0$  otherwise. In this work, we disregard cases of unsampled or missing data. The information in  $X_t$  can be condensed into the set of non-zero time stamps  $S := \{s_i | x_{s_i} = 1\}_{i=1,\dots,N_S}$ .

If the positive time stamps occur at regular intervals, the time series exhibits periodic behavior, and these intervals correspond to periodicities or periods. We frame the periodicity detection problem as the search for the set of periodicities that explain the intervals between time stamps in  $S$ .

We are particularly interested in the set of *prime* periodicities, defined as the smallest integer frequency that describes the intervals. For instance, for a time stamp set  $S = \{12, 23, 34, 45, 56, 67\}$ , several intervals could be explained by a periodicity of 22 or 33 but 11 would be the prime period, as it explains the data best, and 22 and 33 are integer multiples of this prime period. The set of underlying prime periods in  $X_t$  is denoted by  $\mu^* = \{\mu_p^*, p = 1, \dots, P\}$ . Additionally, we assume that for most real-world applications, the interval between two consecutive time stamps associated with a periodicity  $\mu_p^*$  in  $S$  will typically vary around  $\mu_p^*$  with a variance denoted by  $\sigma_p^{*2}$ , i.e.,  $s_{i+1}^p - s_i^p \in [\mu_p^* - \sigma_p^*, \mu_p^* + \sigma_p^*]$ .

If the time series  $X_t$  is generated by a single, stationary periodicity  $\mu_1^*$ , we can compute  $\mu_1^*$ , and thus the prime periodicity  $\mu^*$ , directly from the data as:

$$\mu^* = \frac{N_T}{|S|} = \frac{\sum_{i=1}^{N_T} s_{i+1} - s_i}{|S|}. \quad (1)$$

The first equality in (1) represents the ratio between the length of the time series and total number of events. The second equality in (1) describes the “average interval” between two

adjacent time stamps and holds for a time series with a single, stationary periodicity without noise. The associated variance  $\sigma^{*2}$  can be estimated as the square of the standard deviation.

However, estimation of  $\mu^*$  and  $\sigma^{*2}$  using Equation (1) is insufficient when (i) the time series  $X_t$  is generated by multiple, overlapping periodicities, (ii) the time series  $X_t$  is noisy (i.e., contains false positives), (iii) there are missing values (false negatives), or (iv) there are varying patterns of periodic behavior over time (non-stationarity).

Our generative model addresses challenges (i) and (ii). Specifically, we formulate a generative model of the positive time stamps  $S$  with multiple periodicities, incorporating an explicit noise term and a loss function that enables inference of the model parameters.

Assume the set of positive time stamps  $S$  can be generated by a function  $f$  as:

$$S = f(\mu^*, \sigma^*, \beta, \alpha, M), \quad (2)$$

where we have the following:

- $\mu^*$  is the set of  $P$  prime periodicities in the time series;
- $\sigma^*$  is the set  $P$  variances of the periodic intervals;
- $\beta$  is the rate of false positive events, i.e., noise;
- $\alpha_p$  is the starting point of periodicity  $p$ ;
- $M$  is the generative model scheme.

The generative model scheme  $M$  is characterized by the priors for the distribution of the intervals, their mean values  $\mu^*$ , and their variances  $\sigma^{*2}$ . Following the generative approach in Equation (2), we assume that each event  $s_i$  is generated according to one periodicity (except in the case of overlaps) or false positive noise  $\beta$ . Thus, the set  $S$  is the union of subsets  $S^p$  of positive time stamps  $s_i$  associated with periodicity  $\mu_p$  or random noise  $\beta$ :

$$S = S^{\mu_1^*} \cap S^{\mu_2^*} \dots \cap S^{\mu_P^*} \cap S^\beta. \quad (3)$$

Without loss of generality, we parameterize the distribution of the intervals using the Gaussian distribution; any other distribution, such as those of the exponential family, would also be appropriate. In Sections 2.1 and 2.2, we formulate two different model schemes: the *Clock Model* ( $M = C$ ) and the *Random Walk Model* ( $M = RW$ ).

### 2.1. Clock Model

The “*Clock Model*” describes periodic behavior governed by a fixed period  $\mu_p^*$  with Gaussian noise, which does not rely on information from previous positive time stamps to compute the occurrence of the next event. For  $p = 1, \dots, P$  and  $i = 1, \dots, N_T$ , the events in  $S^p$  are generated by:

$$s_i^p = \alpha_p + (i \cdot \mu_p^*) + \epsilon, \epsilon \sim N(0, \sigma_p^{*2}). \quad (4)$$

The number of events associated with uniformly distributed false positive noise is given by  $\beta * |S^{\mu^*}|$  in the interval  $[0, N_T]$ .

In the *Clock Model*, the location of any event depends solely on its position in the time series and Gaussian noise around some regular interval but not on the previous time steps. Consequently, one can predict any future time step  $s_{i+m}^p$  for  $m > 0$  with equal accuracy. This formulation generalizes the generative models found in much of the previous work on periodicity detection, such as in [15] and [12]. In their models, the objective is to identify a time slot  $si$  as a pair of a period ( $l$ ) and an offset  $i$ , denoted by  $[l : i]$ . This approach is equivalent to finding a period  $\mu_p^*$  and a starting point  $\alpha_p$ , with  $\sigma^* = 0$ . However, this might be limiting in real applications, as it does not account for variability in event locations within the time series. To address this, we introduce Gaussian noise  $\sigma^*$ , with the case  $\sigma^* = 0$  being a special instance of the *Clock Model*.

However, the notion of a pacemaker, a component that imposes regular timing signals to synchronize events, is realistic only for certain systems, thereby motivating the development of the *Random Walk Model*.

## 2.2. Random Walk Model

The Random Walk Model exhibits the Markov property, meaning the temporal location of the next event depends on the current event's temporal location and Gaussian noise. For  $p = 1, \dots, P$  the events in  $S^p$  are generated as follows:

$$s_{i+1}^p = s_i^p + \mu_p^* + \epsilon, \epsilon \sim N(0, (i\sigma_p)^{*2}). \quad (5)$$

The number of events associated with uniformly distributed false positive noise is given as  $\beta * |S^{\mu^*}|$  in the interval  $[0, N_T]$ .

As the noise is Gaussian (and thus is identically distributed), the series of event time stamps in  $S^{\mu_p}$ , for  $p = 1, \dots, P$ , describes a random walk. Therefore, the formulation in Equation (5) is referred to as the Random Walk Model (RWM).

The RWM has the characteristic that the variances  $\sigma^{*}$ 's accumulate with each subsequent time step. Consequently, the variance of the expected location of an event increases linearly with the distance from the current event. This assumption is essential and realistic for many real-life systems lacking a pacemaker, where predictability decreases with the number of steps. For example, given  $s_i$ , we can predict  $s_{i+1}$  more accurately than  $s_{i+10}$ .

## 2.3. Inference

Given an event time series  $X_t$ , a straightforward approach to extract possible periodicities is to study the empirical histogram of all pairwise, forward-order inter-event intervals. For each event, we consider not only the interval to the next event (onset to onset) but also to all subsequent events.

The possible range of the intervals is defined by  $(0, N_T)$ . Please note, that in real applications, the actual range is smaller, as an interval needs to be observed a minimal number of times to be significant. We can estimate a histogram of expected forward-order inter-event intervals based on the generative models defined by Equations (4) and (5). This histogram is obtained by (i) analytically estimating the expected number of intervals for each  $\mu_p^* \in \mu^*$ , (ii) incorporating all intervals between any pair of events associated with different prime periodicities  $\mu_p^*$  and  $\mu_q^*$ , and (iii) incorporating intervals due to noise (which can be performed analytically if the noise source is known; otherwise, an estimate is required). Comparing the empirical histogram with the parametric expectation defines the loss function used to identify the optimal underlying periodicities.

For every  $\mu \in (0, N_T)$ , we define the function of interval counts  $D(\mu)$  as:

$$D(\mu) = \sum_{i,m>0} \mathbb{1}_{s_{i+m}^p - s_i^p = \mu}. \quad (6)$$

Evaluating  $D(\mu)$  for a given  $X_t$  results in a histogram of all pairwise inter-event intervals.

The generative models provide a statistical model for the intervals. Therefore, we can estimate the expected number of intervals for  $\mu \in (0, N_T)$  in reference to a fixed periodicity  $\mu_p^*$  and variance  $\sigma_p^*$  as:

$$\mathbb{E}[D(\mu)]_{\mu_p^*} = \sum_{i,m>0} \mathbb{E}[\mathbb{1}_\mu] \quad (7)$$

$$= \sum_{i,m>0} \mathbb{P}[s_{i+m}^p - s_i^p = \mu], \quad (8)$$

where equality in Equation (7) is due to linearity of expectation and that in Equation (8) is due to the fact that for a random variable  $A$ ,  $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}[A]$ . The distribution of all  $m$ -th order inter-event intervals depends on the specific generative model and can be written as

$$\mathbb{P}[s_{i+m}^p - s_i^p = \mu] = \frac{1}{\sqrt{2\pi\sigma_p^{*2}}} \exp\left[-\frac{(\mu - m\mu_p^*)^2}{2\sigma_p^{*2}}\right] \quad (9)$$



for the Clock Model, and as

$$\mathbb{P}[s_{i+m}^p - s_i^p = \mu] = \frac{1}{\sqrt{2\pi(m\sigma_p^*)^2}} \exp\left[-\frac{(\mu - m\mu_p^*)^2}{2(m\sigma_p^*)^2}\right] \quad (10)$$

for the Random Walk Model. For the latter, the variance grows linearly with the number of steps between events.

Further, we assume the starting point is zero, i.e.,  $\alpha_p = 0$ . For a time series of length  $N_T$ , Equation (8) can be rewritten in a more explicit form by expressing the indicator function as a definite quantity. Assuming no missing values in  $(0, N_T)$ , we should observe  $\frac{N_T}{\mu_p^*}$  first-order intervals ( $m = 1$ ) in the time series, distributed according to the Gaussian probability density function (PDF) parametrized by  $\sigma^*$  and  $\mu_p^*$ . For the second-order intervals ( $m = 2$ ), the scaling factor would be  $(\frac{N_T}{\mu_p^*} - 1)$ , for  $m = 3$ ,  $(\frac{N_T}{\mu_p^*} - 2)$ , and so on. Thus, for a single periodicity  $\mu_p^*$ , the expected value of  $D(\mu)$  can be written for the Clock Model as:

$$\mathbb{E}[D(\mu)]_{\mu_p^*} = \sum_{m=1}^{\frac{N_T}{\mu_p^*}} \frac{\text{const}}{\sqrt{2\pi\sigma_p^{*2}}} \exp\left[-\frac{(\mu - m\mu_p^*)^2}{2\sigma_p^{*2}}\right] \quad (11)$$

and for the Random Walk Model as

$$\mathbb{E}[D(\mu)]_{\mu_p^*} = \sum_{m=1}^{\frac{N_T}{\mu_p^*}} \frac{\text{const}}{\sqrt{2\pi(m\sigma_p^*)^2}} \exp\left[-\frac{(\mu - m\mu_p^*)^2}{2(m\sigma_p^*)^2}\right] \quad (12)$$

with  $\text{const} = \frac{N_T}{\mu_p^*} - (m - 1)$ . Equations (12) and (11) therefore provide the expected values of the function  $D(\mu)$  for counting all order intervals that might be observed for a single periodicity  $\mu_p^*$  for the Random Walk and Clock Models, respectively.

In the case of multiple, overlapping periods, and/or false positive noise, the set of positive time stamps  $S$  consists of multiple sets:  $S^{\mu_1^*} \cap S^{\mu_2^*} \dots \cap S^{\mu_p^*} \cap S^\beta$ . Since the affiliation of events to periodicities is unknown, we adapt our definition of  $D(\mu)$  in Equation (6) by removing the superscript  $p$ :

$$D(\mu) = \sum_{\forall i, m > 0} \mathbb{1}_{s_{i+m} - s_i = \mu}. \quad (13)$$

This modified operator  $D(\mu)$  now counts the intervals not only between events from the same periodicity set but also between events in different sets and/or between noise events. We call the latter two “interaction intervals” and denote their contribution to  $D(\mu)$  by:

$$\zeta(\mu) = \sum_{\forall i, m > 0} \mathbb{1}_{s_{i+m} - s_i = \mu}. \quad (14)$$

This includes three possible scenarios (or their combination): (i) intervals between events from different periodicity sets, i.e.,  $s_i \in S^{\mu_p^*}$  and  $s_{i+m} \in S^{\mu_q^*}$ ; (ii) intervals between events from any periodicity set and noise, i.e.,  $s_i \in S^{\mu_p^*}$  and  $s_{i+m} \in S^\beta$ ; and (iii) intervals between noise events, i.e.,  $s_i \in S^\beta$  and  $s_{i+m} \in S^\beta$ .

The estimates in Equations (11) and (12) do not include these interaction intervals. Next, we discuss how to estimate  $\zeta(\mu)$  and to account for the three cases explicitly. The distribution of the interaction intervals for all the three cases can be obtained in closed form by applying the convolution formula, which provides the distribution of the sum/difference of two interdependent discrete or continuous random variables [19]. For the Clock and the Random Walk Models, we obtain the following: For case (i), the interaction intervals between two periods, denoted as  $\zeta_{pq}(\mu)$ , are Gaussian distributed with a mean  $\mu_{pq} = \mu_p - \mu_q$  and a variance  $\sigma_{pq}^2 = \sigma_p^2 + \sigma_q^2$ , where  $\mu_p > \mu_q$  without loss of generality. For case (ii), the interaction intervals, denoted as  $\zeta_{p\beta}(\mu)$ , follow a Gaussian-like distribution, adjusted for the corresponding uniform support. For case (iii), the forward interaction

intervals, denoted as  $\zeta_\beta(\mu)$ , follow the right sight of a triangle distribution. In this context, we can write  $\zeta(\mu)$  as

$$\zeta(\mu) = \zeta_{pq}(\mu) + \zeta_{p\beta}(\mu) + \zeta_\beta(\mu). \quad (15)$$

In real applications, the amount of noise in the data is unknown, and the exact formulation of (15) is not available, necessitating an approximation of  $\zeta_\beta(\mu)$ . We assume that the events contributing to the interaction intervals are uniformly distributed, an assumption that is not overly restrictive as demonstrated in Appendix A.1. Please note that we formulate the task of finding the optimal estimates for the periodicities by minimizing the distance (loss) between the observed interval differences  $D(\mu)$  and the expected interval differences  $\mathbb{E}[D(\mu)]$  with respect to the chosen generative model. To obtain the expected differences, which include both interaction intervals and noise (whether for a Clock or Random Walk Model), we must explicitly approximate the interaction intervals caused by noise, as the amount of noise in real-world applications is unknown. In the next proposition, we provide an approximation, and hereafter, we derive a loss function that enables the estimation of multiple periodicities.

