Analytic Long-Term Forecasting with Periodic Gaussian Processes

Nooshin Haji Ghassemi
School of Computing
Blekinge Institute of Technology
Sweden

Marc Peter Deisenroth
Department of Computing
Imperial College London
United Kingdom

Department of Computer Science
TU Darmstadt, Germany

Abstract

Gaussian processes are a state-of-the-art method for learning models from data. Data with an underlying periodic structure appears in many areas, e.g., in climatology or robotics. It is often important to predict the long-term evolution of such a time series, and to take the inherent periodicity explicitly into account. In a Gaussian process, periodicity can be accounted for by an appropriate kernel choice. However, the standard periodic kernel does not allow for analytic long-term forecasting. To address this shortcoming, we re-parametrize the periodic kernel, which, in combination with a double approximation, allows for analytic long-term forecasting of a periodic state evolution with Gaussian processes. Our model allows for probabilistic long-term forecasting of periodic processes, which can be valuable in Bayesian decision making, optimal control, reinforcement learning, and robotics.

1 Introduction

Modeling, prediction, and decision making play an important role not only in machine learning, but also in other disciplines, such as control, signal processing, or climatology. Periodic or quasi-periodic behaviors appear almost everywhere, e.g., in robotics, the joint angle of a rotating robotic arm naturally follows a periodic pattern. Periodic time series appear in climate science, where temperature and CO_2 emissions follow quasi-periodic patterns. Moreover, (seasonal) rainfall, famines, sleeping patterns, or traffic congestion possess periodic trends. To make informed decisions, it is often necessary to forecast the system's evolution a long time ahead. With good models that account for the inherent periodicity of the data we can make well-informed long-term predictions and, thus, decisions.

In the context of regression, non-parametric Gaussian processes (GPs) (O'Hagan, 1978; Neal, 1997; Williams and Rasmussen, 1996) are the state-of-the-art method since they allow for flexible modeling while expressing uncertainties in a consistent way. Assumptions regarding the system's characteristics, e.g., smoothness or periodicity, can be encoded explicitly into the kernel of the GP (Rasmussen and Williams, 2006). Periodic GPs are used by Durrande et al. (2013) to detect periodically expressed genes or by Reece and Roberts (2010) in the context of target tracking.

Multiple-step ahead predictions, i.e., long-term forecasts with GPs, require approximations for nonlinear kernel functions, such as the Gaussian kernel. These approximations can be based either on stochastic sampling or on analytic closed-form computations, such as linearization (Girard et al., 2003) or moment matching (Quiñonero-Candela et al., 2003; Deisenroth et al., 2014), which approximate the multiple-step ahead predictive distribution by a Gaussian. Monte Carlo methods are straightforward to implement and flexible, but they can be computationally prohibitive in high dimensions. Although long-term forecasting and uncertainty propagation in terms of moment matching can be done with the Gaussian kernel, it is analytically intractable with the common periodic kernel.

In this paper, we propose a "double approximation", which allows for analytic long-term predictions of periodic patterns with GPs. The key idea is to exploit an equivalent representation of the standard stationary periodic kernel, which is based on a trigonomet-

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

ric transformation $\boldsymbol{u}$ of the original inputs $\boldsymbol{x}$. This re-parametrization allows us to analytically propagate Gaussian distributions $p(\boldsymbol{x}_t)$ through the GP to approximate the moments of the predictive distribution $p(\boldsymbol{x}_{t+1})$. This process requires two analytic approximations: First, we require a Gaussian approximation $p(\boldsymbol{u}_t)$ of the trigonometrically transformed input $p(\boldsymbol{x}_t)$. Second, the predictive distribution $p(\boldsymbol{x}_{t+1})$ is approximated by a Gaussian.

The paper is organized as follows: In Section 2, we provide an overview of Gaussian processes and kernels. In Section 3, we introduce our double approximation for approximate inference with periodic kernels. Section 4 provides an empirical evaluation of the method, and Section 5 summarizes the paper.

2 Gaussian Processes

In the following, we cover the background on GPs. For a comprehensive introduction, we refer to MacKay (1998) or the book by Rasmussen and Williams (2006).

