Explicit Link Between Periodic Covariance Functions and State Space Models

Arno Solin
arno.solin@aalto.fi
Department of Biomedical Engineering
and Computational Science
Aalto University

Simo Särkkä
simo.sarkka@aalto.fi
Department of Biomedical Engineering
and Computational Science
Aalto University

Abstract

This paper shows how periodic covariance functions in Gaussian process regression can be reformulated as state space models, which can be solved with classical Kalman filtering theory. This reduces the problematic cubic complexity of Gaussian process regression in the number of time steps into linear time complexity. The representation is based on expanding periodic covariance functions into a series of stochastic resonators. The explicit representation of the canonical periodic covariance function is written out and the expansion is shown to uniformly converge to the exact covariance function with a known convergence rate. The framework is generalized to quasi-periodic covariance functions by introducing damping terms in the system and applied to two sets of real data. The approach could be easily extended to non-stationary and spatio-temporal variants.

1 INTRODUCTION

In Bayesian non-parametric machine learning, Gaussian processes (GPs, [1]) are commonly used modeling tools. In GP regression the model functions are assumed to be realizations of a Gaussian process random prior with a given covariance function, into which the prior assumptions are encoded. One very commonly encountered phenomenon in applications is periodicity, and in GP regression this is incorporated through periodic covariance functions.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

Whilst being very flexible and convenient modeling tools, computationally the direct GP methodology does not conveniently lend itself to long (or unbounded) time series. This is due to the prohibitive $\mathcal{O}(n^3)$ scaling of computational cost, which grows unbearable when the number of training samples n grows large. Several general sparse approximation schemes have been proposed for this problem (see, *e.g.*, [2] for a review). For periodic latent force models [3] one option is to use a set of basis functions and estimate the model variances as a part of the state [4].

In the case of temporal models computational savings can be made by converting the GP into state space form and do inference using Kalman filtering methods [5]. This connection is well established (see [6, 7]), and recently it has gained a lot of interest. Certain classes of stationary covariance functions can be directly converted into state space models by representing their spectral densities as rational functions [8, 9]. This scheme is, however, not suitable for periodic covariances, where the spectrum is set up by Dirac delta peaks. Therefore this paper seeks an alternative way to approximate periodic covariance functions by introducing the connection to stochastic resonators.

Periodic structure in time series data can be modeled by second-order differential equations. More complex periodical variation can be accounted for by adding harmonics to the model, and quasi-periodic (almost periodic) behavior by extending the model to a stochastic differential equation (SDE) [10]. This formulation fits under Bayesian state space estimation and has been employed in [11, 12]. However, the theory linking stochastic resonators to GP models has been lacking, and constructing this theory enables direct conversion of periodic GP models into computationally efficient state space form.

The structure of this paper is as follows. In Section 2 the state space methodology for Gaussian process regression is briefly reviewed. In Section 3 a novel way

of approximating periodic and quasi-periodic covariance functions in state space form is introduced, and the accuracy and convergence of the approximation is analyzed. Section 4 contains experimental evaluation of the computational requirements and application of the methods to two real data sets.

2 METHODS

2.1 Gaussian Process Regression

GP regression is concerned with predicting an unknown scalar output $f(\mathbf{x}_*)$ associated with a known input $\mathbf{x}_* \in \mathbb{R}^d$, given a training data set $\mathcal{D} = \{(\mathbf{x}_k, y_k) \mid k = 1, 2, \ldots, n\}$. The model function f is assumed to be a realizations of a Gaussian random process prior and the observations corrupted by Gaussian noise

$$f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')) \tag{1}$$
$$y_k = f(\mathbf{x}_k) + \varepsilon_k, \tag{2}$$

where $\varepsilon_k \sim \mathcal{N}(0, \sigma_\text{n}^2)$. The direct solution to the GP regression problem gives predictions $p(f(\mathbf{x}_*) \mid \mathbf{x}_*, \mathcal{D}) = \mathcal{N}(\mathbb{E}[f(\mathbf{x}_*)], \mathbb{V}[f(\mathbf{x}_*)])$. This can be computed in closed-form as [1]

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}_*)] &= \mathbf{k}_*^\mathsf{T}(\mathbf{K} + \sigma_\text{n}^2\mathbf{I})^{-1}\mathbf{y}, \\ \mathbb{V}[f(\mathbf{x}_*)] &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\mathsf{T}(\mathbf{K} + \sigma_\text{n}^2\mathbf{I})^{-1}\mathbf{k}_*,\end{aligned} \tag{3}$$

where $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}_*$ is an n-dimensional vector with the ith entry being $k(\mathbf{x}_*, \mathbf{x}_i)$, and $\mathbf{y}$ is a vector of the n observations. The computational complexity comes from the $n \times n$ matrix inversion in (3).

A common way to learn the hyperparameters $\boldsymbol{\theta}$ of the covariance function ($k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta})$, suppressed earlier in the notation for brevity) and the noise variance σ_n^2 is by maximizing the marginal likelihood function using a suitable optimizer (see, *e.g.*, [1]).

2.2 Representing the GP as a Stochastic Differential Equation

For temporal GPs, instead of directly working with the kernel formalism of the Gaussian process $f(t)$, certain classes of covariance functions allow to work with the mathematical dual [9], where the Gaussian process is constructed as a solution to a mth order linear stochastic differential equation (SDE). The corresponding inference problem can be solved with Kalman filtering type of methods [13], where the computational complexity is $\mathcal{O}(m^3 n)$. If the number of observations $n \gg m$, as typically is the case in temporal modeling, this formulation is very beneficial.

The state space model corresponding to the GP regression problem (1) can be given as

$$\begin{aligned}\frac{\mathrm{d}\mathbf{f}(t)}{\mathrm{d}t} &= \mathbf{F}\mathbf{f}(t) + \mathbf{L}\mathbf{w}(t) \\ y_k &= \mathbf{H}\mathbf{f}(t_k) + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma_\text{n}^2),\end{aligned} \tag{4}$$

where $\mathbf{f}(t) = \big(f_1(t), f_2(t), \ldots, f_m(t)\big)^\mathsf{T}$ holds the m stochastic processes, and $\mathbf{w}(t)$ is a multi-dimensional white noise process with spectral density $\mathbf{Q}_\text{c}$. The model is defined by the feedback matrix $\mathbf{F}$ and the noise effect matrix $\mathbf{L}$.