**Proposition 1.** For uniformly distributed events on  $[1, N_T]$ , the expected number of interaction intervals on the interval  $[1, N_T]$  is given by:

$$\mathbb{E}[\zeta(\mu)] = z \cdot (1 - \frac{\mu}{N_T}) \quad (16)$$

with a constant  $z$ , for every  $\mu \in [1, N_T]$ .

**Proof.** Consider two noise events  $s_i^\beta, s_j^\beta$  each with a uniform probability mass function  $\mathbb{P}_S$  on the support  $[1, \dots, N_T]$ . The difference between the events  $s_j^\beta - s_i^\beta = \mu \in [-N_T, N_T]$  is a random variable whose probability mass function can be derived using the convolution formula for distributions [19]:

$$\mathbb{P}[s_j^\beta - s_i^\beta = \mu] = \sum_{s_j} \mathbb{P}_S[s_j - \mu] \mathbb{P}_S[s_j]. \quad (17)$$

Given that  $\mathbb{P}_S$  is defined on  $[1, \dots, N_T]$ , the probability of  $\mathbb{P}_S[s_j - \mu < 1]$  and  $\mathbb{P}_S[s_j - \mu > N_T]$  is zero. Therefore, we obtain:

$$\sum_{s_j} \mathbb{P}_S[s_j - \mu] \mathbb{P}_S[s_j] = (N_T - \mu) \frac{1}{N_T} \frac{1}{N_T} \quad (18)$$

Finally, the probability mass function is a decaying function of the difference:

$$\mathbb{P}[s_j^\beta - s_i^\beta = \mu] = \frac{1}{N_T} (1 - \frac{\mu}{N_T}). \quad (19)$$

If focusing on  $|s_j^\beta - s_i^\beta| = \mu$ , the right-hand side of Equation (19) must be multiplied by 2 due to symmetry. Next, we estimate the expectation of  $\zeta(\mu)$ :

$$\mathbb{E}[\zeta(\mu)] = \mathbb{E}[\sum_{\forall i, m > 0} \mathbb{1}_{s_{i+m} - s_i = \mu}] \quad (20)$$

$$= \sum_{\forall i, m > 0} \mathbb{E}[\mathbb{1}_{s_{i+m} - s_i = \mu}] \quad (21)$$

$$= \sum_{\forall i, m > 0} \mathbb{P}[s_j^\beta - s_i^\beta = \mu]. \quad (22)$$

The equality in Equation (21) is due to linearity of expectation, and that in Equation (22) is due to the fact that for a random variable  $A$ , the following equality holds:  $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}[A]$ . Inserting Equation (19) into Equation (22) results in:

$$\mathbb{E}[\zeta(\mu)] = \sum_{\forall i, m > 0} \frac{1}{N_T} (1 - \frac{\mu}{N_T}). \quad (23)$$

The number of all pairwise, forward-order differences for the noise events, with  $n = N_T\beta = |S^\beta|$ , is given as:

$$\sum_i (n - i) = n^2 - \frac{(n^2 + n)}{2}, \quad (24)$$

thus, we obtain:

$$\mathbb{E}[\zeta(\mu)] = \frac{2n^2 - (n^2 + n)}{2N_T} \left(1 - \frac{\mu}{N_T}\right). \quad (25)$$

By setting  $z = \frac{2n^2 - (n^2 + n)}{2N_T}$ , we obtain Equation (16).  $\square$

In real applications, the constant  $z$  cannot be estimated as the amount of false positives in the unknown a priori. An approximation  $\hat{z}$  will be inferred from the data in Appendix A.1. For now, the expected number of interaction intervals is approximated via:

$$\mathbb{E}[\hat{\zeta}(\mu)] = \hat{z} \cdot \left(1 - \frac{\mu}{N_T}\right). \quad (26)$$

In the case of multiple periodicities, due to linearity of expectation, the expected number of intervals over multiple periods is the sum of the expectations for  $D(\mu)$  for each periodicity  $\mu_p^*$  present in the data, plus the expected number of interaction intervals approximated by (14):

$$\mathbb{E}[D(\mu)] = \sum_{p=1}^P \mathbb{E}[D(\mu)]_{\mu_p^*} + \mathbb{E}[\zeta(\mu)]. \quad (27)$$

Hereinafter, the first addend on the right-hand side is denoted as the deterministic parametric function  $G_M(\mu; \hat{\mu}, \hat{\sigma})$  for the Clock Model:

$$G_C(\mu; \hat{\mu}_p, \hat{\sigma}_p) = \sum_{p=1}^P \sum_{m=1}^{\frac{N_T}{\hat{\mu}_p}} \frac{c_p}{\sqrt{2\pi\hat{\sigma}_p^2}} \exp\left[-\frac{(\mu - m\hat{\mu}_p)^2}{2\hat{\sigma}_p^2}\right], \quad (28)$$

and the Random Walk Model:

$$G_{RW}(\mu; \hat{\mu}_p, \hat{\sigma}_p) = \sum_{p=1}^P \sum_{m=1}^{\frac{N_T}{\hat{\mu}_p}} \frac{c_p}{\sqrt{2\pi(m\hat{\sigma}_p)^2}} \exp\left[-\frac{(\mu - m\hat{\mu}_p)^2}{2(m\hat{\sigma}_p)^2}\right] \quad (29)$$

with  $c_p = (\frac{N_T}{\hat{\mu}_p} - (m - 1))$ .

Once we obtain estimates  $\hat{\mu}_p$  and  $\hat{\sigma}_p$  for the true periodicities  $\mu_p^*$  and variances  $\sigma_p^*$ , and given a prior on the generating function (either Random Walk or Clock), we can write a loss function for our estimates as the difference between the empirical  $D(\mu)$  and the parametric  $G_M(\mu; \hat{\mu}, \hat{\sigma})$ . The loss function can be either the absolute error or a quadratic loss; since we have deterministic expectations, we focus on the absolute error as follows:

$$\mathcal{L} = \sum_{\mu=1}^{N_T} |D(\mu) - \mathbb{E}[D(\mu)]|, \quad (30)$$

$$= \sum_{\mu=1}^{N_T} |D(\mu) - ([\sum_{p=1}^P \mathbb{E}[D(\mu)]_{\mu_p^*}] + \mathbb{E}[\zeta(\mu)]|, \quad (31)$$

$$\approx \sum_{\mu=1}^{N_T} |D(\mu) - G_M(\mu; \hat{\mu}, \hat{\sigma}) - \mathbb{E}[\hat{\zeta}(\mu)]|. \quad (32)$$

Finally, for either the Clock Model or the Random Walk Model, the aim is to find a set of periodicities and variances that minimize the corresponding loss. A straightforward approach would consider all possible combinations of acceptable periodicities and variances, where the optimal combination minimizes the loss.

However, such an approach is computationally infeasible. Therefore, the following section outlines the Gaussian Mixture Periodicity Detection Algorithm (GMPDA).

### 3. GMPDA

Given an event time series  $X_t \in \mathbb{R}^{(1 \times N_T)}$  the aims are (i) to extract an estimate  $\hat{\mu}$  of the true generating periodicities  $\mu^*$ , (ii) to infer  $\sigma^*$ , and (iii) to test the fit of the chosen generative model  $M$ . The GMPDA provides a method to learn the parameters of the generative function of  $X_t$  accurately and efficiently by minimizing the loss  $\mathcal{L}$  defined in Equation (32). The GMPDA is open-source and available on <https://github.com/nnaisense/gmpda> (accessed on 18 July 2024). The GMPDA is based on comparing  $D(\mu)$ , the empirical distribution of the intervals observed in the time series  $X_t$ , with parametrized estimates of its generative function  $G_M(\hat{\mu}, \hat{\sigma})$ , plus the contribution coming from the interaction intervals, using the loss function (32). The main steps of the GMPDA for estimating the optimal parameters  $\hat{\mu}, \hat{\sigma}$  are outlined in Algorithm 1.

---

#### Algorithm 1: Main Steps of the GMPDA.

---

- 1 Extract event time stamps:  $S \leftarrow$  where  $X_t = 1$
  - 2 Compute intervals  $D(\mu)$  from  $S$ , with respect to Equation (A2) and subtract  $\zeta(\mu)$ , estimated with respect to Equation (A1)
  - 3 Identify candidates periods using integral convolution
  - 4 Initialize and optimize variance for candidates periods
  - 5 Find optimal combination of periodicities, which minimize the loss defined in (32)
  - 6 Update loss and sigma with respect to optimal periodicities
- 

After extracting the event time stamps, the GMPDA computes  $D(\mu)$  with respect to Equation (A2) and subtracts the approximated contribution from the interaction events (step 2 of Algorithm 1). The approximation of the length of interaction intervals is either limited by the minimal expected periodicity or by a user-defined parameter, denoted as *noise\_range*. The estimation of the approximation is outlined in Appendix A.1. For the estimation of  $D(\mu)$  and the loss, the range for  $\mu$  is limited by the parameter *loss\_length*, mainly due to the flattening of the Gaussian distribution with increasing variance for the Random Walk Model. A detailed discussion can be found in Appendix A.2.

In the *third step*, the GMPDA estimates a set of candidate periodicities using a heuristic approach since computing  $G_M(\hat{\mu}_p, \hat{\sigma})$  for all possible  $\hat{\mu}_p$  is computationally expensive. The heuristic approach iteratively searches for periodicities  $\hat{\mu}$  by performing “integral convolutions” on  $D(\mu)$  in each iteration. The convolution smooths the function for extracting periods that explain the time series. The maximum number of candidates is controlled by parameter *max\_candidates*, and the maximum number of iterations by the parameter *max\_iterations*. This heuristic approach is described in detail in Appendix A.2.

In the *fourth step*, the GMPDA performs least-squares curve fitting to refine the initial guess for  $\hat{\sigma}$ . This step is optional and can be controlled by the parameter *curve\_fit*. The curve fitting procedure is described in Appendix A.3.

In the *fifth step*, the GMPDA computes the function  $G_M(\hat{\mu}_p, \hat{\sigma})$  for all combinations of candidate periodicities and corresponding variances. It then selects the set of “prime periodicities”  $\hat{\mu}^*$  that minimizes the loss, defined and explained in Appendix A.4. In the final, sixth step, the loss and  $\hat{\sigma}$  are updated with respect to optimal periodicities.

### 4. Performance Evaluation on Test Cases

This section evaluates the capacity of the GMPDA to detect periodicities  $\mu^*$  and variances  $\sigma^*$  on synthetic time series generated according to Clock and Random Walk Models. The performance of the GMPDA in detecting periodicities  $\mu^*$  was compared with other periodicity detection algorithms, including FFT, autocorrelation with FFT, histogram with FFT, and E-periodicity (the alternative algorithms were implemented in MATLAB). For all

algorithms, the minimal and maximal considered period lengths were set to 10 and 350, respectively. The corresponding code is available on <https://github.com/nnaisense/gmpda> (accessed on 18 July 2024)). These specific algorithms are described below.

**GMPDA:** We used the baseline algorithm described in Algorithm 1 with  $\hat{\sigma}$  set equal to  $\sigma^*$  (i.e.,  $\sigma^* = \log(\mu)$ ) and no non-linear curve fitting.

**GMPDA  $\sigma^*$  unknown:** The algorithm is initialized with  $\hat{\sigma} = \text{int}(\log(10))$ , which is the minimal possible value for  $\mu^*$ . For the above GMPDA configurations, the algorithm searches for maximal  $|\mu^*| + 2$  periodicities. Please note that we have chosen here to use a different sigma for the application of the GMPDA with curve fitting (i.e.,  $\hat{\sigma} = \sigma^*$ ) compared to the GMPDA without curve fitting ( $\hat{\sigma} = \text{int}(\log(10))$ ). In real applications, sigma is unknown and would be supplied as the best guess available. The GMPDA with curve fitting tries to optimize the initial estimate of sigma once the candidate periodicities are identified. If no non-linear curve fitting is deployed, we suggest running the algorithm multiple times for a range of possible  $\hat{\sigma}$  values, and the optimal  $\hat{\sigma}$  can then be chosen with respect to the lowest loss.

**FFT:** This is a Power Spectral Density Estimates approach [1]. In the case of a single periodicity, the frequency with the highest spectral power is selected as the prime periodicity. In the case of multiple periodicities, the frequencies  $|\mu^*|$  with the highest spectral power are selected as the true periodicities.

**Autocorrelation with FFT:** The Autocorrelation Function (ACF) estimates how similar a sequence is to its previous sequence for different lags and then uses the lag that maximizes ACF as the predicted period [5]. Since all integer multiples of true periods will have the same function value, an FFT is applied to the ACF to select the frequencies with the highest spectral power as the true periodicities. In the case of multiple periodicities, the frequencies with the highest spectral power are selected as the true periodicities.

**Histogram with FFT:** This is an FFT applied to the histogram of all forward differences in the time series  $D(\mu)$  [20]. In the case of multiple periodicities, the frequencies with the highest spectral power are selected as the true periodicities.

**E-Periodicity:** We implement the method presented in [12], which computes a “discrepancy score” for each possible periodicity, i.e., the number of intervals between events that are equal to the candidate periodicity. To detect multiple periods, we select the top  $|\mu^*|$  candidate periods from the discrepancy function.

There are several conceptual differences and similarities between the GMPDA and the alternative algorithms: the GMPDA, like all the methods listed above, computes frequencies/periodicities based on the observed intervals between positive observations in the time series. For data that follow a Clock Model, variance in intervals can be handled using regression frameworks for ACF and spectral methods for FFT. However, for very small or large variations in intervals, parametrized by  $\sigma^*$  in the Random Walk Model, these methods may struggle—particularly for multiple periodicities—due to the linear increase in variance. E-periodicity and histogram methods are likely to show decreased performance for variable intervals, as they lack specific mechanisms to handle interval variance, which is particularly problematic for time series following the Random Walk Model.