2.1 Model and Predictions

GPs are a state-of-the-art non-parametric regression method. A GP is a probability distribution over functions f. More formally, a GP is a collection of random variables $f = f_1, f_2, \ldots$, any finite number of which is Gaussian distributed Rasmussen and Williams (2006). A Gaussian process is fully determined by a mean m and a kernel/covariance function k, such that

$$f \sim \mathcal{GP}(m, k), \tag{1}$$
$$p(f(x_1), \ldots, f(x_n)|x_1, \ldots, x_n) = \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}), \tag{2}$$

where $\boldsymbol{K}$ is the full covariance matrix of the function values $f(x_1), \ldots, f(x_n)$, and $\boldsymbol{m}$ is the corresponding mean vector. Throughout this paper, we consider a prior mean function that is zero everywhere, i.e., $m \equiv 0$. This means, all relevant structure of the function f is expressed by the kernel k. The kernel encodes high-level assumptions of the underlying function f. The kernel trick (Schölkopf and Smola, 2002) allows us to compute the covariance between function values $f(\boldsymbol{x}_i)$ and $f(\boldsymbol{x}_j)$ by evaluating the kernel at the corresponding inputs, i.e., $\mathbb{C}[f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)] = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Given a training data set $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$, where $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$, and $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, the GP predictive distribution of $f(\boldsymbol{x}_*)$ at a test point $\boldsymbol{x}_*$ is Gaussian and given by $p(f(\boldsymbol{x}_*)|\mathcal{D}) = \mathcal{N}(\mu(\boldsymbol{x}_*), \sigma^2(\boldsymbol{x}_*))$, where

$$\mu(\boldsymbol{x}_*) = k(\boldsymbol{x}_*, \boldsymbol{X})(\boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I})^{-1}\boldsymbol{y} = \boldsymbol{k}_*^\top \boldsymbol{\beta}, \tag{3}$$
$$\sigma^2(\boldsymbol{x}_*) = k_{**} - \boldsymbol{k}_*^\top (\boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I})^{-1} \boldsymbol{k}_*, \tag{4}$$

where $k_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$ is the variance of the function value $f(\boldsymbol{x}_*)$ at the test input $\boldsymbol{x}_*$, $\boldsymbol{k}_* = k(\boldsymbol{X}, \boldsymbol{x}_*)$, and $\boldsymbol{\beta} = (\boldsymbol{K} + \sigma_\varepsilon^2 \boldsymbol{I})^{-1}\boldsymbol{y}$. The matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix whose entries are given by $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $i, j = 1, \ldots, n$.

We train the Gaussian processes by maximizing the marginal likelihood with respect to the hyper-parameters of the kernel and the measurement noise variance σ_ε^2 (Rasmussen and Williams, 2006).

2.2 Kernels

Kernels impose characteristics on the underlying function to be modeled by the GP. For example, a Gaussian kernel implies that the modeled function f is smooth, whereas other kernels encode lower degrees of differentiability (Matérn) or periodicity. In the following, we introduce two kernel functions, the common Gaussian (squared exponential) kernel and the periodic kernel, both of which will be used later.

The Gaussian kernel is defined as

$$k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \alpha^2 \exp\Big(-\tfrac{1}{2}\sum\nolimits_{d=1}^{D} \tfrac{(x_d - x'_d)^2}{l_d^2}\Big), \tag{5}$$

where the hyper-parameters are the signal variance α^2 and the characteristic length-scales l_d, which control the relevance of each input dimension $d = 1, \ldots, D$.