The Gaussian process can be reconstructed by defining the observation matrix $\mathbf{H}$ such that $f(t) = \mathbf{H}\mathbf{f}(t)$. In this form, the spectral density $S(\omega)$ of $f(t)$ can be written using the state representation as

$$S(\omega) = \mathbf{H}(\mathbf{F} - \mathrm{i}\,\omega\mathbf{I})^{-1}\mathbf{L}\mathbf{Q}_\text{c}\mathbf{L}^\mathsf{T}\left[(\mathbf{F} + \mathrm{i}\,\omega\mathbf{I})^{-1}\right]^\mathsf{T}\mathbf{H}^\mathsf{T}. \tag{5}$$

In a stationary state, the covariance function of $f(t)$ is the inverse Fourier transform of its spectral density:

$$k(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)\exp(-\mathrm{i}\,\omega\tau)\,\mathrm{d}\omega. \tag{6}$$

This can be written with the help of the state space matrices:

$$k(\tau) = \begin{cases}\mathbf{H}\mathbf{P}_\infty\boldsymbol{\Phi}(\tau)^\mathsf{T}\mathbf{H}^\mathsf{T}, & \text{if } \tau \geq 0 \\ \mathbf{H}\boldsymbol{\Phi}(-\tau)\mathbf{P}_\infty\mathbf{H}^\mathsf{T}, & \text{if } \tau < 0,\end{cases} \tag{7}$$

where $\boldsymbol{\Phi}(\tau) = \exp(\mathbf{F}\tau)$ is the matrix exponential of the feedback matrix. $\mathbf{P}_\infty$ is the stationary covariance of $\mathbf{f}(t)$ that is the solution to the corresponding matrix Riccati equation:

$$\frac{\mathrm{d}\mathbf{P}_\infty}{\mathrm{d}t} = \mathbf{F}\mathbf{P}_\infty + \mathbf{P}_\infty\mathbf{F}^\mathsf{T} + \mathbf{L}\mathbf{Q}_\text{c}\mathbf{L}^\mathsf{T} = \mathbf{0}. \tag{8}$$

The continuous-time linear time-invariant model (4) can be solved for discrete points. This is the closed-form solution to the SDE at the specified time points, and it is given as

$$\mathbf{f}_{k+1} = \mathbf{A}_k\mathbf{f}_k + \mathbf{q}_k, \quad \mathbf{q}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_k), \tag{9}$$

where $\mathbf{f}(t_k) = \mathbf{f}_k$, and the state transition and process noise covariance matrices can be solved analytically (see, *e.g.*, [9]). They are given as

$$\mathbf{A}_k = \boldsymbol{\Phi}(\Delta t_k) \tag{10}$$

$$\mathbf{Q}_k = \int_0^{\Delta t_k} \boldsymbol{\Phi}(\Delta t_k - \tau)\mathbf{L}\mathbf{Q}_\text{c}\mathbf{L}^\mathsf{T}\boldsymbol{\Phi}(\Delta t_k - \tau)^\mathsf{T}\,\mathrm{d}\tau \tag{11}$$

where $\Delta t_k = t_{k+1} - t_k$. The inference problem is now directly solvable using Kalman filtering type of methods [13, 14]. The Kalman filtering scheme naturally lends itself to hyperparameter optimization, as the marginal likelihood comes out as a by-product of the filtering step. Analytic gradients for conjugate gradient optimization can also be calculated.

3 PERIODIC COVARIANCE FUNCTIONS

3.1 Periodic Covariance Functions Through Warping

Consider a stationary covariance function such that with $\mathbf{x} \in \mathbb{R}^d$

$$k(\mathbf{x}, \mathbf{x}') \triangleq k(\mathbf{x} - \mathbf{x}'), \tag{12}$$

where a one-argument notation of the stationary covariance is introduced. A general way of constructing non-stationary covariance functions is to introduce a non-linear mapping (or warping) $\boldsymbol{u}(t)$ of the input t and then use a stationary covariance function in the $\boldsymbol{u}$ space. Using this warping method [15], periodic stationary kernels can be constructed by setting $\mathbf{x}(t) = \boldsymbol{u}(t)$ for some periodic function $\boldsymbol{u}(t) : \mathbb{R} \to \mathbb{R}^d$ for which the resulting covariance

$$k(t, t') = k(\boldsymbol{u}(t) - \boldsymbol{u}(t')) \tag{13}$$

becomes stationary. A typical choice in GP context (see, *e.g.*, [1, 15]) is $\boldsymbol{u}(t) = \big(\sin(t), \cos(t)\big)^\mathsf{T}$, which has the property

$$\begin{aligned} &\|\boldsymbol{u}(t) - \boldsymbol{u}(t')\|^2 \\ &\quad = \big(\sin(t) - \sin(t')\big)^2 + \big(\cos(t) - \cos(t')\big)^2 \\ &\quad = 4 \sin^2\left(\frac{t - t'}{2}\right) \end{aligned} \tag{14}$$

and thus results in a stationary covariance function for isotropic $k(\cdot)$. An example of such periodic processes is shown in Figure 1.