The GMPDA is designed for multiple periodicity detection, and its loss function explicitly targets finding all periodicities present in the data. Once the set of candidate periodicities is identified, the GMPDA checks all possible combinations of periodicities and selects the one with the smallest loss.

ACF and FFT are accepted methods for hierarchical frequency detection, but they lack a “stopping criteria” to determine the number of significant periodicities in the time series. This can lead to the over- or underestimation of the number of periodicities. In our test cases, we always selected the top  $|\mu^*|$  frequencies as the true periodicities, likely overestimating the accuracy of these methods since there is no way to know this number without prior knowledge of the generative mechanism. Additionally, the E-periodicity and histogram methods are not explicitly designed for multiple periodicity detection and lack mechanisms for handling noise in intervals.



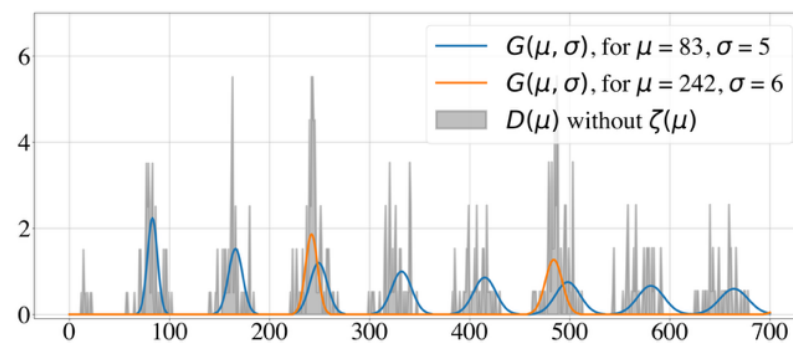
Conceptually, the GMPDA differs from classical Gaussian Mixture and Hidden Markov (HMM) approaches. All three methods aim to fit the shape of a distribution described by the corresponding histogram of the data. However, the GMPDA generative models account for peaks in the histogram at prime periods and their integer multiples, combining these peaks for a better estimate. Classical MM/HMM models do not use this information; instead, they try to fit all peaks individually if the number of mixture models  $K$  is large, or average them if  $K$  is small, leading to biased results.

In the following, we compare the performance of the above-described algorithms on a large set of generated test cases.

#### 4.1. Test Cases

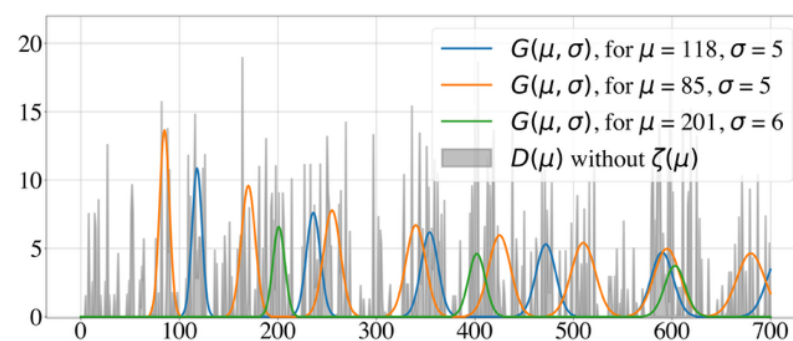
The performance of the GMPDA was evaluated on a wide range of test cases for the Clock Model and the Random Walk Model. The test cases systematically varied the following model parameters: periodicity  $\mu$ , variance  $\sigma$ , noise  $\beta$ , and the number of events  $n$ . These generative model parameters influence the histogram of inter-event intervals, which is the input data for all applied algorithms.

To better understand how these model parameters influence the histogram, we show two illustrative test cases with different model parameters. Figure 1 shows a well-posed test case, with two underlying periodicities and no noise, while Figure 2 displays an ill-posed test case, where the signal-to-noise ratio is 1:2. Identifying underlying periodicities in the latter case requires advanced analysis of the histogram.



**Figure 1.** Example of a well-posed test case with two underlying periodicities and no noise. Histogram of the intervals  $D(\mu) - \hat{z}$  and generative curves  $G(\mu, \sigma)$  for the Random Walk Model with  $n = 30$  and  $\beta = 0$ .

The following analyses examined an extensive range of test cases to study the limitations of the presented GMPDA and the alternatives described in Section 4.



**Figure 2.** Example of an ill-posed test case with a signal-to-noise ratio of 1:2. Histogram of the intervals  $D(\mu) - \hat{z}$  and generative curves  $G(\mu, \sigma)$  for the Random Walk Model with  $n = 100$  and  $\beta = 2$ .

## Configurations

For the configuration of the test cases, we considered the following values for the model parameters:

- $\sigma^* \in \{1, \log(\mu), \frac{\mu}{p}\}$ , with  $p = 16, 8, 4, 3$ ,
- $n \in \{10, 30, 50, 100, 300, 500\}$ ,
- $\beta \in \{0, 0.1, 0.5, 0.7, 1, 2, 4, 8\}$ .

Please note that test cases with  $\sigma = \log(\mu)$  represent scenarios where no  $\sigma$  optimization is required, as  $\sigma = \log(\mu)$  is the default initialization in the GMPDA. Small values for  $n$  and large values for  $\sigma, \beta$  were chosen to investigate the limits of the periodicity detection algorithms.

For every combination of  $\sigma^*$ ,  $\beta$ , and  $n$ , we generated 100 event time series with randomly drawn  $\mu^* \in [10, 350]$ . For test cases with multiple periodicities, we enforced the difference between the involved periodicities to be bigger than  $\log(\mu)$ . Otherwise, the generative curves become indistinguishable too quickly, making multiple periodicity detection too ill-posed.

The combination of the above model parameter settings resulted in 28,800 test cases for each generative model. All algorithms were applied to identify the underlying periodicities for every generated test case.

An identified periodicity is considered correct if it lays within  $\mu^* \pm 0.5 \cdot \sigma$ , where  $\mu^*$  is the true periodicity and  $\sigma$  the corresponding variance. For instance, for the cases  $\mu^* = 15$ ,  $\sigma = 2$  and  $\mu^* = 350$ ,  $\sigma = 44$ , a guess of  $\mu$  within  $15 \pm 1$  and  $350 \pm 22$ , respectively, would be considered an accurate detection.

Thus, for a fixed configuration of the parameters  $\sigma$ ,  $\beta$ , and  $n$ , the performance of the algorithms is measured by accuracy, which is the average number of correctly identified periodicities (across the 100 generated test cases) with a value between zero and one.

In the following, we first present the results for  $|\mu^*| = 1$  and identify valid ranges for  $n$ ,  $\beta$ , and  $\sigma$ . Second, within the valid range, we compare the performance of the GMPDA to that of the other algorithms for  $|\mu^*| = 1, 2, 3$ .

### 4.2. Performance with respect to $|\mu^*| = 1$

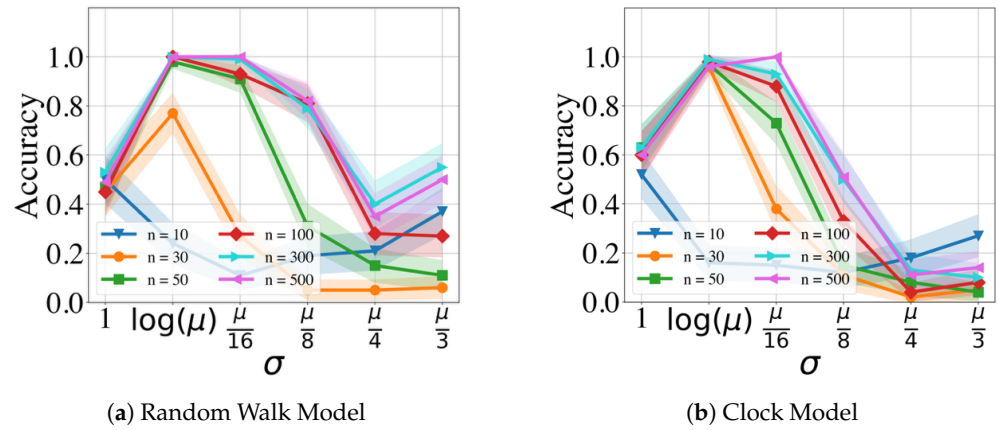
#### 4.2.1. GMPDA Performance

In this section, we focus on the performance of the GMPDA with respect to  $|\mu^*| = 1$  to determine realistic limits for  $\sigma$ ,  $\beta$ , and the number of events.

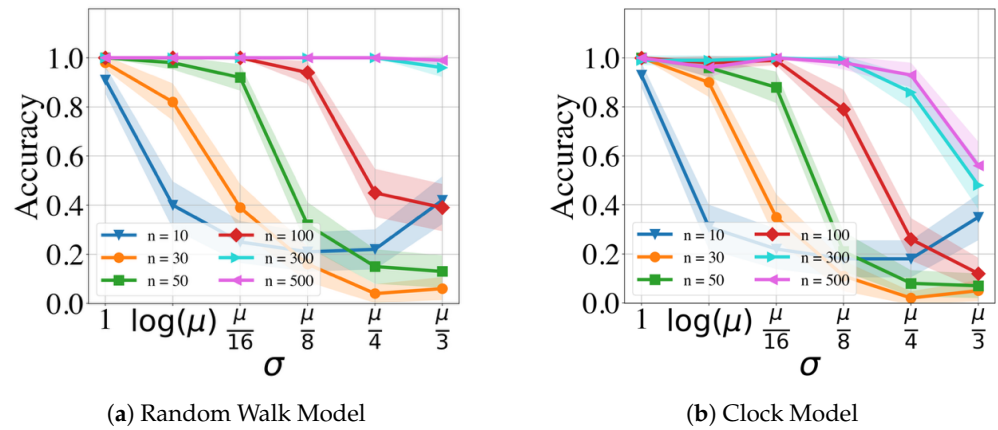
Figures 3 and 4 display the performance of the GMPDA for fixed  $\beta = 0$  and  $|\mu| = 1$  with varying values of  $\sigma$  and different numbers of events  $n$ , without and with curve fitting, respectively. Curve fitting is an optional step in the GMPDA that helps optimize the estimate for  $\sigma$  once the algorithm has identified the candidate periodicities. The confidence intervals (CI) in all the following figures (if present) are estimated as  $\bar{x} \pm 1.96 \text{ SEM}$ , where  $\bar{x}$  is the mean and SEM is the standard error of the mean.

The results in Figures 3 and 4 show, as expected, that accuracy is decreased with increasing  $\sigma$  and decreasing the number of events. In other words, with increasing variance, more events are required for an accurate detection.

The figures also compare the performance of the GMPDA with and without curve fitting. The GMPDA without curve fitting performed worse, except in the case of  $\sigma = \log(\mu)$ . This behavior can be explained as follows: in the algorithm, the default initialization value of  $\sigma$  is  $\log(\mu)$ , and therefore for this configuration, the GMPDA without curve fitting worked with a known sigma. In all the other cases, the GMPDA with curve fitting provided better results.



**Figure 3.** Performance of the GMPDA *without* curve fitting for the Random Walk Model (a) and for the Clock Model (b), with  $\beta = 0$  and  $|\mu| = 1$  and varying number of events ( $n$ ).

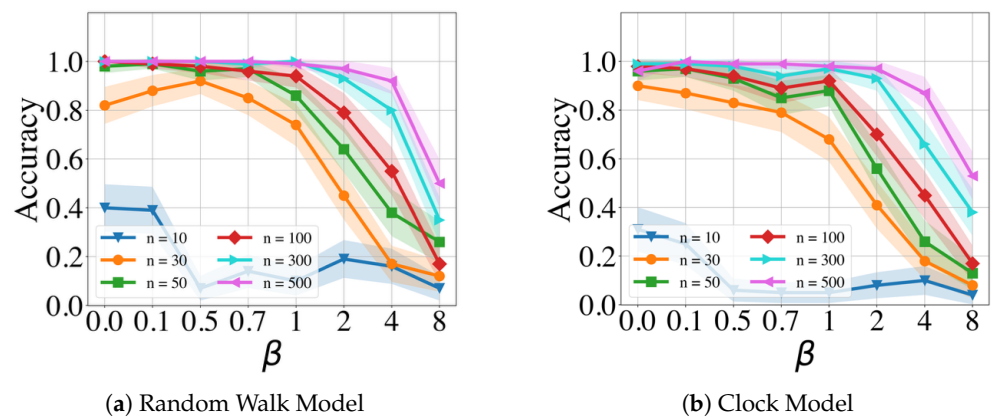


**Figure 4.** Performance of the GMPDA *with* curve fitting for the Random Walk Model (a) and for the Clock Model (b), with  $\beta = 0$  and  $|\mu| = 1$  and varying number of events ( $n$ ).

Next, to compare the effect of noise, we restricted our evaluation to the GMPDA with curve fitting due to its better performance. Please note that the comparison between results with and without curve fitting can be found in Appendix B.1. Further, we focused on the case of  $|\mu^*| = 1$  and known  $\sigma$ , which can be viewed as an *ideal* scenario, as only  $\mu$  needs to be estimated. For this ideal case, we compared the effect of varying noise levels across a varying number of events on detection accuracy. Figure 5 shows the performance of the different algorithms with respect to increasing the amounts of noise in the time series, for the case with  $|\mu| = 1$  and  $\sigma = \log(\mu)$ , and separately for Random Walk and the Clock Models.

For the Random Walk Model (Figure 5a), performance was acceptable for signals with  $n \geq 300$  and noise up to  $\beta = 4$ ; for  $n \leq 300$ , performance dropped below 0.75 already for  $\beta \geq 2$ . In comparison, the Clock Model was substantially more sensitive to noise (Figure 5b) with acceptable results only for  $\beta \leq 1$ .

In summary, in cases where the actual variance is unknown, GMPDA with curve fitting outperformed the GMPDA without curve fitting. The GMPDA was not suited for cases with fewer than 50 events. The GMPDA performance increased with the number of events. The GMPDA could also handle moderate to high amounts of noise, and we show in the next section how this compares to other periodicity detection algorithms.



**Figure 5.** Performance of the GMPDA with curve fitting for the Random Walk Model (a), and for the Clock Model (b), with  $|\mu| = 1$  and  $\sigma = \log(\mu)$  across varying levels of uniform noise beta and number of events.