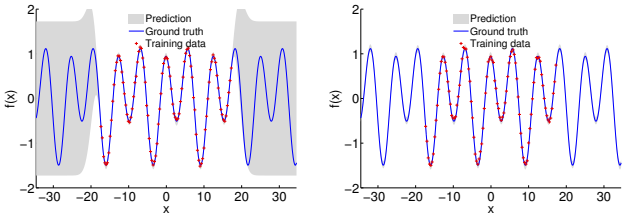
The Gaussian kernel in (5) is well suited for modeling smooth functions. However, it cannot capture periodicity. For this purpose, MacKay (1998) proposed the periodic kernel

$$k_{\mathrm{per}}(\boldsymbol{x}, \boldsymbol{x}') = \alpha^2 \exp\left(-\tfrac{1}{2}\sum\nolimits_{d=1}^{D}\left(\tfrac{\sin(b(x_d - x'_d))}{l_d}\right)^2\right). \tag{6}$$

The periodicity hyper-parameter is denoted by b[1], the signal variance by α^2, and l_i are the length-scales as in the case of the Gaussian kernel in (5).

Figure 1 illustrates the difference between a Gaussian and a periodic kernel in a GP that models a periodic function. The training targets are shown in red, the periodic latent function in blue. The shaded areas represent the $\pm 2\sigma$ bounds of the predictive distribution around the predicted mean values of $f(x_*)$. Figure 1a displays the predictive performance of the GP with the Gaussian kernel. Close to the training data the model is confident. However, since the periodicity of the signal is not encoded in the kernel, the model cannot extrapolate with confidence and falls back to the prior, indicated by the increasing error bars for $|x_*| > 17$. On the other hand, the GP with a periodic kernel extrapolates the periodic signal with confidence as shown in Figure 1b.

[1] More specifically, π/b determines the distance between repititions of the function.



(a) Gaussian kernel. (b) Periodic kernel.

Figure 1: Comparison of GPs with a Gaussian and b periodic kernels to model a periodic signal. Red crosses denote the training targets y_i and the blue line the true latent function. The shaded areas represent the $\pm 2\sigma$ bound of the marginal predictive distribution of the GP models. Outside the training data, the GP with the Gaussian kernel falls back to the prior, whereas the GP with the periodic kernel tracks the signal with high confidence.

3 Long-Term Forecasting

In order to predict long-term state evolutions $p(\boldsymbol{x}_1), p(\boldsymbol{x}_2), \ldots$ we iteratively concatenate one-step predictions. For a deterministic input $\boldsymbol{x}_t$, the GP predictive distribution $p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$ is given in (3)–(4), where $\boldsymbol{x}_t$ plays the role of the test input $\boldsymbol{x}_*$ and $\boldsymbol{x}_{t+1}$ plays the role of $f(\boldsymbol{x}_*)$.[2] In the case of long-term predictions, however, the inputs $\boldsymbol{x}_t$ are typically not given deterministically but by a probability distribution $p(\boldsymbol{x}_t)$, which we assume to be Gaussian. The predictive distribution

$$p(\boldsymbol{x}_{t+1}) = \iint p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)p(\boldsymbol{x}_t)d\boldsymbol{x}_t df \tag{7}$$

requires to integrate out both $\boldsymbol{x}_t \sim p(\boldsymbol{x}_t)$ and the function $f \sim \mathcal{GP}$, which is analytically intractable for nonlinear kernels k. Therefore, approximations of the predictive distribution $p(\boldsymbol{x}_{t+1})$ are required. We focus on Gaussian approximations by means of moment matching (Quiñonero-Candela et al., 2003), where we compute the mean and the variance of $p(\boldsymbol{x}_{t+1})$ analytically. Therefore, the predictive distributions $p(\boldsymbol{x}_1), p(\boldsymbol{x}_2), \ldots$ can be computed in closed form by repeated application of this Gaussian approximation.

In the following, we will derive the high-level steps for moment matching and identify the integrals, which cannot be computed in closed form when we use a periodic kernel. Subsequently, we will detail our double-approximation scheme to sidestep this difficulty to allow long-term forecasting with periodic Gaussian processes.

[2]To keep notation uncluttered, we tacitly ignore the Gaussian likelihood arising from the noise ε.

3.1 Moment Matching with Gaussian Processes

For moment matching with GPs, we compute the predictive mean and variance of $p(\boldsymbol{x}_{t+1})$ in (7). We assume that $p(\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and that $f \sim \mathcal{GP}$.