In terms of the original GP, $f(\mathbf{x})$, the above means that the values of the GP are evaluated on a certain periodic curve $\mathbf{x}(t) = \boldsymbol{u}(t)$. An interesting question is now that under what conditions does $k(t, t')$ then become stationary. For all t the following must hold:

$$\begin{aligned} k(t + \tau, t) &= k(\boldsymbol{u}(t + \tau) - \boldsymbol{u}(t)) \\ &= k(\boldsymbol{u}(\tau) - \boldsymbol{u}(0)). \end{aligned}$$

Assume now that $k(\cdot)$ is invariant with respect to some parametrized set of invertible time-invariant linear transforms $\mathbf{T}(s)$:

$$\mathbf{x}^* = \mathbf{T}(s)\, \mathbf{x} \tag{15}$$

for some s. Consequently, this should always result in $\boldsymbol{u}(\tau) = \mathbf{T}(s)\boldsymbol{u}(t + \tau)$. This seems to imply that $\boldsymbol{u}(\tau)$ and $\mathbf{T}(\tau)$ can actually be identified by suitable selection of scaling for s. Thus $\mathbf{T}$ needs to be the transition matrix of $\boldsymbol{u}(t)$. In other words, $k(\cdot)$ needs to be invariant with respect to the transition matrix of $\boldsymbol{u}$. For isotropic functions, any orthogonal matrix will do (provided that it is periodic).

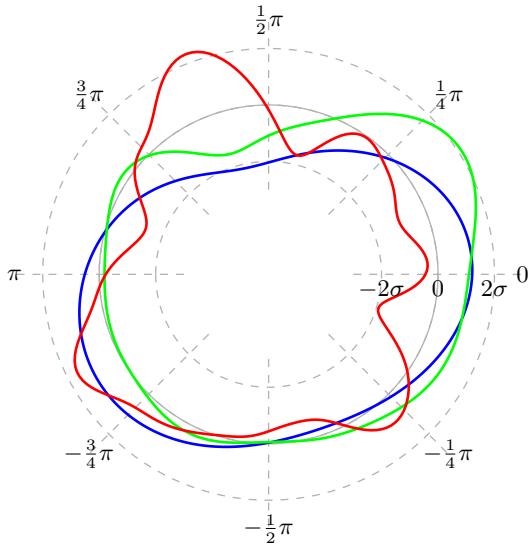


Figure 1: Random draws from a periodic GP prior with different length-scales: $\ell = 1$ (——), $\ell = 0.5$ (——), and $\ell = 0.25$ (——).

3.2 Writing Periodic Covariances in State Space Form

Let $k(t, t') = k(\boldsymbol{u}(t) - \boldsymbol{u}(t'))$ be a stationary, periodic, and valid covariance function set up by the procedure in the previous section. There exists a symmetric and periodic $k_\mathrm{p}(\tau)$ such that

$$k(t, t') = k_\mathrm{p}(t - t'). \tag{16}$$

As $k_\mathrm{p}(\cdot)$ is a periodic and continuous even function, $k_\mathrm{p}(\tau) = k_\mathrm{p}(-\tau)$, it can be expanded into a (almost everywhere) convergent Fourier series

$$k_\mathrm{p}(\tau) = \sum_{j=0}^{\infty} q_j^2 \cos(j\, \omega_0\, \tau), \tag{17}$$

where ω_0 is the angular frequency defining the periodicity. Formally its spectral density consists of delta function peaks at the characteristic frequencies, which can be written as

$$S_\mathrm{p}(\omega) = \sum_{j=0}^{\infty} q_j^2 \pi \left[\delta(\omega - j\, \omega_0) + \delta(\omega + j\, \omega_0)\right]. \tag{18}$$

As pointed out by Reece *et al.* [4], this spectral density does not conveniently fit under the framework by Särkkä *et al.* [9], where the spectral density was approximated by rational functions. Instead, this paper takes an alternative approach to come up with the explicit state space representation of the periodic covariance function in terms of resonator models.

Each periodic term j in the series (17) can be considered separately as a pair of processes stacked in $\mathbf{f}_j(t) =$

$\big(x_j(t), y_j(t)\big)^\mathsf{T}$ with initial conditions $\mathbf{f}_j(0) \sim \mathcal{N}(\mathbf{0}, q_j^2\mathbf{I})$ and satisfying the differential equations

$$\begin{cases} \dfrac{\mathrm{d}x_j(t)}{\mathrm{d}t} = -j\,\omega_0\,y_j(t) \\ \dfrac{\mathrm{d}y_j(t)}{\mathrm{d}t} = \quad j\,\omega_0\,x_j(t) \end{cases} \tag{19}$$

which define a harmonic oscillator. This second-order ordinary differential equation can be solved, and the processes are given in closed-form as

$$\mathbf{f}_j(t) = \begin{pmatrix} \cos(\omega_0\,j\,t) & -\sin(\omega_0\,j\,t) \\ \sin(\omega_0\,j\,t) & \cos(\omega_0\,j\,t) \end{pmatrix} \mathbf{f}_j(0). \tag{20}$$

These processes are random only in the sense that the initial state is drawn from a Gaussian. The trajectories themselves are deterministic. The covariance of x_j for $\tau > 0$ is given by

$$\begin{aligned} &\mathbb{E}[x_j(t)\,x_j(t+\tau)] \\ &= \mathbb{E}\big[(x_j(0)\cos(\omega_0\,j\,t) - y_j(0)\sin(\omega_0\,j\,t)) \\ &\qquad (x_j(0)\cos(\omega_0\,j\,(t+\tau)) - y_j(0)\sin(\omega_0\,j\,(t+\tau)))\big] \\ &= q_j^2\cos(\omega_0\,j\,\tau). \end{aligned}$$

Therefore the covariance function of the sum of statistically independent resonators, $\sum_{j=0}^{\infty} x_j(t)$, with $\big(x_j(0), y_j(0)\big)^\mathsf{T} \sim \mathcal{N}(\mathbf{0}, q_j^2\,\mathbf{I})$, is

$$k_\mathrm{p}(\tau) = \sum_{j=0}^{\infty} q_j^2\cos(j\,\omega_0\,\tau). \tag{21}$$

The question now remains how to determine q_j^2 from given $k(\cdot)$ and $\boldsymbol{u}(\cdot)$, or equivalently from a given covariance function $k_\mathrm{p}(\cdot)$. One way to determine the coefficients is via projection to the cosine basis

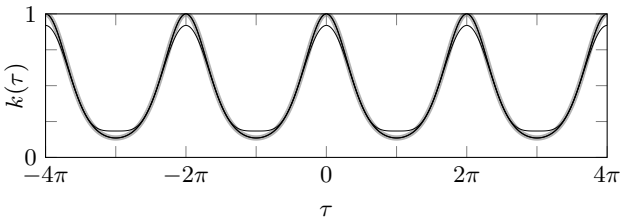
$$q_j^2 = \frac{\omega_0}{\pi}\int_{-\pi/\omega_0}^{\pi/\omega_0} k_\mathrm{p}(\tau)\cos(j\,\omega_0\,\tau)\,\mathrm{d}\tau, \tag{22}$$

for $j = 1, 2, \ldots$. However, the coefficients can be matched in other ways as well, as will be demonstrated in the next section.