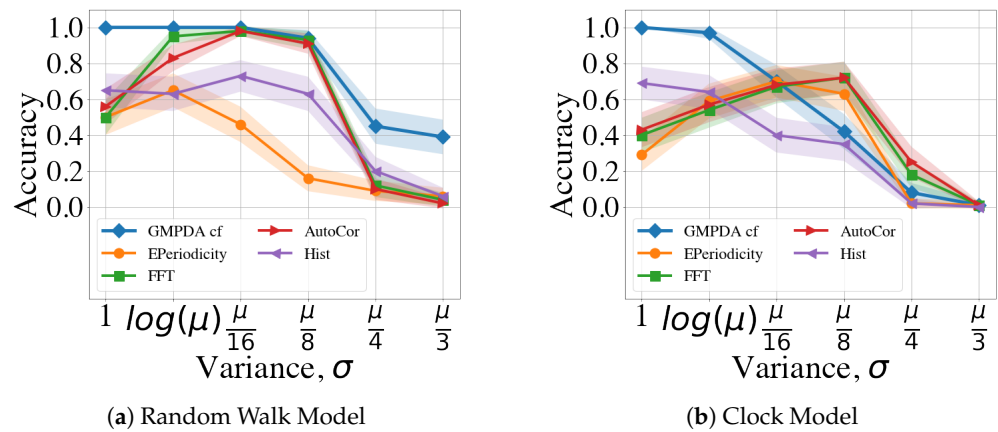
#### 4.2.2. Comparison with Alternative Periodicity Detection Algorithms

Next, we compared the GMPDA (with curve fitting) algorithm to other periodicity detection algorithms regarding their performance under varying conditions of noise and variance. As noise and variance increase, the histograms of the inter-event intervals analyzed by all algorithms become less informative, making the peaks that indicate periodicities less identifiable. Therefore, we investigated the sensitivity to noise and different variances used for generating the periodicities. We first examined the effect of varying levels of variance  $\sigma$  for cases where no noise was present, i.e.,  $\beta = 0$ .

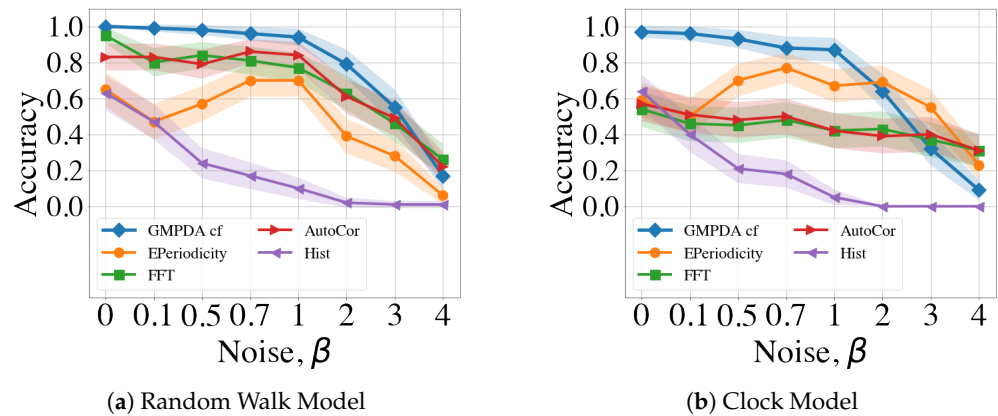
The results for all algorithms and  $n = 100$  are shown in Figure 6. The results for different numbers of  $n$ , averaged over all levels of  $\beta$ , can be found in Appendix B.3. For the Random Walk Model, the GMPDA was very accurate up to  $\sigma = \frac{\mu}{8}$ . Interestingly, all other algorithms performed worse when variance was very small ( $\sigma = 1$  and  $\sigma = \log(\mu)$ ), a case where the GMPDA excelled. FFT and AutoCor converged to the accuracy bound given by the GMPDA for  $\sigma > 1$ , while the accuracy of E-periodicity and Hist had its maximum of about 0.8. For all methods, the performance dropped for  $\sigma \geq \frac{\mu}{8}$ . This behavior is distinctive for the Random Walk Model, where the variance increases with every step, causing the generative distributions to start overlapping more quickly with larger variance. Performance was generally lower for the Clock Model, which was also more sensitive to increases in the variance. The GMPDA was sufficiently accurate only for  $\sigma = 1$  and  $\sigma = \log(\mu)$ , with a distinct drop in performance with increased variance. For the other algorithms, except the histogram method, performance initially increased with increasing variance up to  $\sigma = \frac{\mu}{8}$  and then declined sharply.

Next, we evaluated the performance of all methods with respect to increasing levels noise, with the results shown in Figure 7. For these analyses, the variance was fixed to  $\sigma = \log(\mu)$  and number of events to  $n = 100$ . The plots for all numbers of  $n$  can be found in Appendix B.2. For the Random Walk Model, the GMPDA was insensitive to noise up to  $\beta = 1$ , with performance decreasing linearly thereafter. The performance of FFT and AutoCor mirrored that of the GMPDA with slightly lower levels of accuracy. Notably, E-periodicity's performance increased up to  $\beta = 1$  and then declined, while Hist was very sensitive to all levels of noise and performed worse than all other algorithms.

For the Clock Model, the GMPDA behaved similarly, while the performance of the other methods was more sensitive to noise, and accuracy was generally lower than for the Random Walk Model.



**Figure 6.** Comparison of the GMPDA to alternative algorithms: for the Random Walk Model (a), and for the Clock Model (b). Accuracy is plotted for different levels of variance  $\sigma$  for cases with one period ( $|\mu| = 1$ ), no noise ( $\beta = 0$ ) and number of events,  $n = 100$ .



**Figure 7.** Comparison of the GMPDA to alternative methods: for the Random Walk Model (a), and for the Clock Model (b). Accuracy is plotted against increasing levels of noise  $\beta$  for cases with one period ( $|\mu| = 1$ ), known variance, i.e.,  $\sigma = \log(\mu)$  and number of events,  $n = 100$ .

The presence of moderate noise (i.e., with  $\beta \in [0.1, 0.7]$ ) did not affect performance, except for E-periodicity, where performance increased for noise levels up to  $\beta = 2$ .

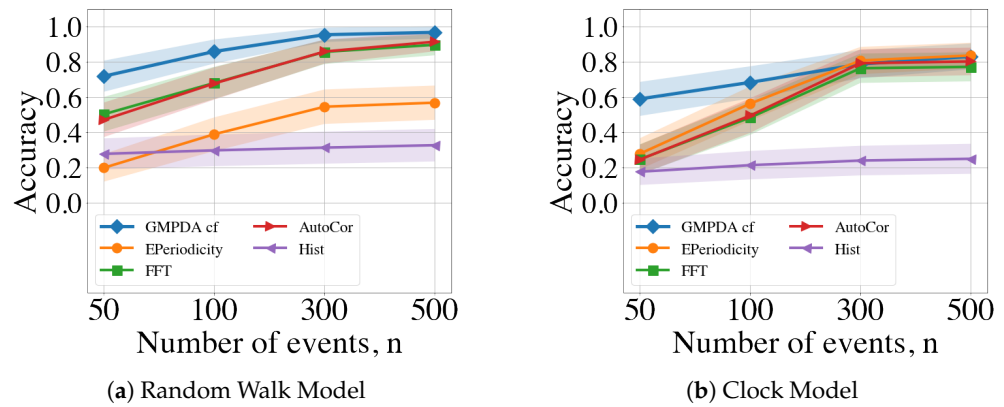
The maximal noise levels that the algorithms could handle were not higher than two  $\beta \leq 2$ , i.e., a signal-to-noise ratio of 1:2, one periodic event to two noise events.

In conclusion, we averaged performance over all acceptable values of noise and variance (i.e.,  $\sigma = \{1, \log(\mu), \frac{\mu}{16}, \frac{\mu}{8}\}$  and  $\beta \leq 2$ ). The results are shown in Figure 8. Overall, the detection of a single periodicity was increasingly accurate with an increasing number of events for all methods and both the Random Walk and Clock Models (see Figure 8). For both models, the periodicity detection with the Hist algorithm had very low accuracy with a maximal performance of less than 0.4.

For the Random Walk Model, the GMPDA outperformed alternative approaches, with accuracy converging to one as the number of events increased, and even for  $n = 30$ , its performance was larger than 0.75. FFT / Autocor achieved similar performance when the number of events was larger than 300. In contrast, EPeriodicity's performance for the Random Walk Model was relatively poor, with a maximum of 0.6 for 500 events.

For the Clock Model, the GMPDA outperformed alternatives when the number of events was smaller than 300. For more than 300 events, the performance of all approaches, except Hist, became equally good.





**Figure 8.** Comparison of the GMPDA to alternative methods for the Random Walk Model (a), and for the Clock Model (b). Accuracy is plotted against the number of events averaged over  $\sigma = \{1, \log(\mu), \frac{\mu}{16}, \frac{\mu}{8}\}$  and  $\beta \leq 2$ .

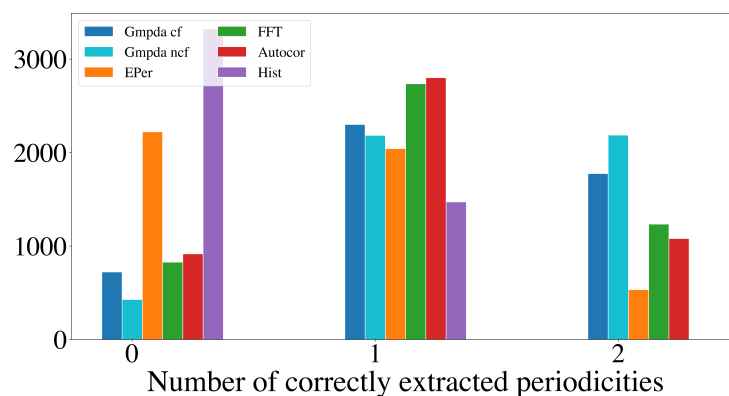
#### 4.3. Performance with Respect to $|\mu^*| > 1$

This section compares the performance of the GMPDA (with and without curve fitting) to that of the alternative methods for multiple periodicity detection, focusing on the set of sensible simulation parameters identified in Section 4.2.2. These parameters are  $n = 50, 100, 300, 500$ ,  $\sigma = \{1, \log(\mu), \frac{\mu}{16}, \frac{\mu}{8}\}$ , and  $\beta \leq 1$ , resulting in 8000 test cases for each setting of  $|\mu| = 2$  and  $|\mu| = 3$  for every generative model. For comparison, the performance is summarized over  $n, \mu, \sigma$ , and  $\beta$ , visualized here as a histogram, where the  $x$ -axis displays the number of correctly detected periodicities and the  $y$ -axis the number of test cases.

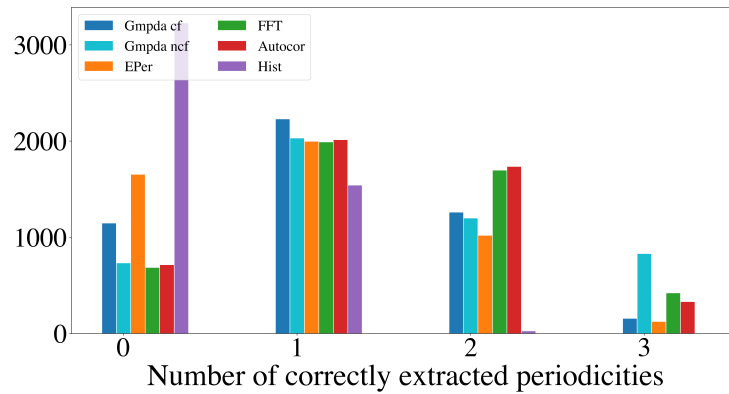
Figures 9 and 10 show the results for the Random Walk Model, and Figures 11 and 12 show the results for the Clock Model for  $|\mu| = 2$  and  $|\mu| = 3$ , respectively.

For the case with two periodicities,  $|\mu| = 2$ , the GMPDA outperformed the alternative methods, both with and without curve fitting. Interestingly, the GMPDA without curve fitting performed slightly better, suggesting that the current sigma optimization might require further development.

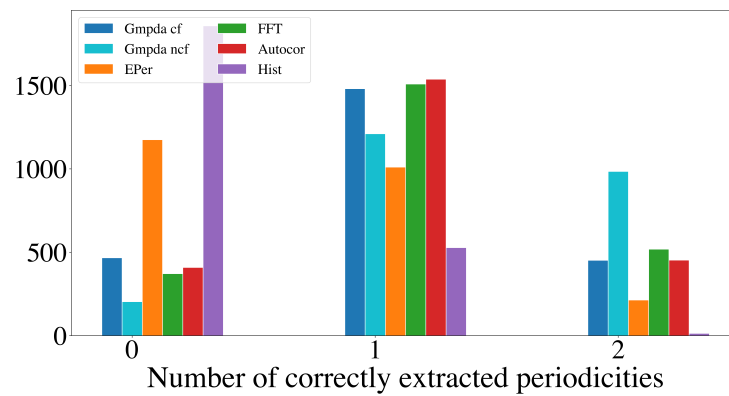
The detection of three periodicities,  $|\mu| = 3$ , was challenging for all methods as shown in Figures 10 and 12. One possible explanation is that with more periodicities, there are more interaction intervals, i.e., intervals between the periodic events from different periodicities. Furthermore, at least for the Random Walk Model, the histogram becomes less identifiable as  $\sigma$  grows with each subsequent step, flattening out the distribution responsible for the events. This effect is amplified when more than one periodicity is present. We conclude that the GMPDA in the current version is not well suited for detecting more than two periodicities.



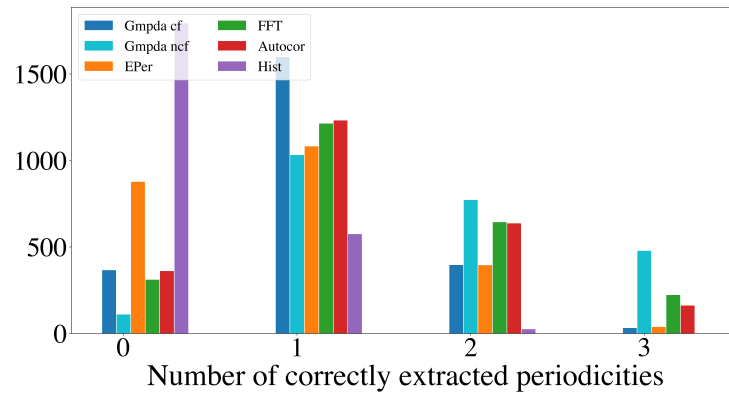
**Figure 9.** Detection of multiple periodicities ( $|\mu| = 2$ ) quantified as the number of correctly extracted periodicities by the GMPDA and alternative methods for the Random Walk Model.