The exact predictive mean $\boldsymbol{\mu}_{t+1}$ is obtained by applying the law of iterated expectations and given by

$$\boldsymbol{\mu}_{t+1} = \mathbb{E}_{\boldsymbol{x}_t}[\mathbb{E}_f[\boldsymbol{x}_{t+1}|\boldsymbol{x}_t]] = \mathbb{E}_{\boldsymbol{x}_t}[m(\boldsymbol{x}_t)]\,, \tag{8}$$

where $m(\boldsymbol{x}_t)$ is the (posterior) mean function of the GP evaluated at $\boldsymbol{x}_t$. By plugging in (3) for the predicted mean, we obtain

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\beta}^\top \int k(\boldsymbol{X}, \boldsymbol{x}_t)\mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)d\boldsymbol{x}_t\,, \tag{9}$$

where $\boldsymbol{\beta} = (\boldsymbol{K} + \sigma_\varepsilon^2\boldsymbol{I})^{-1}\boldsymbol{y}$ and $\boldsymbol{X}, \boldsymbol{y}$ are the training inputs and targets, respectively.

Similarly, the predictive variance is given as

$$\begin{aligned}\boldsymbol{\Sigma}_{t+1} &= \mathbb{E}_{\boldsymbol{x}_t}[\mathbb{V}_f[\boldsymbol{x}_{t+1}|\boldsymbol{x}_t]] + \mathbb{V}_{\boldsymbol{x}_t}[\mathbb{E}_f[\boldsymbol{x}_{t+1}|\boldsymbol{x}_t]] \\ &= \mathbb{E}_{\boldsymbol{x}_t}[\sigma^2(\boldsymbol{x}_{t+1})] + \mathbb{E}_{\boldsymbol{x}_t}[m(\boldsymbol{x}_t)m(\boldsymbol{x}_t)^\top] \\ &\quad - \boldsymbol{\mu}_{t+1}\boldsymbol{\mu}_{t+1}^\top\,,\end{aligned} \tag{10}$$

where $\sigma^2(\boldsymbol{x}_t)$ is the predictive GP variance at $\boldsymbol{x}_t$, see (4). The last term in (10) is the predictive mean $\boldsymbol{\mu}_{t+1}$, which is computed in (8). By plugging in the GP mean and variance from (3) and (4), respectively, the first two terms in (10) are given as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_t}[\sigma^2(\boldsymbol{x}_t)] &= \int k(\boldsymbol{x}_t, \boldsymbol{x}_t)p(\boldsymbol{x}_t)d\boldsymbol{x}_t \\ &\quad - \int k(\boldsymbol{x}_t, \boldsymbol{X})(\boldsymbol{K} + \sigma_\varepsilon^2\boldsymbol{I})^{-1}k(\boldsymbol{X}, \boldsymbol{x}_t)p(\boldsymbol{x}_t)d\boldsymbol{x}_t\end{aligned} \tag{11}$$

and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_t}[m(\boldsymbol{x}_t)m(\boldsymbol{x}_t)^\top] &= \int m(\boldsymbol{x}_t)m(\boldsymbol{x}_t)^\top p(\boldsymbol{x}_t)d\boldsymbol{x}_t \\ &= \boldsymbol{\beta}^\top \int k(\boldsymbol{X}, \boldsymbol{x}_t)k(\boldsymbol{x}_t, \boldsymbol{X})p(\boldsymbol{x}_t)d\boldsymbol{x}_t\boldsymbol{\beta}\,.\end{aligned} \tag{12}$$

The integrals in (9), (11), and (12) depend on the choice of the kernel k. For polynomial kernels or Gaussian kernels these integrals can be computed analytically (Quiñonero-Candela et al., 2003; Deisenroth et al., 2012). However, for the periodic kernel in (6), they cannot be computed in closed form, rendering the problem of analytic moment matching for long-term forecasting of periodic GPs intractable.

To address this issue, we propose a re-parametrization of the periodic kernel in (6), which allows for an analytic approximate solution to the integrals in (9), (12), and (11). In particular, we propose a double approximation to analytically compute these integrals by exploiting the fact that these expressions can be solved analytically for the Gaussian kernel.