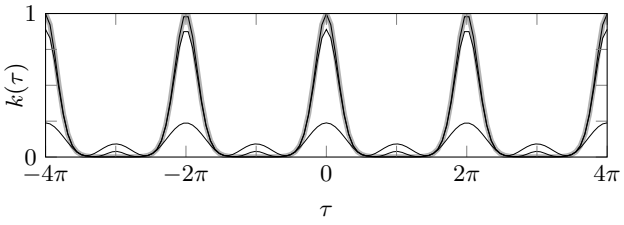
3.3 The Canonical Periodic Covariance Function in State Space Form

In machine learning, the most commonly encountered periodic covariance function (see, *e.g.*, [1, 15]) corresponds to the squared exponential, $k(\mathbf{x}, \mathbf{x}') = \sigma^2\exp(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\ell^2))$, covariance function in $\boldsymbol{u}$-space as given by (14). In this paper, this covariance is referred to as the *canonical periodic covariance function*:

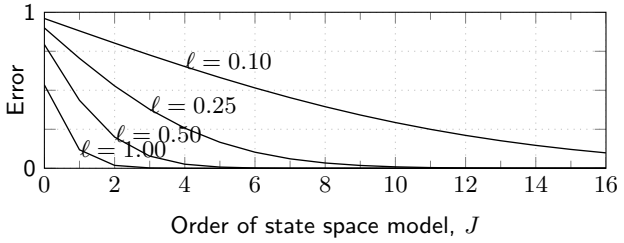
$$k_\mathrm{p}(t, t') = \sigma^2\exp\left(-\frac{2\sin^2\left(\omega_0\frac{t-t'}{2}\right)}{\ell^2}\right), \tag{23}$$



(a) Periodic covariance function with $\ell = 1$.



(b) Peridic covariance function with $\ell = 0.5$.



(c) Upper bound for approximation error.

Figure 2: Approximations to the canonical periodic covariance function with two length-scales. The degree of approximation is $J = 2, 6, 10$, growing with the line thickness. The grey line (——) represents the exact covariance.

where, without loss of generality, the magnitude scale σ^2 and frequency scale ω_0 parameters are assigned unit values to simplify the expressions that will follow. The scaling can always be restored by replacing τ with $\omega_0\,\tau$ and multiplying the spectral density coefficients with σ^2. The parameter ℓ defines the characteristic length-scale of the covariance. Figure 1 shows three random draws on the unit circle from the canonical periodic GP prior (23) with different length-scales.

In order to convert the covariance into state space form, the exponential expression in (23) can be decomposed by the identity $2\sin^2(\tau/2) = 1 - \cos(\tau)$. Writing out the Taylor series expansion of the exponential function gives:

$$k_\mathrm{p}(\tau) = \exp(-\ell^{-2})\sum_{j=0}^{\infty}\frac{1}{j!}\cos^j(\tau). \tag{24}$$

Now consider truncating this series at J, and recall that the powers of cosine can be expressed as sums of cosines with multiplied angles. Similarly, each power of cosine can be rewritten as a sum of first-order cosine

terms with multiplied angles. Expanding the powers of cosine and applying reduction formulas lead to the expression that is of the form given by (21) meeting the requirements of the previous subsection:

$$k_{\mathrm{p},J}(\tau) = \sum_{j=0}^{J} \tilde{q}^2_{j,J} \cos(j\,\tau), \tag{25}$$

where the coefficients for $\cos(j\,\tau)$ are given by

$$\tilde{q}^2_{j,J} = \frac{2}{\exp(\ell^{-2})} \sum_{i=0}^{\lfloor \frac{J-j}{2} \rfloor} \frac{(2\ell^2)^{-j-2i}}{(j+i)!\,i!}, \tag{26}$$

where $j = 1, 2, \ldots, J$ and $\lfloor \cdot \rfloor$ denotes the floor round-off operator. $\tilde{q}^2_{0,J}$ obeys the above formula, but is divided by 2. These coefficients always return a valid covariance function. Note that each $\tilde{q}^2_{j,J}$ depends on the chosen truncation index J. These coefficients ensure that Equation (25) is always a valid covariance function, as the terms are coupled in growth by J.

If the requirement of a valid covariance function is relaxed and only an optimal series approximation is required, taking the limit $J \to \infty$ in the sub-sums (26) gives the following spectral densities (or variances coefficients)

$$q^2_j = \frac{2\,\mathrm{I}_j(\ell^{-2})}{\exp(\ell^{-2})}, \text{ for } j = 1, 2, \ldots, J, \tag{27}$$

and $q^2_0 = \mathrm{I}_0(\ell^{-2})/\exp(\ell^{-2})$, where $\mathrm{I}_\alpha(z)$ is the modified Bessel function [16] of the first kind of order α. This is also the solution to the corresponding integral in (22). Note that the terms in (26) are bounded from above such that $\tilde{q}^2_{j,J} < q^2_j$ for any J.