**Figure 10.** Detection of multiple periodicities ( $|\mu| = 3$ ) quantified as the number of correctly extracted periodicities by the GMPDA and alternative methods for the Random Walk Model.



**Figure 11.** Detection of multiple periodicities ( $|\mu| = 2$ ) quantified as the number of correctly extracted periodicities by the GMPDA and alternative methods for the Clock Model.



**Figure 12.** Detection of multiple periodicities ( $|\mu| = 3$ ) quantified as the number of correctly extracted periodicities by the GMPDA and alternative methods for the Clock Model.

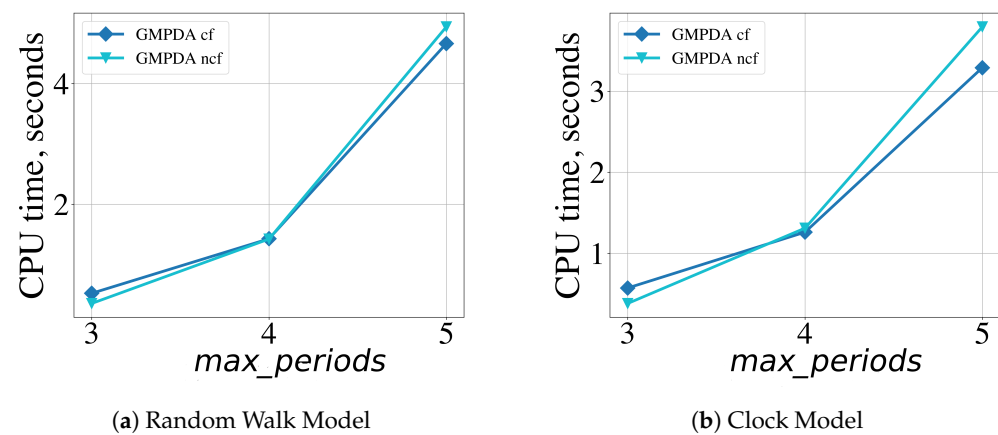
#### 4.4. Computational Performance

The computational performance (CPU time) of the GMPDA was evaluated across different experiments. For this purpose, time series were generated for every combination of the following model parameters:  $|\mu^*| = [1, 2, 3]$ , events per periodicity =  $[50, 100, 300, 500]$ ,  $\sigma^* = [\log(\mu)]$ ,  $\beta = [1]$ . The GMPDA was executed for each time series, and the computational/execution time was determined using the Python module `timeit` with 100 executions. For the generated test cases, we tested them with the following GMPDA configurations (described in Section 3):  $loss\_length = [400, 800, 1200]$  and  $max\_periods = [|\mu| + 2]$ . The remaining parameters were fixed at  $L_{min} = 5$ ,  $max\_iterations = 5$ ,  $max\_candidates = 15$ ,  $noise\_range = 5$ ,  $loss\_tol\_change = 0.01$ .

Our analysis shows that the computational performance strongly depended on the maximum number of allowed periodicities, *max\_periods*. The CPU time for both models (averaged over the number of executions, number of events  $n$ , and loss length) is shown in Figure 13. All other parameters had a comparatively minor influence on the performance. In additional experiments not shown here, we also investigated the influence of noise  $\beta$  on the computational performance of the algorithm.

The results indicated that although, on average, the CPU time increased slightly with increasing noise  $\beta$ , the influence was minimal when compared to the maximum number of allowed periodicities, *max\_periods*. Finally, the maximal number of candidates periods *max\_candidates* also affected the CPU time: a lower *max\_candidates* resulted in faster execution time but decreased the accuracy of the algorithm.

The proposed GMPDA is optimized by vectorizing all major computations. Due to the hierarchical structure of the algorithm, the computational time will depend on the maximum number of periodicities. The computationally costly part arises from the curve fitting optimization, which is negligible as shown in Figure 13. The algorithm's memory requirements are linearly dependent on the length of the considered time series. Therefore, the algorithm is scalable and applicable for real data applications.



**Figure 13.** Effect of the GMPDA parameter *max\_periods*, the maximum number of searched for periodicities, on computational performance averaged over 1200 executions for the GMPDA, with (dark blue symbols) and without curve fitting (light blue symbols).

#### 4.5. Summary

We have evaluated the performance of the GMPDA across a large set of test cases, covering different configurations of the Random Walk and the Clock Models. Our main findings are as follows: First, for time series following the Random Walk Model, the GMPDA outperformed alternative algorithms. Second, for time series following the Clock Model, the GMPDA outperformed alternative methods in cases with low variance of the inter-event intervals. Third, all algorithms struggled to identify more than two periodicities.

Additionally, we analyzed the sensitivity to critical simulation parameters across the different algorithms and found that both sigma and the number events emerged as the strongest determinants of periodicity detection accuracy. The details of the sensitivity analysis can be found in Appendix B.4.

### 5. Real Application

Finally, we applied the GMPDA to real data, specifically to the recording of leg movements during sleep from the publicly available MrOS data set [21–25].

From 2905 available sleep recordings in community-dwelling men 67 years or older (median age 76 years), we considered all recordings with at least 4 h of sleep, a minimum of 10 leg movements and 10 arousals, and adequate signal quality based on various parameters

in the MrOS database. This resulted in 2650 recordings satisfying our inclusion criteria, from which we randomly selected 100 recordings for this real application case.

We chose to examine leg movements during sleep because it is known that in a relatively large proportion of the population (up to 23% [26]), these leg movements tend to occur in a periodic pattern, known as periodic leg movements during sleep (PLMS) [27], with a typical inter-movement interval around 20 to 40 s [28]. We, therefore, expected to find some amount of periodicity in this data set, making this analysis a real-life positive control.

We applied the GMPDA to both raw and preprocessed data. In the preprocessing step, the time series of leg movements for each subject was segmented into *sleeping* bouts according to the following criteria: Each bout (i) contained only sleep interrupted by not more than 2 min of wakefulness, (ii) lasted at least 5 min, and (iii) contained at least four leg movements. This resulted in 579 sleep bouts from the 100 recordings where the GMPDA was applied independently to each bout. The number of events was less than 100 for 85% of the bouts, and for those, the average bout length was 2572 s.

### 5.1. GMPDA Configurations

The following GMPDA parameters were fixed for both data sets (i.e., whole night data and sleep bout data):  $L_{min} = 5$ ,  $L_{max} = 200$ ,  $max\_iterations = 5$ ,  $max\_candidates = 15$ ,  $loss\_length = 400$ ,  $max\_periods = 5$ ,  $noise\_range = 5$ ,  $loss\_tol\_change = 0.1$ . We chose a tolerance value for a decrease in the loss of 0.1, meaning additional periodicities are only considered if their inclusion results in a change in loss greater than this tolerance value. This value is substantially higher than in the simulated examples (0:01) because, in this first real-life application, we aimed to generate robust results given the expected noise in the data. In this context, the results presented here and the periodicities identified can be seen as “low-hanging fruit”. Moreover, the detection of additional periodicities would be expected with different GMPDA parameters.

For the MrOS data set, we assumed a Random Walk Model, which we applied both with and without the curve fitting of the variance parameter  $\hat{\sigma}$ . Consistent across all single records, the curve fitting approach identified periodicities with a lower loss, so we will describe only the curve fitting results in the following. The GMPDA loss with and without curve fitting is compared in Appendix B.5, Figure A7.

### 5.2. Reference Loss

The GMPDA identifies the periodicity with minimal loss. However, even if minimal, this loss might still be numerically significant. In a real-life application where it can be assumed that some of the time series do not contain periodic events, it is necessary to identify loss values that do not support the existence of periodicities in the data. We address this issue: we constructed a reference loss, derived from the minimal GMPDA loss returned for times series that contain only random noise.

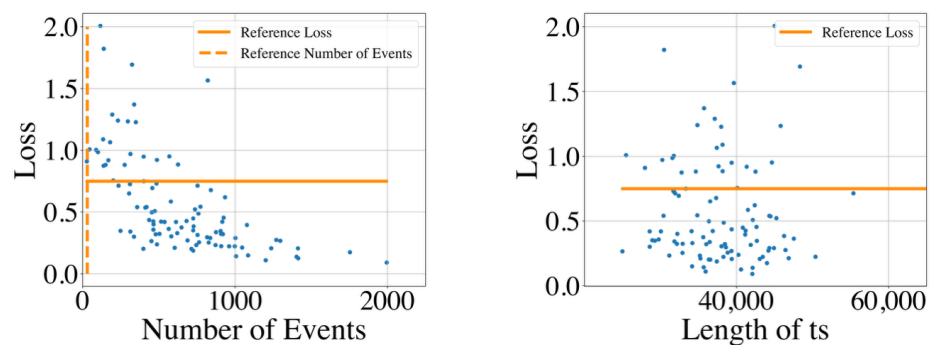
For the MrOS data set, the length of the included bouts and the number of events ranged from 300 to 24,000 seconds and 5 to 430 events, respectively. To obtain an overall reference loss, we constructed 100 noisy bouts with uniformly distributed events for all different combinations of the number of events [10; 30; 50; 100; 200; 400] and length of the bout [500; 1000; 2000; 4000; 8000; 16,000]. Applying the GMPDA to each combination, we obtained an empirical distribution of loss values for cases where the events were generated randomly and did not exhibit any clear periodic pattern. The global MrOS reference loss is set to the 0.01 quantile of this distribution, corresponding to a value of 0.74468, rounded to 0.75 in the following.

Additionally, we estimated a local reference loss for each bout in the MrOS data set by generating 100 time series with the bout-specific length and the number of events and taking 0.01 quantile of the resulting loss distribution. A significant periodicity was identified when the GMPDA loss for this bout was lower than the local reference loss. However, the significant periodicities obtained with local and global reference losses did

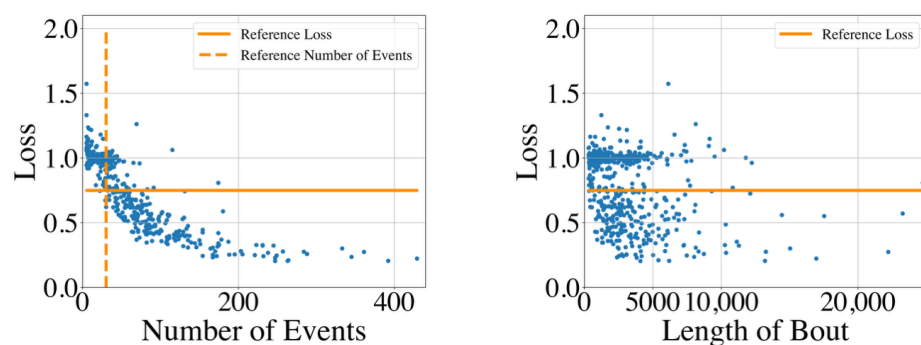
not differ significantly, and for simplicity, we focus on the results obtained for a global reference loss of 0.75.

### 5.3. Results

The distribution of the GMPDA model loss for all time series is shown in Figure 14 for the whole night recording and in Figure 15 for the single sleep bouts. The figures suggest that the GMPDA loss did not systematically change with the length of the times series. However, the loss tended to decrease with the number of events in the time series. More specifically, as already seen in the simulation experiments, for time series with a low number of events, the resulting loss was not distinguishable from the loss found for non-periodic time series. The left panel of Figure 15, which shows the distribution of the loss for the number of events in the MrOS data set, could also suggest a minimum number of events needed for the GMPDA to detect a significant periodicity in this data set. For the records selected here, no significant periodicity was detected for any bout with fewer than 30 events (see reference number of events in Figure 15). Further analysis with other records from the same data set and new data sets is needed to determine whether this reference number constitutes an absolute threshold for biomedical event data.



**Figure 14.** The GMPDA loss for 100 whole night time series plotted against the number of events (**left panel**) and length of time series (ts, in seconds, **right panel**).

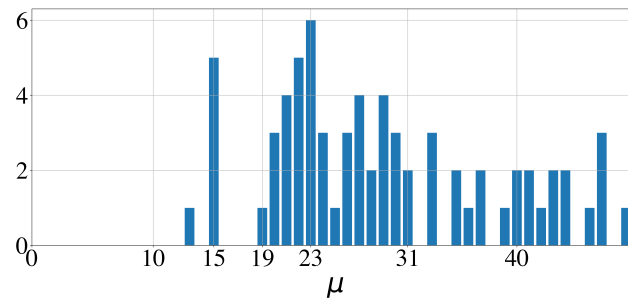


**Figure 15.** The GMPDA loss for 579 sleep bouts of at least 5 min plotted against the number of events (**left panel**) and the length of the sleep bout (in seconds, **right panel**).

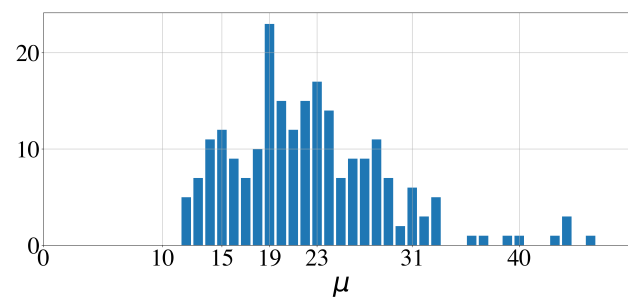
Out of the 579 sleep bouts and out of the 100 whole night time series, 183 (31.6%) and 75, respectively, had a loss below 0.75. The corresponding histograms of the significant periodicities extracted from the signals by the GMPDA are shown in Figures 16 and 17. In both figures, the expected peak in periodicities is around 20. Another minor, rather unexpected, peak is at 15. Significant periodicities ranged from 10 to 33 seconds (except two bouts with a periodicity of 49 and 192 s). Periodicities around 20, i.e.,  $\mu \in [17, 18, 19, 20]$ , were present in 95 bouts (out of 183) from 77 subjects (out of 100). Periodicities around 15, i.e.,  $\mu \in [12, 13, 14]$ , were present in 30 bouts from 18 subjects.