3.2 Double Approximation for Analytic Inference with Periodic GPs

In the following, we derive a kernel, which is equivalent to the periodic kernel in (6). After this, we will exploit this new kernel representation for a double approximation, which enables analytic inference.

3.2.1 Kernel Re-parametrization

Let us consider scalar inputs x in the following. The extension to multivariate inputs is straightforward and given by Haji Ghassemi (2013). Our periodic kernel uses a trigonometric transformation u of the inputs x and is given by

$$\begin{aligned} k_{\text{per}}(x, x') &= k_{\text{SE}}(u(x), u(x')) \\ &= \alpha^2 \exp\left(-\tfrac{1}{2} \boldsymbol{z}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{z} \right), \end{aligned} \tag{13}$$

where

$$\begin{aligned} u(x) &= \begin{bmatrix} \sin(ax) & \cos(ax) \end{bmatrix}^\top, \\ \boldsymbol{z} &= u(x) - u(x') = \begin{bmatrix} \sin(ax) - \sin(ax') \\ \cos(ax) - \cos(ax') \end{bmatrix}, \end{aligned}$$

and $\boldsymbol{\Lambda} = \text{diag}[\lambda_1^2, \lambda_2^2]$, where we assume that $\lambda_1 = \lambda = \lambda_2$, such that the sin and cos terms are scaled by the same value. The kernel in (13) is effectively a Gaussian kernel wrapped around a complex representation $u(x) \in \mathbb{R}^2$ of the input $x \in \mathbb{R}$. Therefore, this kernel is valid (MacKay, 1998).

In the following, we show that the kernel in (13) is equivalent to the periodic kernel in (6). For this purpose, let us ignore the diagonal scaling matrix $\boldsymbol{\Lambda}$ in (13) for a moment. Multiplying out $\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z}$ yields

$$\tfrac{1}{2}\boldsymbol{z}^\top \boldsymbol{z} = 1 - \sin(ax)\sin(ax') - \cos(ax)\cos(ax'). \tag{14}$$

With the identity

$$\cos(x - x') = \cos(x)\cos(x') + \sin(x)\sin(x')$$

we obtain $\frac{1}{2}\boldsymbol{z}^\top \boldsymbol{z} = 1 - \cos(a(x - x'))$. Now, we apply the identity $\cos(2x) = 1 - 2\sin^2(x)$ and obtain

$$\tfrac{1}{2}\boldsymbol{z}^\top \boldsymbol{z} = 2\sin^2\left(\tfrac{a(x-x')}{2}\right).$$

Incorporating the diagonal scaling $\boldsymbol{\Lambda}$ from (13) yields

$$\exp\left(-\tfrac{1}{2}\boldsymbol{z}^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{z}\right) = \exp\left(-\frac{2\sin^2\left(\frac{a(x-x')}{2}\right)}{\lambda^2} \right).$$

With $\lambda^2 = 2l^2$ and $a = 2b$, we see that the kernel in (13) is equivalent to the periodic kernel in (6).

3.2.2 Approximate Inference

In the following, we detail how the kernel in (13) can be used for long-term forecasting, where we approximate the intractable integrals in (9), (11), and (12).

The high-level idea is to use a two-step approximation (double approximation) to compute a Gaussian approximation to the desired predictive distribution $p(\boldsymbol{x}_{t+1})$ from $p(\boldsymbol{x}_t)$. First, we analytically compute a Gaussian approximation $p(u(\boldsymbol{x}_t))$ of the trigonometrically augmented state $\boldsymbol{x}_t$. Second, we analytically compute a Gaussian approximation to $p(\boldsymbol{x}_{t+1})$ by exploiting the fact that we can map the Gaussian $p(u(\boldsymbol{x}_t))$ through a Gaussian kernel. Figure 2 illustrates this procedure. The top row of the figure shows the desired path, which is intractable. The two steps at the bottom of the figure correspond to our proposed double approximation using the periodic kernel in (13). First, the input distribution $p(\boldsymbol{x}_t)$ is mapped to the trigonometric space $p(u(\boldsymbol{x}_t))$, which is subsequently mapped through a GP with a Gaussian kernel. In the following, we discuss both steps in detail.