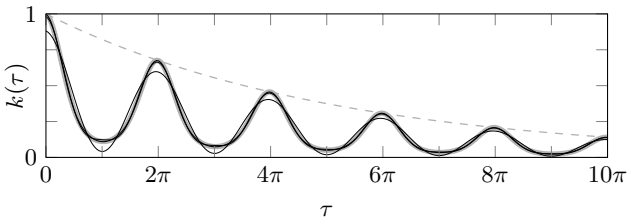
In the notation of Section 2 the corresponding state space model is now: $\mathbf{F}$, $\mathbf{L}$, and $\mathbf{P}_\infty$ are block-diagonal matrices, where block $j = 0, 1, \ldots, J$ is set up by the statistically independent feedback matrices

$$\mathbf{F}^{\mathrm{p}}_j = \begin{pmatrix} 0 & -\omega_0\, j \\ \omega_0\, j & 0 \end{pmatrix}, \tag{28}$$

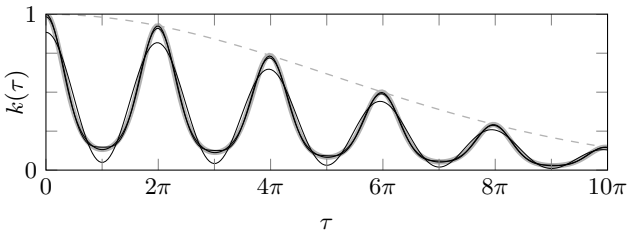
noise effect matrices $\mathbf{L}^{\mathrm{p}}_j = \mathbf{I}_2$, and stationary covariances $\mathbf{P}^{\mathrm{p}}_{\infty,j} = q^2_j\,\mathbf{I}_2$, respectively. Because the process does not have a diffusion term, $\mathbf{Q}^{\mathrm{p}}_{\mathrm{c}} = \mathbf{0}$, but the noise effect matrix is written out for compatibility. The measurement model matrix $\mathbf{H}$ is a block-row vector of $\mathbf{H}^{\mathrm{p}}_j = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

3.4 Approximation Error

Recall that the approximation in (25) is the result of a Jth order truncation of the Taylor series representation (24) of the exponential at origin. By Taylor's theorem the residual error at x is given by



(a) Quasi-periodic covariance function with damping smoothness $\nu = \frac{1}{2}$ (exponential covariance function).



(b) Quasi-periodic covariance function with damping smoothness $\nu = \infty$ (squared exponential).

Figure 3: Approximations to the quasi-periodic covariance function with two length-scales. The order of the model is $J = 1, 2, 3$, growing with the line thickness. The grey line (—) represents the exact covariance.

$R_J(x) = \frac{1}{(J+1)!} \exp(z)\, x^{J+1}$, for some $z \in [0, x]$. Thus the residual error in (25) can be written as $\tilde{R}_J(\tau) = \exp(-\ell^{-2}) \frac{1}{(J+1)!} \exp(\cos(\tau)) \cos^{J+1}(\tau)$ for some $\tau \in [0, 2\pi]$. Because $|\cos(\tau)| \leq 1$, the error bound is $|\tilde{R}_J(\tau)| \leq \frac{1}{(J+1)!} \exp(1-\ell^{-2})$, which thus also shows that $k_{\mathrm{p},J}(\tau) \to k_{\mathrm{p}}(\tau)$ uniformly, when $J \to \infty$. It is also easy to show that the series expansion obtained by replacing the terms $\tilde{q}^2_{j,J}$ with q^2_j in (25) converges to $k_{\mathrm{p}}(\tau)$ uniformly, when $J \to \infty$.

Figures 2a and 2b show the canonical periodic covariance function (23) for two different length-scales ℓ (in grey). The black lines correspond to approximations defined by the state space model truncated at different values of J. It is apparent that smaller length-scales correspond to rougher processes, and thus to longer tails in the spectrum.

Taking the analysis a step further gives an upper bound for the approximation error. Using recurrence relations of $\mathrm{I}_\alpha(z)$ (see [17]), $\sum_{j=0}^{\infty} \mathrm{I}_j(z) = [\exp(z) + \mathrm{I}_0(z)]/2$, and thus $\sum_{j=0}^{\infty} q^2_j = 1$. Because $|\cos(\omega_0\, j\, \tau)| \leq 1$, a rough upper bound for the truncation error can be given as $\epsilon(J) = 1 - \sum_{j=0}^{J} q^2_j$. This is visualized for various length-scales in Figure 2c.

3.5 Quasi-Periodic Covariance Functions

It is often desirable to allow for seasonable periodic variation, allowing the shape of the periodic effect to

change over time. This is known as *quasi-periodicity*. New covariance functions can be constructed as products of existing covariances. If $k_{\mathrm{p}}(\mathbf{x}, \mathbf{x}')$ and $k_{\mathrm{q}}(\mathbf{x}, \mathbf{x}')$ are both covariance functions of the same space, then so is $k(\mathbf{x}, \mathbf{x}') = k_{\mathrm{p}}(\mathbf{x}, \mathbf{x}')\, k_{\mathrm{q}}(\mathbf{x}, \mathbf{x}')$. A common way (see, *e.g.*, [1]) of constructing quasi-periodic covariances is to take the product of a periodic covariance function $k_{\mathrm{p}}(\tau)$ with a covariance function $k_{\mathrm{q}}(\tau)$ with rather long characteristic length-scale, allowing the covariance to decay away from exact periodicity.

Even though this is very straight-forward under the classical GP scheme, this is not trivial when using the state space form. In the state space model, the state transition matrix needs to factorize such that

$$\begin{aligned} \mathbf{A}_k &= \exp(\mathbf{F}^{\mathrm{p}} \Delta t_k) \exp(\mathbf{F}^{\mathrm{q}} \Delta t_k) \\ &= \exp\left((\mathbf{F}^{\mathrm{p}} + \mathbf{F}^{\mathrm{q}}) \Delta t_k\right), \end{aligned} \tag{29}$$

where $\mathbf{F}^{\mathrm{p}}$ is the feedback matrix corresponding to covariance function $k_{\mathrm{p}}(\tau)$ and $\mathbf{F}^{\mathrm{q}}$ the matrix corresponding to $k_{\mathrm{q}}(\tau)$. This is not true in general, and in order to factorize as above, the feedback matrices need to commute ($\mathbf{F}^{\mathrm{p}}\mathbf{F}^{\mathrm{q}} = \mathbf{F}^{\mathrm{q}}\mathbf{F}^{\mathrm{p}}$). This also ensures that the matrices preserve each others eigenspaces.