Although the minimal periodicity and noise range were set to 5,  $\mu = 12$  was the smallest periodicity identified by the algorithm for significant bouts.





**Figure 16.** Histogram of significant periodicities identified in 100 whole night time series by the GMPDA.



**Figure 17.** Histogram of significant periodicities identified in 579 sleep bouts, 5 min or longer, by the GMPDA.

## 6. Conclusions

In this paper, we developed the Gaussian Mixture Periodicity Algorithm (GMPDA) to address the challenge of detecting overlapping periodicities in noisy data. The GMPDA is based on a novel generative model scheme that explicitly accounts for both a Clock Model and a Random Walk Model. The Clock Model describes periodic behavior in systems where variances do not change over time due to a governing pacemaker, such as scheduled or seasonal behavior like traffic patterns or migration patterns. In contrast, the Random Walk Model describes systems where variances increase over time, making distant temporal predictions difficult or impossible, such as biological behaviors like footsteps or gene expression where events depend only on the interval to the last event.

The primary entry point for the GMPDA is the empirical histogram of all forward-order inter-event intervals. This histogram contains information about the underlying prime periodicities, interaction noise between events associated with different periodicities, and false positive noise. We approximate the overall noise using an explicit formulation under the assumption that the noise is uniformly distributed. This approximation accounts for all interaction intervals, whose lengths are limited by a user-defined parameter in the GMPDA. After subtracting this noise, the GMPDA hierarchically extracts multiple overlapping periodicities by minimizing the loss, defined as the absolute difference between the parametrized histogram obtained by the generative scheme and the empirical histogram.

The GMPDA is implemented in a computationally efficient manner and is available as open-source software on <https://github.com/nnaisense/gmpda> (accessed on 18 July 2024). We have demonstrated its performance on a set of test cases, including scenarios with up to three overlapping periodicities, different values for Gaussian noise, and varying number of events. For the Random Walk Model, the GMPDA outperformed the FFT and autocorrelation-based approaches as well as the E-periodicity algorithm in identifying true prime periodicities. For the Clock Model, the GMPDA outperformed other algorithms in cases with low variance of intervals.

The GMPDA performed well in the presence of noise with a signal-to-noise ratio of 1:1 and performed adequately up to a ratio of 1:2, given an appropriate number of events. This appropriate number of observed events depends on the signal-to-noise ratio,

but more than 30 actual periodic events are generally required for the GMPDA to identify any periodicity.

Finally, we applied the GMPDA to extract significant periods in real data, focusing on leg movements during sleep. The main results here were (i) that the GMPDA was able to identify the expected periodicities around 20 s, (ii) we introduced a procedure to identify a data set-dependent reference loss (of 0.75) to distinguish significant from spurious periodicities, and (iii) our results suggest that there is a minimal number of events (30) required for the GMPDA to perform periodicity detection successfully in biomedical data.

The GMPDA has demonstrated robust performance in detecting periodicity within the framework of the Clock and Random Walk Models. These models are effective for a broad range of scenarios; however, we acknowledge that this focus introduces a limitation in environments where these models may not adequately capture the underlying dynamics of the signal, such as in certain biological systems or financial time series. Nevertheless, the general nature of the generative framework and the formulation of the GMPDA allows for alternative statistical parametrization for the event data. An extension could involve modeling events as a Poisson process, which for multiple periodic generative functions could be modeled as a sum of scaled probability density functions. Additionally, the GMPDA could be extended to periodicity extraction in non-stationary event time series. One approach could involve dividing the time series into locally stationary segments and using a bottom-up segmentation strategy to estimate optimal switching points and prime periodicities for each segment. Another approach could incorporate a Monte Carlo-based particle approach for adaptive periodicity detection as presented in [18]. These extensions remain for future work.

**Author Contributions:** Conceptualization, O.K. and S.F.; Formal analysis, E.B., O.K. and S.F.; Methodology, E.B., O.K., J.M., E.C.W. and S.F.; Project administration, O.K.; Software, E.B., O.K. and S.F.; Supervision, O.K. and S.F.; Visualization, O.K. and S.F.; Writing—original draft, E.B., O.K. and S.F.; Writing—review and editing, E.B., O.K., J.M., E.C.W. and S.F. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original data taken from the MrOS Sleep Study are openly available at the National Sleep Research Resource (NSRR) at <https://doi.org/10.25822/kc27-0425> (accessed on 17 July 2014).

**Acknowledgments:** Stephany Fulda is supported by Swiss National Science Foundation (SNSF) grants No. 320030 160009 and 320030 179194. The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, “Outcomes of Sleep Disorders in Older Men,” under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation Function
FFT	Fast Fourier Transform
GMPDA	Gaussian Mixture Periodicity Detection Algorithm
GMPDA cf	GMPDA with curve fitting
ts	time series

## Appendix A. GMPDA

In this section, we outline in more detail the steps involved in the GMPDA.

### Appendix A.1. Approximation of $|\zeta(\mu)|$

In Equation (16), we estimate the number of interaction intervals as a decreasing linear function  $\mathbb{E}[\hat{\zeta}(\mu)] = z(1 - \frac{\mu}{N_T})$ . An approximation of  $z$  is required, as the information about the amount of noise and the number of true periodicities is unknown a priori. Here, we propose the following approximation:

$$\hat{z} = \frac{1}{z_{min}} \sum_{i=1}^{z_{min}} D(\mu)(i). \quad (A1)$$

This approximation follows the idea that first,  $z$  should be close to the maximal value of  $\mathbb{E}[\hat{\zeta}(\mu)]$  and that second, all non-zero contributions in  $D(\mu)$  for  $\mu < \underset{\mu}{\operatorname{argmin}} \mu^*$  are due to the interaction intervals, asymptotically distributed for an increasing number of events and/or increasing number of involved prime periodicities. For the GMPDA, we set default  $z_{min} = L_{min} - 1$ . Thus, in the algorithm, the length of the interaction intervals is limited either by the minimal expected periodicity, or it can be adjusted by the user.

We verified this approximation empirically on a set of 50,000 test cases with and without noise and a priori known two randomly chosen prime periodicities  $\mu_{*1/2} \in [10, 60]$ :  $\mathbb{E}[\hat{\zeta}(\mu)]$  with  $\hat{z}$  provided on average a good linear fit, since the distribution of the mean errors between  $\mathbb{E}[\hat{\zeta}(\mu)]$  and  $\mathbb{E}[\zeta(\mu)]$  was centered at zero and was approximately normal.

However, it must be stressed that the assumption of an uniform distribution of interaction intervals between the periodic and the noise events may be unrealistic in real life data sets, and there is currently no alternative available for estimating zeta from the observed data. This remains an area of improvement for the GMPDA.

### Appendix A.2. Candidate Period Identification

The proposed algorithm hierarchically extracts a set of candidate periodicities, which can explain  $D(\mu)$  using an integral convolution approach. The method works by iteratively selecting periodicities, which explain many of the observed intervals, and then subtracting the integer multiple intervals which can be explained by these periodicities.

The algorithm takes as input some guess  $\hat{\sigma}$ , a range in which to search for periodicities  $\{L_{min}, L_{max}\}$ , a number of hierarchical periodicity extraction steps *max\_iterations*, and a maximal number of periodicities to extract at each hierarchical iteration, *max\_candidates*.

Recall that  $D(\mu)$  counts the number of times a given interval  $\mu$  appears between any two events in the time series

$$D(\mu) = \sum_{i,m>0} \mathbb{1}_{s_{i+m}-s_i=\mu}. \quad (A2)$$

One tempting method would be to select  $\operatorname{argmax}_{\mu} D(\mu)$  as the first prime period  $\hat{\mu}_1$ . However, for small or noisy real-world data,  $\operatorname{argmax}$  may not be the prime period.

We introduce the notion of an *integral convolution*, in which we integrate around a fixed  $\mu$  to capture how much of the observed intervals in  $D(\mu)$  are explainable by that particular “mean” periodicity, which also acts to smooth  $D(\mu)$ . We therefore define a function  $\tau(\hat{\mu}, \hat{\sigma}, D(\mu))$ , which will act as a symmetric convolution kernel across  $D(\mu)$ , centered at the candidate means  $\hat{\mu}$ . This function provides a point-wise estimate of the explained data for a given  $\mu$  in  $D(\mu)$ :

$$\tau(\hat{\mu}, \hat{\sigma}, D(\mu)) = \int_{\hat{\mu}-(1.96*\hat{\sigma})}^{\hat{\mu}+(1.96*\hat{\sigma})} D(\mu) \partial\mu. \quad (A3)$$

Because we do not want to calculate the full loss function (32) at this stage, due to computational expense, we approximate our loss function with the function  $\tau(\hat{\mu}, \hat{\sigma}, D(\mu))$  for the Clock Model:

$$\mathbb{E}(\hat{\mu}_p) = \frac{\hat{\mu}_p}{N_T} \cdot \left[ \sum_{i=1}^{\frac{N_T}{\hat{\mu}_p}} \tau(i \cdot \hat{\mu}_p, \hat{\sigma}, D(\mu)) \right], \quad (\text{A4})$$

and for the Random Walk:

$$\mathbb{E}(\hat{\mu}_p) = \frac{\hat{\mu}_p}{N_T} \cdot \left[ \sum_{i=1}^{\frac{N_T}{\hat{\mu}_p}} \tau(i \cdot \hat{\mu}_p, i \cdot \hat{\sigma}, D(\mu)) \right]. \quad (\text{A5})$$

Functions (A4) and (A5) approximate, for each candidate prime period, how much of the data can be explained by this periodicity in some confidence interval about  $\mu_p$ . If a periodicity is present and is persistent through the time series, integer multiples of the periodicity  $\hat{\mu}_p$  will also frequently appear in (A4) and (A5); we can use this information to select the periodicity which explains the most data.

Once the first periodicity  $\hat{\mu}_1$  has been identified, we remove the intervals found in  $D(\mu)$ , which can be explained by  $\hat{\mu}_1$ , i.e., integer multiples of  $\hat{\mu}_1$ . We realize this by setting  $D(\mu)$  to zero for  $\mu \in [i \cdot \hat{\mu}_1 + \hat{\sigma}, i \cdot \hat{\mu}_1 - \hat{\sigma}]$ ,  $i = 1, \dots$ , and recompute  $\tau(\mu)$  from  $D(\mu)$  missing these intervals.

The GMPDA performs reasonably well at identifying the true periods  $\mu^*$  as  $\hat{\mu}$  without the use of the loss function or adjustments to  $\hat{\sigma}$ . But without a measure of relative goodness of this estimates, we have no stopping criteria for finding multiple periodicities. Instead, we repeat this procedure for *max\_iterations* iterations.

Once we have initialized (hierarchically) a set of candidate prime periods  $\hat{\mu}^{init}$  using this “fast” method, we compute a better estimate of the variance and loss using methods which are elaborated in the following sections.

#### Appendix A.3. Non-Linear Least Squares Fitting for $\hat{\sigma}$

We can improve our guess of the variance  $\hat{\sigma}$  by formulating a non-linear least squares curve fitting optimization problem, in which our set of parameters comprises those of a Gaussian PDF. That is, here we consider  $D(\mu)$ , which can be modeled as the sum of Gaussian PDFs, and a set of candidate means  $\hat{\mu}$  for those Gaussian PDFs. For a fixed set  $\hat{\mu}_p$ , we initialize guesses for  $\hat{\sigma}_p$ , for  $p = 1, \dots, P$ , and deploy the Trust Region Reflective algorithm to obtain an update for the guesses of  $\hat{\sigma}_p$ . It is implemented with `curve_fit()` from Scipy’s optimization package.

#### Appendix A.4. Selecting True Parameters: Loss Function

The parameter estimates  $\hat{\mu}^{init}$ ,  $M$ ,  $\hat{\sigma}$  are assessed with respect to  $D(\mu)$ —the observed intervals between events in the time series—using a loss function. This loss function describes the proportion of intervals in the data, which can be explained with (i) the parametrized “generative function”  $G_M(\hat{\mu}, \hat{\sigma})$  implicated by the estimates, which is asymptotically the same as the expectation of  $D(\mu)$  if  $\hat{\mu} = \mu^*$  and  $\hat{\sigma} = \sigma^*$  and (ii) the noise approximation.

Computing the loss function (32) is expensive because in the case of the Random Walk Model,  $G_{RW}$  requires computation at increasingly large intervals since the variance terms grow linearly, and thus the area covered with some density by a single Gaussian distribution grows at the same rate. Thus, we only want to compute  $G_M$  for a few very probable periodicities (the set  $\hat{\mu}^{init}$  computed in the fast algorithm), using the optimal variance guesses  $\hat{\sigma}$ , and only across a limited range of intervals specified by *loss\_length* chosen a priori.

We also adjust the scaling factor of the generative function to account for sections of the time series which may not have any events, for instance, missing values or large intervals of the time series with no observations. This concerns the scaling factor  $c_p$ , for  $p = 1, \dots, P$ ,

of  $G_M(\hat{\mu}, \hat{\sigma})$  for the Clock Model and Random Walk Model in Equations (28) and (29), respectively. The adjusted scaling factor is

$$c_p = \frac{N_T(\hat{\mu}_p)}{\hat{\mu}_p} - (m - 1), \quad (\text{A6})$$

where  $N_T(\hat{\mu}_p)$  is the sum of intervals which are smaller than  $\hat{\mu}_p + (\hat{\sigma}_p \cdot 2)$ . This correction ensures that we only count “possible appearances” in the time series on sections which actually have events. Without this scaling factor, missing values would bias our results towards higher frequencies, and the scaling factor would be far too large for lower frequencies which may appear in the time series but with intervals of no events.