Step 1: Mapping to Trigonometric Space When mapping a Gaussian distribution $p(\boldsymbol{x})$ through u, we obtain a non-Gaussian distribution $p(u(\boldsymbol{x})) = p(\sin(a\boldsymbol{x}), \cos(a\boldsymbol{x}))$, which cannot be computed analytically. In this paper, we propose a Gaussian approximation to $p(u)$, which will be convenient for the purpose of long-term forecasting. It turns out that the mean and variance of the trigonometrically augmented variable $u(\boldsymbol{x}) \in \mathbb{R}^{2D}$ can be computed analytically. For notational convenience, we will detail the computations in the following for scalar variables $x \in \mathbb{R}^D$. The extension to multivariate $\boldsymbol{x}$ is detailed by Haji Ghassemi (2013).

Let us assume that $p(x) = \mathcal{N}(x|\mu, \sigma^2)$. The mean $\tilde{\boldsymbol{\mu}}$ and covariance $\tilde{\boldsymbol{\Sigma}}$ of $p(u(x))$ are given as

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \mathbb{E}[\sin(ax)] \\ \mathbb{E}[\cos(ax)] \end{bmatrix}, \tag{15}$$

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \mathbb{V}[\sin(ax)] & \mathbb{C}[\sin(ax), \cos(ax)] \\ \mathbb{C}[\cos(ax), \sin(ax)] & \mathbb{V}[\cos(ax)] \end{bmatrix}, \tag{16}$$

where $\mathbb{C}$ denotes the covariance between two variables.

Using results from convolving trigonometric functions with Gaussians (Gradshteyn and Ryzhik, 2000), we obtain

$$\mathbb{E}[\sin(ax)] = \exp(-\tfrac{1}{2}a^2\sigma^2)\sin(a\mu), \tag{17}$$

$$\mathbb{E}[\cos(ax)] = \exp(-\tfrac{1}{2}a^2\sigma^2)\cos(a\mu), \tag{18}$$

which allows us to compute the mean $\tilde{\boldsymbol{\mu}}$ in (15) analytically.