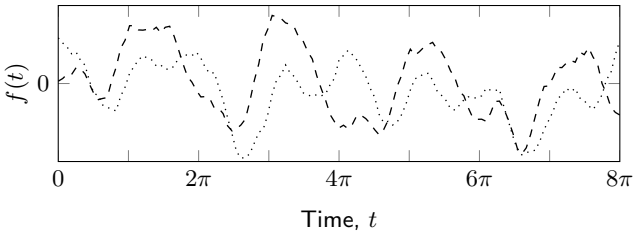
Consider the second covariance function to be of the Matérn class (*e.g.* [1]), which is a class of stationary isotropic covariance functions that are widely used in many applications and their parameters have understandable interpretations. A Matérn covariance function can be expressed as:

$$k(\tau) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\, \frac{\tau}{\ell}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\, \frac{\tau}{\ell}\right), \tag{30}$$

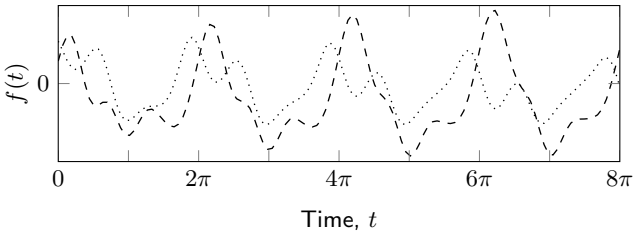
where $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is the modified Bessel function of the second kind [16]. The covariance function is characterized by three parameters: a smoothness parameter ν, a distance scale parameter ℓ, and a strength (magnitude) parameter σ^2, all of which are positive.

For example, if $\nu = 1/2$, this gives the exponential (Ornestein–Uhlenbeck) covariance function, $k_{\exp}(\tau) = \exp(-\lambda|\tau|)$, where $\lambda = 1/\ell$. The dashed line in Figure 3a shows the function values as $\ell = 16$. As explained in [8] it is straight-forward to form the corresponding state space model for this function, and in the notation of Section 2 this model is $F = -\lambda$, $L = 1$, and $Q_c = 2\lambda$. This is not an approximation, but the exact representation of the process in state space form.

However, the model does not commute with the periodic state space model defined by (28). But writing the Ornstein–Uhlenbeck process in terms of two separate stochastic processes gives the following state space presentation of the product between the periodic and



(a) With exponential damping ($\nu = 1/2$).



(b) With squared exponential damping ($\nu \to \infty$).

Figure 4: Random draws from quasi-periodic GP priors with different damping covariance functions, $k_{\mathrm{q}}(\tau)$, of the Matérn class.

the exponential covariance functions:

$$\mathbf{F}_j = \begin{pmatrix} 0 & -\omega_0\, j \\ \omega_0\, j & 0 \end{pmatrix} + \begin{pmatrix} -\lambda & 0 \\ 0 & -\lambda \end{pmatrix} = \begin{pmatrix} -\lambda & -\omega_0\, j \\ \omega_0\, j & -\lambda \end{pmatrix} \tag{31}$$

and $\mathbf{L}_j = \mathbf{I}_2$ and $\mathbf{Q}_{\mathrm{c},j} = 2\lambda q_j^2\, \mathbf{I}_2$, $\mathbf{P}_{\infty,j} = q_j^2\, \mathbf{I}_2$. Note that when $\ell \to \infty$, this reverts back to the fully periodic model. The original quasi-periodic covariance function and the resulting state space approximation of it is visualized in Figure 3a. The dashed line shows the exponential covariance function, and the black lines the approximation to the quasi-periodic covariance (in tick grey).

Generalizing the approach, consider the periodic state space model to be represented by the matrices $\mathbf{F}_j^{\mathrm{p}}$, $\mathbf{P}_{\infty,j}^{\mathrm{p}}$, $\mathbf{H}_j^{\mathrm{p}}$, and the q-dimensional model matrices corresponding to the second covariance function (*e.g.* of the Matérn class) to be denoted $\mathbf{F}^{\mathrm{q}}$, $\mathbf{L}^{\mathrm{q}}$, $\mathbf{Q}_{\mathrm{c}}^{\mathrm{q}}$, $\mathbf{P}_{\infty}^{\mathrm{q}}$. The joint model corresponding to the quasi-periodic product of the two covariance functions can then be given in the block-form similar to Section 3.3:

$$\begin{aligned} \mathbf{F}_j &= \mathbf{F}^{\mathrm{q}} \otimes \mathbf{I}_2 + \mathbf{I}_q \otimes \mathbf{F}_j^{\mathrm{p}}, \\ \mathbf{L}_j &= \mathbf{L}^{\mathrm{q}} \otimes \mathbf{L}_j^{\mathrm{p}}, \\ \mathbf{Q}_{\mathrm{c},j} &= \mathbf{Q}_{\mathrm{c}}^{\mathrm{q}} \otimes q_j^2\, \mathbf{I}_2, \\ \mathbf{P}_{\infty,j} &= \mathbf{P}_{\infty}^{\mathrm{q}} \otimes \mathbf{P}_{\infty,j}^{\mathrm{p}}, \\ \mathbf{H}_j &= \mathbf{H}^{\mathrm{q}} \otimes \mathbf{H}_j^{\mathrm{p}}, \end{aligned} \tag{32}$$

where '$\otimes$' denotes the Kronecker product of two matrices. The way of setting up $\mathbf{F}_j$ is also known as the Kronecker sum of matrices $\mathbf{F}^{\mathrm{q}}$ and $\mathbf{F}_j^{\mathrm{p}}$, which makes the matrix exponential factor as the Kronecker product of the corresponding two matrix exponentials [18].