Our final loss function will therefore be constructed using  $D(\mu)$  computed from the real data,  $\mathbb{E}[\hat{\zeta}(\mu)]$ ,  $G_M(\hat{\mu}_{init}, \hat{\sigma})$ , and one additional parameter, *loss\_length*. This parameter manages high variance at high integer multiples and decreases the computational complexity. In the Random Walk Model, for high integer multiples of a periodicity, the implied Gaussian distributions of intervals begin to have large tails and the distribution density mean decreases. Meanwhile for the Clock Model, estimating many integer multiples is not actually necessary to compute the true periods. Therefore, the loss we compute in the algorithm is:

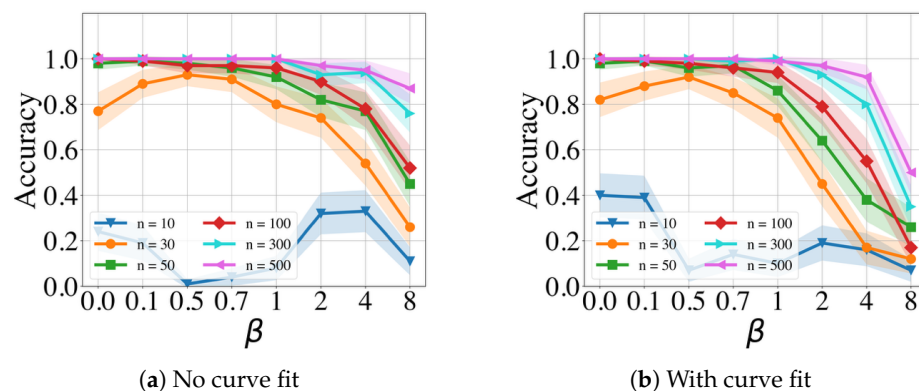
$$\hat{\mathcal{L}} = \sum_{\mu=0}^{\text{loss\_length}} |D(\mu) - \hat{\zeta}(\mu) - G_M(\hat{\mu}) \cdot \hat{\sigma}|, \quad (\text{A7})$$

Within the algorithm, we compute  $G_M$  for all combinations of the set  $\hat{\mu}^{init}$  up to order *max\_combi*, and select our true set of periodicities as that which minimizes (A7). Please note that the number of true periodicities *max\_combi* is not known a priori; the optimal value of *max\_combi* will minimize the loss. However, in real applications, we might have situations where there are weak peaks in  $D(\mu)$  around very large  $\mu$  due to noise or the influence of large/slow interactions intervals. Adding these to the set of prime periodicities will decrease the loss, but will not contribute to the identification of intrinsic periodicities. To account for this, the GMPDA provides the possibility to control the magnitude of the loss decrease by a parameter *loss\_decrease\_tol*, with the loss being typically of magnitude one and lower; see Section 5. That is, setting this tolerance parameter to a very low number, e.g., *loss\_change\_tol* = 0.001, will result in including more periodicities (that might be due to noise), while a larger number, e.g., *loss\_change\_tol* = 0.1, will be more conservative.

## Appendix B. Performance

### Appendix B.1. $|\mu| = 1$

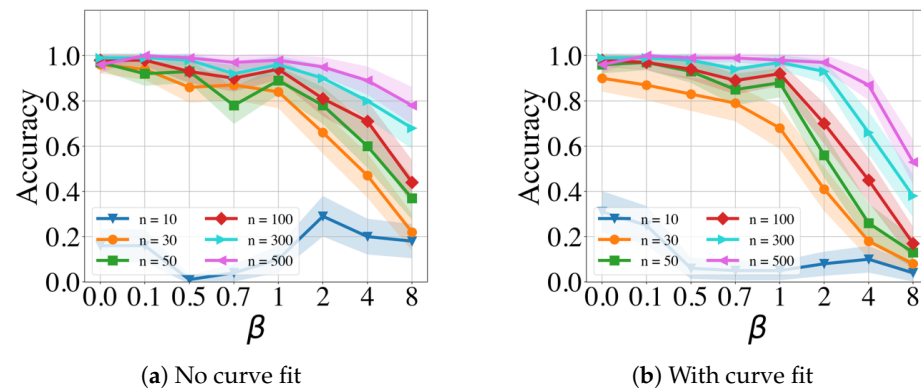
Here, the performance of the GMPDA with respect to noise with curve fitting and without curve fitting is presented.



**Figure A1.** Random Walk Model performance with respect to  $\beta$ , for  $\sigma = \log(\mu)$  and  $|\mu| = 1$ .



The Random Walk Model model exhibits a decay in performance with an increasing noise for  $n > 30$ . For  $n = 30$ , the performance increases until  $\beta = 0.5$ , as the noise, to a certain extent, is acceptable due to the definition of the variance for the Random Walk Model 5.

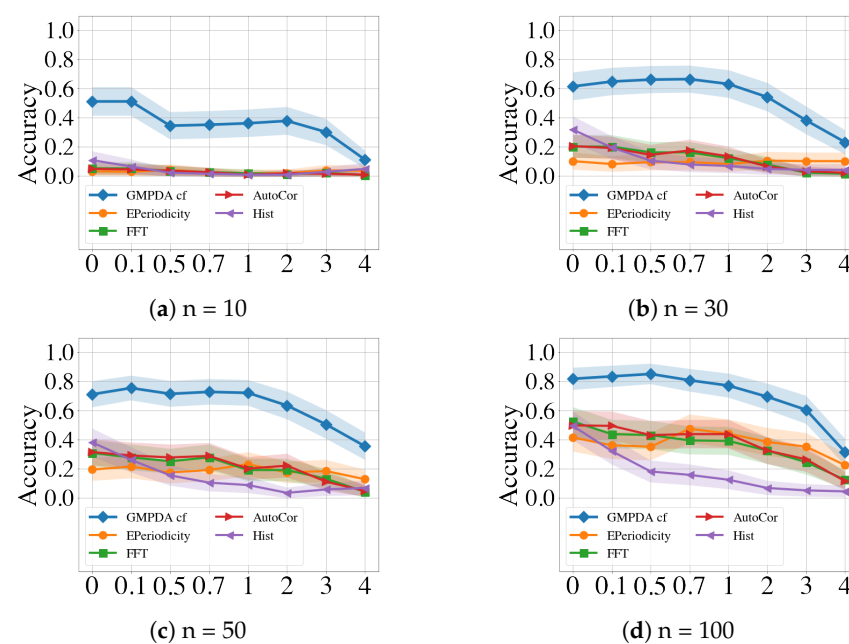


**Figure A2.** Clock Model performance with respect to  $\beta$ , for  $\sigma = \log(\mu)$  and  $|\mu| = 1$ .

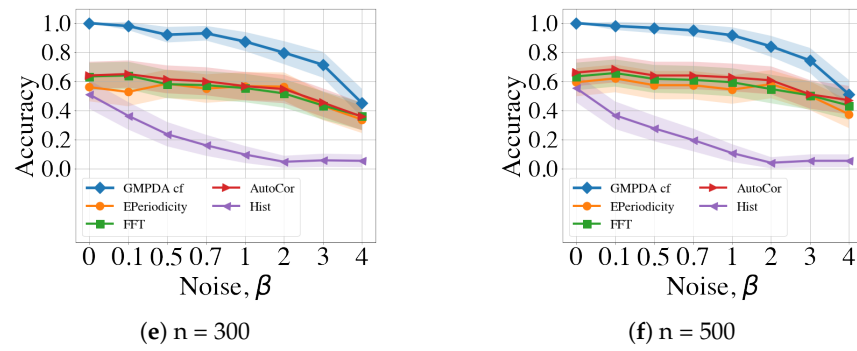
The Clock Model exhibits for  $n > 10$  a decay in performance with an increase in  $\beta$ . For both models, the case  $n = 10$  performs insufficiently, indicating that the number of events must be definitely higher than ten.

#### Appendix B.2. Comparison to Alternative Methods with Respect to Noise $\beta$

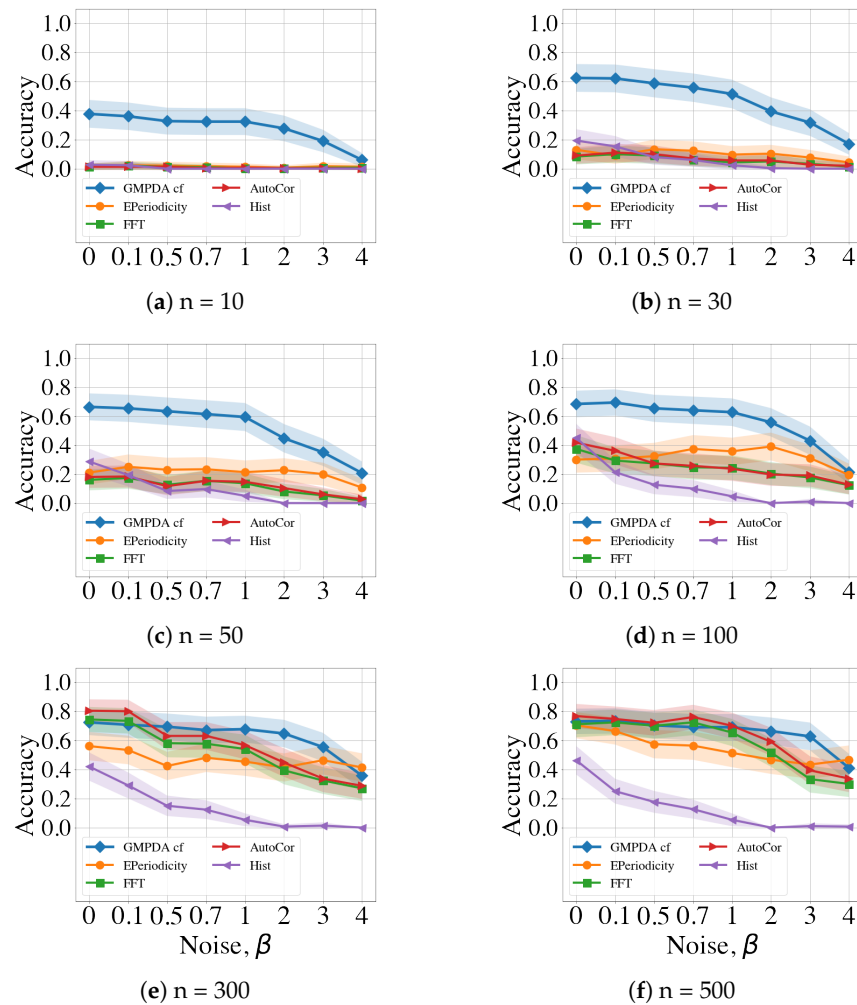
Here, we compare the performance of the GMPDA and of the alternative methods regarding an increase in noise. In the following figures, the accuracy of all the involved methods is plotted for  $|\mu| = 1$ , different number of events  $n$ , while it is averaged over all considered values of variance  $\sigma$ . The GMPDA with curve fitting consistently outperformed the alternative methods for different levels of noise for the Random Walk Model (Figure A3) for up to 500 events, and up to 100 events for the Clock Model (Figure A4).



**Figure A3.** Cont.



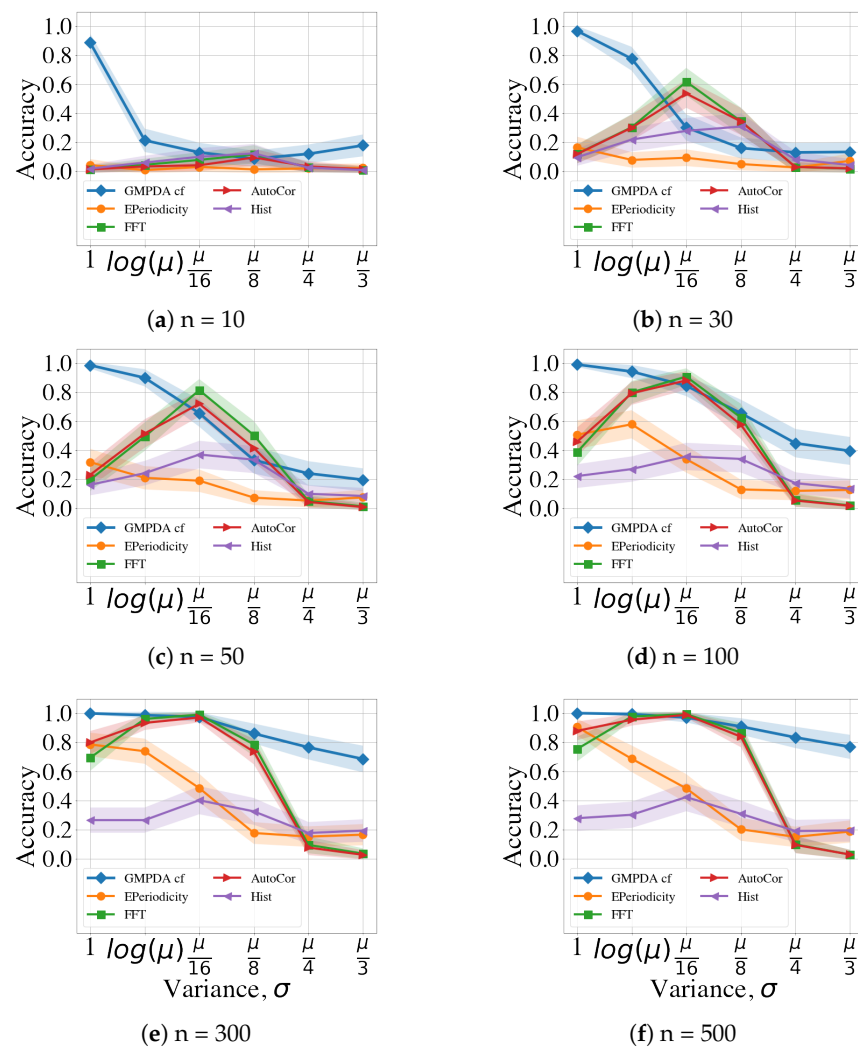
**Figure A3.** Random Walk Model performance with respect to  $\beta$ , averaged over  $\sigma$ ,  $|\mu| = 1$ .



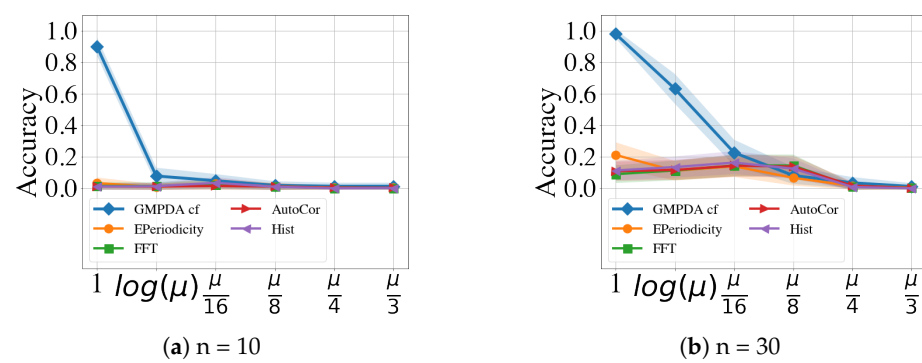
**Figure A4.** Clock Model performance with respect to  $\beta$ , averaged over  $\sigma$ ,  $|\mu| = 1$ .