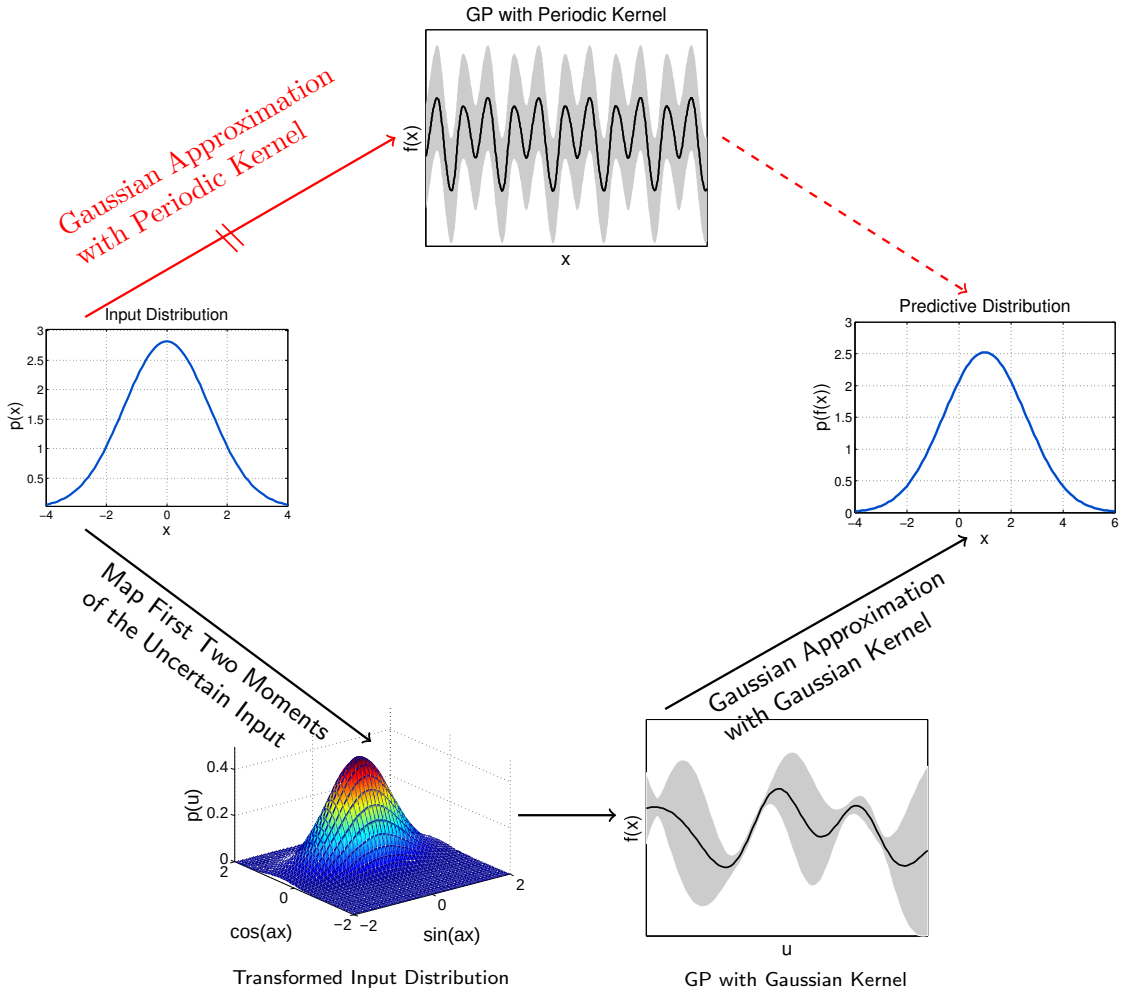


Figure 2: The objective is to map a Gaussian distribution (left) through a periodic GP (top) to compute the exact mean and variance of the predictive distribution (right). This ideal path is analytically intractable. Therefore, we propose a two-step approximation (bottom path). First, the input distribution is mapped into a trigonometrically augmented space via analytic moment matching. Subsequently, the Gaussian approximation in this augmented space is mapped through a GP with a Gaussian kernel to approximate the exact moments of the predictive distribution.

To compute the covariance matrix $\tilde{\boldsymbol{\Sigma}}$ in (16), we need to compute the marginal variances $\mathbb{V}[\sin(ax)]$, $\mathbb{V}[\cos(ax)]$ and the cross-covariance terms $\mathbb{C}[\sin(ax), \cos(ax)]$.

The marginal variance of $\sin(ax)$ is given by

$$\mathbb{V}[\sin(ax)] = \mathbb{E}[\sin^2(ax)] - \mathbb{E}[\sin(ax)]^2\,, \tag{19}$$

where $\mathbb{E}[\sin(ax)]$ is given in (17) and

$$\mathbb{E}[\sin^2(a\boldsymbol{x})] = \int \sin^2(ax)p(x)dx \tag{20}$$

$$= \tfrac{1}{2}(1 - \exp(-2a^2\sigma^2)\cos(2a\mu))\,. \tag{21}$$

Similarly, the marginal variance of $\cos(ax)$ is given by

$$\mathbb{V}[\cos(ax)] = \mathbb{E}[\cos^2(ax)] - \mathbb{E}[\cos(ax)]^2\,, \tag{22}$$

where $\mathbb{E}[\cos(ax)]$ is given in (18) and

$$\mathbb{E}[\cos^2(ax)] = \tfrac{1}{2}(1 + \exp(-2a^2\sigma^2)\cos(2a\mu))\,. \tag{23}$$

The cross-covariance term $\mathbb{C}[\sin(ax), \cos(ax)]$ is

$$\begin{aligned}&\mathbb{C}[\sin(ax), \cos(ax)]\\ &= \mathbb{E}[\sin(ax)\cos(ax)] - \mathbb{E}[\sin(ax)]\mathbb{E}[\cos(ax)]\,,\end{aligned} \tag{24}$$

where $\mathbb{E}[\sin(ax)]$ and $\mathbb{E}[\cos(ax)]$ are given in (17) and (18), respectively. The first term in (24) is computed according to

$$\mathbb{E}[\sin(ax)\cos(ax)] = \tfrac{1}{2}\exp(-2a^2\sigma^2)\sin(2a\mu)\,, \tag{25}$$

where we exploited that $\sin(x)\cos(x) = \sin(2x)/2$.