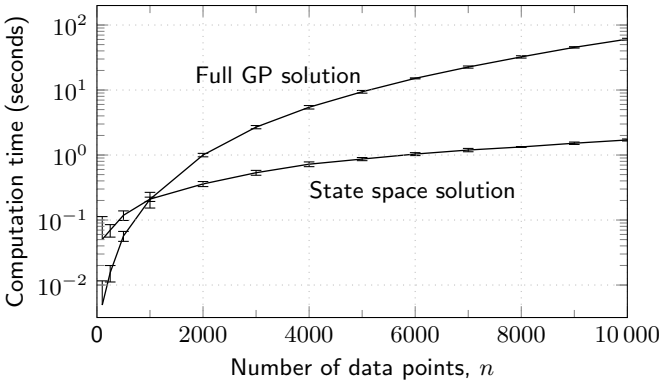


Figure 5: Demonstration of the computational benefits of the state space model in solving a GP regression problem for a number of data points up to 10 000 and with ten repetitions. The state space model execution times grow exactly linearly.

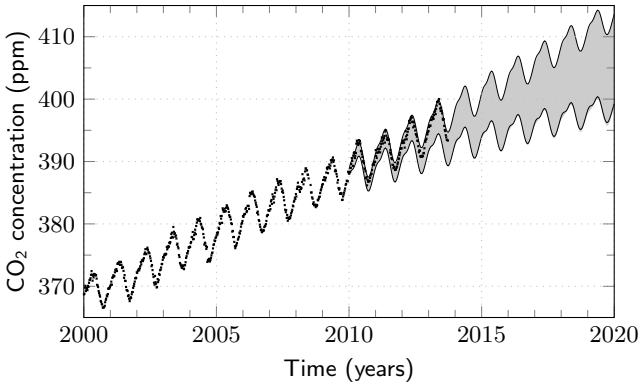


Figure 6: CO_2 concentration observations ($n = 2227$, values for years 1958–2000 not shown in figure) together with the 95% predictive confidence region (the shaded patch is from the state space model, and the thin lines from the exact GP solution).

The approximation still converges uniformly as long as $k_{\mathrm{q}}(0) < \infty$ and the approximation for $k_{\mathrm{p}}(\tau)$ converges uniformly. The matrices (32) result in a very sparse model, and sparse matrix methods can be employed for the matrix exponentials and multiplication.

Figure 3b shows the quasi-periodic covariance function corresponding to squared exponential damping ($\nu \to \infty$) of the periodicity (the squared exponential covariance is represented by the dashed line), and Figure 4b shows draws from the corresponding prior.

4 EXPERIMENTAL RESULTS

In this section the computational efficiency of the method is first demonstrated by applying it to simulated data, after which two empirical sets of data are used to show that the method is feasible in real-world applications.

4.1 Demonstrating the Computational Efficiency

To illustrate the efficiency of the proposed model, let $f(t)$ be a Gaussian process simulated from a GP prior with a periodic covariance function with unit parameters. The state space solution is benchmarked against a naive GP implementation in Mathworks Matlab (implemented as in [1] using the Cholesky decomposition).

Figure 5 shows the results for simulated GP regression problems with the number of observations ranging up to $n = 10\,000$ and with ten repetitions each. The periodic model was truncated at $J = 6$, and yet the worst case root-mean-square error was $\sim 10^{-3}$. As stated in Section 2 the computational complexity truly scales linearly with respect to the number of observations.

4.2 Modeling Carbon Dioxide Concentration

In this section the method is applied to the well-known time series data[1] consisting of atmospheric CO_2 concentration readings in parts per million (ppm) by volume from air samples collected at the Mauna Loa observatory, Hawaii (see, *e.g.*, [1]). The observations are monthly from 1958 to May 1974, after which the observations are weekly, resulting in 2227 measurements altogether. Data collected after year 2010 were retained for validation.

In practical GP modeling problems it is common to combine several simple covariance functions in order to come up with a model structure that meets the requirements of the phenomenon. The following rather simplified model is considered for the covariance:

$$k(\tau) = k_1(\tau) + k_2(\tau)\, k_3(\tau) + k_4(\tau), \qquad (33)$$

where $k_1(\cdot)$ is a squared exponential covariance function for the slow rising trend (hyperparameters σ_1^2, ℓ_1), $k_2(\cdot)$ the canonical periodic covariance function with a period of one year (hyperparameters σ_2^2, ℓ_2), $k_3(\cdot)$ is a covariance function of the Matérn class with $\nu = 3/2$ (hyperparameter ℓ_3), and $k_4(\cdot)$ is a covariance function of the Matérn class with $\nu = 3/2$ (hyperparameters σ_4^2, ℓ_4). The observations are assumed to be corrupted by Gaussian noise with variance σ_{n}^2.

Maximizing the marginal likelihood (quasi-Newton BFGS) with respect to the hyperparameters and predicting 20 years forward gives the results that are shown in Figure 6. The predictive 95% confidence region may be compared to the solid line representing

[1]Data available from `ftp://ftp.cmdl.noaa.gov/ccg/co2/trends/`.

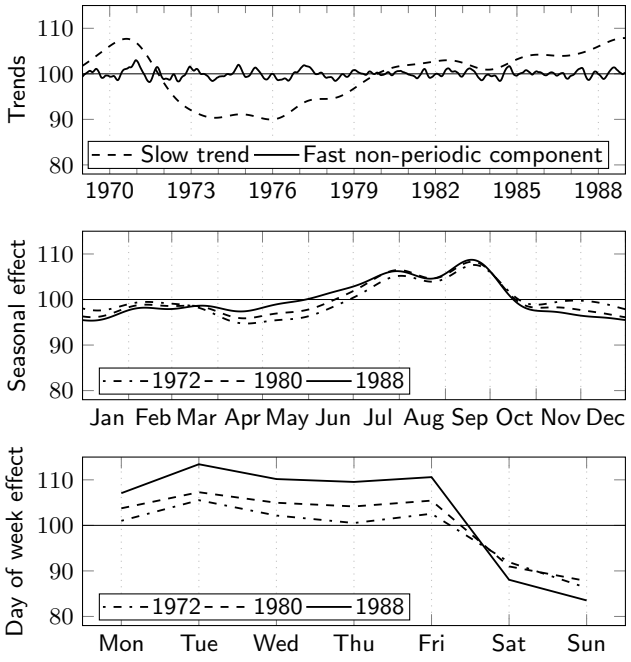


Figure 7: Relative number of births in the US based on daily data between 1969–1988 ($n = 7305$). The first plot shows the non-periodic long-term effects, the two latter the quasi-periodic seasonal and weekly effects.

the region corresponding to the full GP solution. The approximation error is negligible even though both the squared exponential and the periodic covariance function were approximated only by degree $J = 6$.