### Appendix B.3. Comparison to Alternative Methods with Respect to Variance $\sigma$

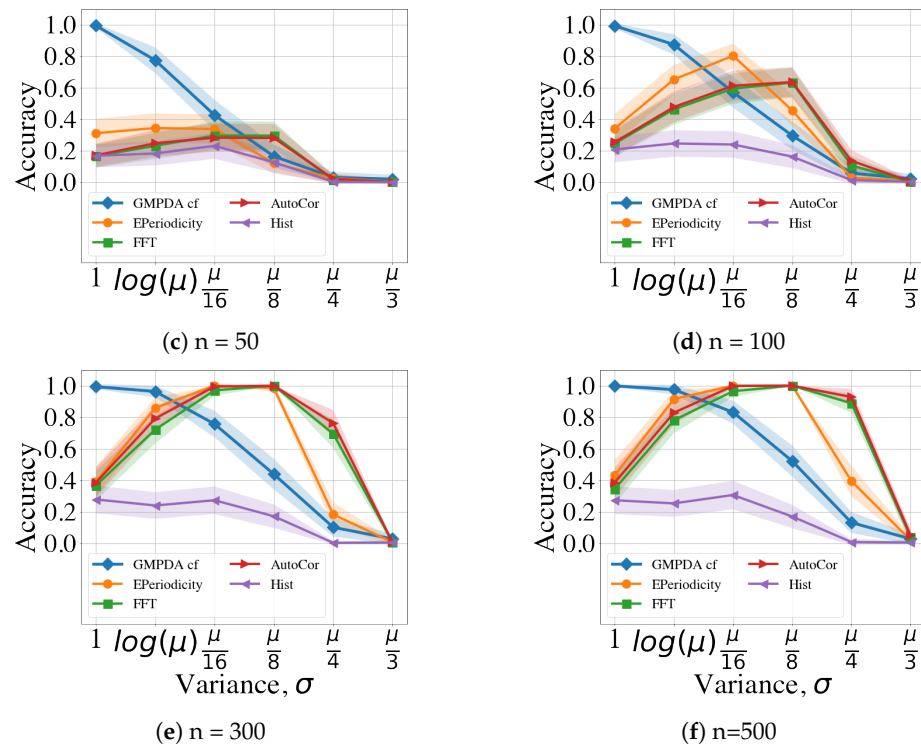
Figures A5 and A6 compare the performance of the GMPDA and the alternative methods for increasing levels of variance. In the following figures, the accuracy of all the involved methods is plotted for  $|\mu| = 1$  and different number of events  $n$ , while it is averaged over all considered values of noise  $\beta$ .



**Figure A5.** Random Walk Model performance with respect to  $\sigma$ , averaged over noise,  $|\mu| = 1$ .



**Figure A6.** Cont.



**Figure A6.** Clock Model performance with respect to  $\sigma$ , averaged over noise,  $|\mu| = 1$ .

#### Appendix B.4. Sensitivity Analysis

We summarize the differences in sensitivity to critical simulation parameters across the different algorithms in Table A1. Based on the simulation results obtained in Section 4, we used generalized linear mixed models with number of periods  $|\mu| = 1, 2, 3$  nested within trials ( $n = 38,400$ ), with the response being the accurate detection of a single periodicity (coded as 1 if the estimate is within an intervals around the true value  $\pm\sigma$ ) and the independent factors being as follows:

- Number of events  $n = 10, 30, 50, 100, 300, 500$ ;
- Number of periods  $|\mu| = 1, 2, 3$ ;
- Variance  $\sigma = 1, \log(\mu), \frac{\mu}{p}$ , with  $p = 3, 8, 16$ ;
- Noise  $\beta = 0, 0.1, 0.5, 0.7, 1, 2, 4, 8$ .

Mixed logistic models were computed separately for the Clock Model and the Random Walk Model and each algorithm. Table A1 lists the ANOVA type II sum of squares (SoS), i.e., the SoS of each main effect after the introduction of all other main effects. While the SoS are not directly comparable between models, their relative contribution is, and suggests that for the Random Walk Model, the number of events had a major effect in all algorithms except the FFT histogram algorithm. The number of periods had a small to moderate effect, except for the E-periodicity, where it did not play a role. Both E-periodicity and the GMPDA with curve fitting were very sensitive to the noise level, and across algorithms, variations in sigma had one of the strongest effects on accuracy, again with the exception of the E-periodicity algorithm.

The results for the Clock Model were largely similar with some notable exceptions. Compared to the Random Walk Models, the E-periodicity algorithm was considerably less sensitive to variations in noise but more sensitive to variations in variance. Overall, the three algorithms E-periodicity, FFT, and FFT autocorrelation showed a similar pattern, with the number of events having the strongest influence, sigma being the second strongest, and the noise and number of periods having only relatively minor effects. For the two GMPDAs, the strongest effect was seen for sigma.

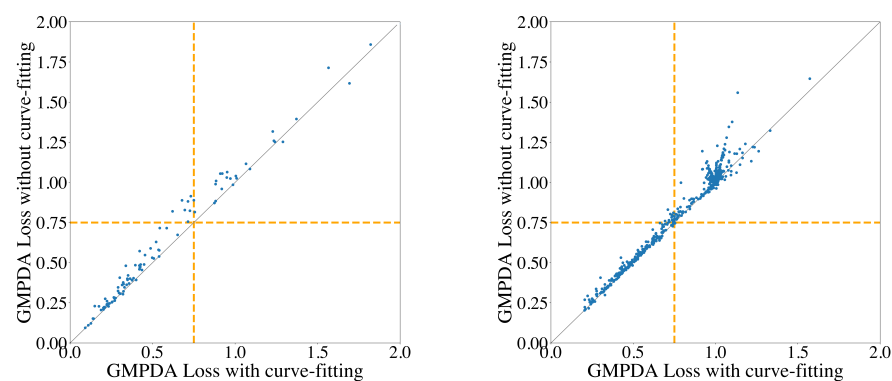
Across all models and algorithms, both sigma and the number events emerged as the strongest determinants of periodicity detection accuracy.

**Table A1.** Differences in sensitivity to critical simulation parameters across the different algorithms. The table lists the ANOVA type II sums, with larger numbers signifying larger sensitivity to the respective parameter.

	Number of Events	Number of Periods	Noise	Sigma Ratio
Df	5	2	7	5
<b>Random Walk Models</b>				
GMPDA with curve fitting	5110	3746	5096	4756
GMPDA w/o curve fitting	3349	1821	1291	7830
E-periodicity	3218	10	5221	438
FFT	5519	1420	1359	6687
FFT Autocorrelation	5506	1252	3058	6013
FFT Histogram	619	1159	958	2994
<b>Clock Models</b>				
GMPDA with curve fitting	2714	2377	3186	7399
GMPDA w/o curve fitting	2089	454	1210	6806
E-periodicity	7659	280	651	4707
FFT	8520	157	1522	5216
FFT Autocorrelation	8480	87	1283	5689
FFT Histogram	438	211	2829	1035

#### Appendix B.5. Real Application: Loss

This section shows the GMPDA loss obtained with and without curve fitting for the MROS data set. Figure A7 shows the loss for 100 recordings in the left panel and the loss comparison for all single bouts in the right panel.



**Figure A7.** Comparison of GMPA loss with and without curve fitting for individual nights (**left panel**) and single sleep bouts (**right panel**).

#### References

1. Welch, P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **1967**, *15*, 70–73. [[CrossRef](#)]
2. Priestley, M.B. *Spectral Analysis and Time Series: Univariate Series*; Academic Press: London, UK, 1981.
3. Madsen, H. *Time Series Analysis*; CRC Press: Boca Raton, FL, USA, 2007.
4. Mitsa, T. *Temporal Data Mining*; CRC Press: Boca Raton, FL, USA, 2010.
5. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*; John Wiley & Sons: New York, NY, USA, 2015.
6. Junier, I.; Hérissou, J.; Képès, F. Periodic pattern detection in sparse boolean sequences. *Algorithms Mol. Biol.* **2010**, *5*, 31. [[CrossRef](#)] [[PubMed](#)]
7. Glynn, E.F.; Chen, J.; Mushegian, A.R. Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. *Bioinformatics* **2006**, *22*, 310–316. [[CrossRef](#)] [[PubMed](#)]

8. Vlachos, M.; Yu, P.; Castelli, V. On Periodicity Detection and Structural Periodic Similarity. In Proceedings of the 2005 SIAM International Conference on Data Mining (SDM), Newport Beach, CA, USA, 21–23 April 2005; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2005; pp. 449–460. [\[CrossRef\]](#)
9. Ahdesmäki, M.; Lähdesmäki, H.; Gracey, A.; Shmulevich, L.; Yli-Harja, O. Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data. *BMC Bioinform.* **2007**, *8*, 233. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Berberidis, C.; Aref, W.G.; Atallah, M.; Vlahavas, I.; Elmagarmid, A.K. Multiple and Partial periodicity mining in time series databases. In Proceedings of the 15th European Conference on Artificial Intelligence, NLD, ECAI'02, Lyon, France, 21–26 July 2002; pp. 370–374.
11. Elfeky, M.G.; Aref, W.G.; Elmagarmid, A.K. STAGGER: Periodicity Mining of Data Streams Using Expanding Sliding Windows. In Proceedings of the Sixth International Conference on Data Mining (ICDM'06), Hong Kong, China, 18–22 December 2006; pp. 188–199. [\[CrossRef\]](#)
12. Li, Z.; Wang, J.; Han, J. ePeriodicity: Mining Event Periodicity from Incomplete Observations. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 1219–1232. [\[CrossRef\]](#)
13. Sheng, M.; Hellerstein, J. Mining partially periodic event patterns with unknown periods. In Proceedings of the 17th International Conference on Data Engineering, Heidelberg, Germany, 2–6 April 2001; pp. 205–214. [\[CrossRef\]](#)
14. Yang, K.J.; Hong, T.P.; Chen, Y.M.; Lan, G.C. Projection-based partial periodic pattern mining for event sequences. *Expert Syst. Appl.* **2013**, *40*, 4232–4240. [\[CrossRef\]](#)
15. Yuan, Q.; Shang, J.; Cao, X.; Zhang, C.; Geng, X.; Han, J. Detecting Multiple Periods and Periodic Patterns in Event Time Sequences. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 617–626. [\[CrossRef\]](#)
16. Van Dongen, H.; Olofsen, E.; Van Hartevelt, J.; Kruij, E. A Procedure of Multiple Period Searching in Unequally Spaced Time-Series with the Lomb–Scargle Method. *Biol. Rhythm. Res.* **1999**, *30*, 149–177. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Parthasarathy, S.; Mehta, S.; Srinivasan, S. Robust periodicity detection algorithms. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM '06, Arlington, VA, USA, 6–11 November 2006; pp. 874–875. [\[CrossRef\]](#)
18. Ghosh, A.; Lucas, C.; Sarkar, R. Finding Periodic Discrete Events in Noisy Streams. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 627–636. [\[CrossRef\]](#)
19. Grinstead, C.M.; Snell, J.L. *Introduction to Probability*, 2nd ed.; American Mathematical Society: Palo Alto, CA, USA, 2013.
20. Auger, F.; Flandrin, P. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Process.* **1995**, *43*, 1068–1089. [\[CrossRef\]](#)
21. Zhang, G.Q.; Cui, L.; Mueller, R.; Tao, S.; Kim, M.; Rueschman, M.; Mariani, S.; Mobley, D.; Redline, S. The National Sleep Research Resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1351–1358. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Dean, D.A.; Goldberger, A.L.; Mueller, R.; Kim, M.; Rueschman, M.; Mobley, D.; Sahoo, S.S.; Jayapandian, C.P.; Cui, L.; Morrical, M.G.; et al. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. *Sleep* **2016**, *39*, 1151–1164. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Blackwell, T.; Yaffe, K.; Ancoli-Israel, S.; Redline, S.; Ensrud, K.E.; Stefanick, M.L.; Laffan, A.; Stone, K.L.; Osteoporotic Fractures in Men Study Group. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: The Osteoporotic Fractures in Men Sleep Study. *J. Am. Geriatr. Soc.* **2011**, *59*, 2217–2225. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Blank, J.B.; Cawthon, P.M.; Carrion-Petersen, M.L.; Harper, L.; Johnson, J.P.; Mitson, E.; Delay, R.R. Overview of recruitment for the osteoporotic fractures in men study (MrOS). *Contemp. Clin. Trials* **2005**, *26*, 557–568. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Orwoll, E.; Blank, J.B.; Barrett-Connor, E.; Cauley, J.; Cummings, S.; Ensrud, K.; Lewis, C.; Cawthon, P.M.; Marcus, R.; Marshall, L.M.; et al. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study—a large observational study of the determinants of fracture in older men. *Contemp. Clin. Trials* **2005**, *26*, 569–585. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Haba-Rubio, J.; Marti-Soler, H.; Marques-Vidal, P.; Tobback, N.; Andries, D.; Preisig, M.; Waeber, G.; Vollenweider, P.; Kutalik, Z.; Tafti, M.; et al. Prevalence and determinants of periodic limb movements in the general population. *Ann. Neurol.* **2016**, *79*, 464–474. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Ferri, R.; Fulda, S.; Allen, R.P.; Zucconi, M.; Bruni, O.; Chokroverty, S.; Ferini-Strambi, L.; Frauscher, B.; Garcia-Borreguero, D.; Hirshkowitz, M.; et al. World Association of Sleep Medicine (WASM) 2016 standards for recording and scoring leg movements in polysomnograms developed by a joint task force from the International and the European Restless Legs Syndrome Study Groups (IRLSSG and EURLSSG). *Sleep Med.* **2016**, *26*, 86–95. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Ferri, R.; Koo, B.B.; Picchietti, D.L.; Fulda, S. Periodic leg movements during sleep: Phenotype, neurophysiology, and clinical significance. *Sleep Med.* **2017**, *31*, 29–38. [\[CrossRef\]](#) [\[PubMed\]](#)

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.