4.3 Modeling Birth Frequencies

Gaussian processes can ultimately be employed as components in a larger model. As demonstrated in [19], analysis of birthday frequencies can be done by considering structural knowledge of population growth and temporal patterns implied by the calendar weeks and years. The data in this example consist of the number of deliveries in the US during the years 1969–1988 (observed daily, $n = 7305$). The data was provided by the US National Vital Statistics System, available from Google BigQuery and pre-processed by Chris Mulligan[2].

Consider the following additive model with four components: a Matérn ($\nu = 5/2$, with hyperparameters σ_1^2, ℓ_1) GP prior for a smooth slow trend, a Matérn ($\nu = 3/2$, with hyperparameters σ_2^2, ℓ_2) prior for the fast non-periodic component, a quasi-periodic covariance function with a period of about one year (365.25 days, $J = 6$, hyperparameters σ_3^2, ℓ_3) and Matérn ($\nu = 3/2$, hyperparameter ℓ_4) damping, and a quasi-periodic covariance function with a period of one week ($J = 6$, hyperparameters σ_5^2, ℓ_5) and Matérn ($\nu = 3/2$, hyperparameter ℓ_6) damping. This is similar to [19], but special days are not considered separately. The observations are assumed to be corrupted by Gaussian noise with variance σ_n^2.

Optimizing (quasi-Newton BFGS) the marginal likelihood with respect to all the 11 hyperparameters gives the results that are shown in Figure 7. All the plots have been scaled in the same way to show differences relative to a baseline of 100. The first subfigure shows the slow trend over the 20-year period and the faster non-periodic component. The two remaining subfigures visualize the periodic yearly and weekly effects for years 1972, 1980, and 1988. The day of week and seasonal effects are clearly quasi-periodic; the rising number of induced births and selective C-sections has affected the day of week effect. The results agree with those of [19], and this can be regarded a successful example of a beneficial reformulation of a GP model in terms of sequential inference.

5 CONCLUSION

This paper has established the explicit connection between periodic covariance functions and stochastic differential equations. This link enables the use of efficient sequential inference methods to solve periodic GP regression problems in $\mathcal{O}(n)$ time complexity.

This reformulation is a 'best of both worlds' approach; it brings together the convenient model specification and hyperparametrization of GPs with the computational efficiency of state space models. As shown in Section 3.4, the approximation converges uniformly and a rough upper bound for the error can be given in closed-form. As demonstrated in the examples, the computational benefits can be accomplished with practically no loss of accuracy. Several extensions could be considered: It is possible to consider time-dependent frequencies (non-stationarity) as was done in [11] for the resonator model. Spatio-temporal extensions could be formulated following [20].

The codes for running the examples in this paper are available on the author's web page: `http://becs.aalto.fi/~asolin/`.

ACKNOWLEDGEMENTS

The authors wish to thank Aki Vehtari for help with the data analysis, and Tomi Peltola, Ville Tolvanen, and Alan Saul for comments on the manuscript. This work was supported by grants from the Academy of Finland (266940, 273475) and the Finnish Funding Agency for Technology and Innovation (40304/09).

[2]Data available from `http://chmullig.com/wp-content/uploads/2012/06/births.csv`.

REFERENCES

[1] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[2] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[3] Mauricio A. Álvarez, David Luengo, and Neil D. Lawrence. Linear latent force models using Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2693–2705, 2013.

[4] Steven Reece, Stephen Roberts, Siddhartha Ghosh, Alex Rogers, and Nicholas Jennings. Efficient state-space inference of periodic latent force models, 2013. arXiv:1310.6319v1 [stat.ML].

[5] Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82:35–45, 1960.

[6] Anthony O'Hagan. Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society. Series B*, 40(1):1–42, 1978.

[7] Yaakov Bar-Shalom, X. Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. John Wiley & Sons, 2004.

[8] Jouni Hartikainen and Simo Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010.

[9] Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

[10] Kevin Burrage, Ian Lenane, and Grant Lythe. Numerical methods for second-order stochastic differential equations. *SIAM Journal on Scientific Computing*, 29(1):245–264, 2008.

[11] Simo Särkkä, Arno Solin, Aapo Nummenmaa, Aki Vehtari, Toni Auranen, Simo Vanni, and Fa-Hsuan Lin. Dynamical retrospective filtering of physiological noise in BOLD fMRI: DRIFTER. *NeuroImage*, 60(2):1517–1527, 2012.

[12] Jouni Hartikainen, Mari Seppänen, and Simo Särkkä. State-space inference for non-linear latent force models with application to satellite orbit prediction. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 903–910, 2012.

[13] Mohinder S. Grewal and Angus P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley-Interscience, second edition, 2001.

[14] Simo Särkkä. *Bayesian Filtering and Smoothing*, volume 3 of *Institute of Mathematical Statistics Textbooks*. Cambridge University Press, 2013.

[15] David J.C. MacKay. Introduction to Gaussian processes. In Christopher M. Bishop, editor, *Neural Networks and Machine Learning*, volume 168 of *NATO ASI Series F Computer and Systems Sciences*, pages 133–166. Springer Verlag, 1998.

[16] Milton Abramowitz and Irene Stegun. *Handbook of Mathematical Functions*. Dover Publishing, New York, 1970.

[17] Frank W.J. Olver. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

[18] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

[19] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Press, third edition, 2013.

[20] Arno Solin and Simo Särkkä. Infinite-dimensional Bayesian filtering for detection of quasiperiodic phenomena in spatiotemporal data. *Physical Review E*, 88:052909, 2013.