These results allow us to analytically compute the mean $\tilde{\boldsymbol{\mu}}_t$ and the covariance matrix $\tilde{\boldsymbol{\Sigma}}_t$ of a trigonometrically transformed variable $u(\boldsymbol{x}_t)$ for $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$.

Step 2: Computing the Predictive Distribution Now we turn to the second step of the double approximation, which is the analytic computation of the terms (9)–(11) with the trigonometrically transformed inputs $\boldsymbol{u}_t = u(\boldsymbol{x}_t)$, where $\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{u}_t|\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$. For this purpose, we also augment the GP training inputs $\boldsymbol{X}$ trigonometrically into $\boldsymbol{U}$ and use results from Quiñonero-Candela et al. (2003) and Deisenroth et al. (2012) to map $p(u(\boldsymbol{x}_t)) \approx \mathcal{N}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$ through a GP with a Gaussian kernel to compute the mean and the covariance of $p(\boldsymbol{x}_{t+1})$.

The predictive mean $\boldsymbol{\mu}_{t+1}$ in (9) can be written as

$$\boldsymbol{\mu}_{t+1} = \tilde{\boldsymbol{\beta}}^\top \int k_{\mathrm{SE}}(\boldsymbol{U}, \boldsymbol{u}_t)\mathcal{N}(\boldsymbol{u}_t|\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)d\boldsymbol{u}_t\,,$$

where we define $\tilde{\boldsymbol{\beta}} = (k_{\mathrm{SE}}(\boldsymbol{U},\boldsymbol{U}) + \sigma_\varepsilon^2\boldsymbol{I})^{-1}\boldsymbol{y} \in \mathbb{R}^n$. Note that the kernel in this integral is no longer a periodic kernel, but a Gaussian (squared exponential) kernel, applied to the trigonometrically transformed inputs $\boldsymbol{u}_t$. We define

$$\boldsymbol{q} = \int k_{\mathrm{SE}}(\boldsymbol{u}_t, \boldsymbol{U})p(\boldsymbol{u}_t)d\boldsymbol{u}_t\,,$$

which can be computed analytically. The elements of $\boldsymbol{q} \in \mathbb{R}^n$ are given by

$$q_j = \frac{\alpha^2}{\sqrt{|\tilde{\boldsymbol{\Sigma}}_t\boldsymbol{\Lambda}^{-1}+\boldsymbol{I}|}} \exp(-\tfrac{1}{2}\boldsymbol{\zeta}_j^\top(\tilde{\boldsymbol{\Sigma}}+\boldsymbol{\Lambda})^{-1}\boldsymbol{\zeta}_j) \qquad (26)$$

for $j = 1,\ldots,n$, where $\boldsymbol{\zeta}_j = (\boldsymbol{u}_j - \tilde{\boldsymbol{\mu}}_t)$.

To compute the predictive covariance matrix $\boldsymbol{\Sigma}_{t+1}$, we need to solve the following integrals, see (11)–(12):

$$\int k_{\mathrm{SE}}(\boldsymbol{u}_t, \boldsymbol{u}_t)p(\boldsymbol{u}_t)d\boldsymbol{u}_t\,, \qquad (27)$$

$$\int k_{\mathrm{SE}}(\boldsymbol{U}, \boldsymbol{u}_t)k_{\mathrm{SE}}(\boldsymbol{u}_t, \boldsymbol{U})p(\boldsymbol{u}_t)d\boldsymbol{u}_t\,. \qquad (28)$$

Note that the second integral in (11) can be expressed in terms of (28) by using $\boldsymbol{a}^\top\boldsymbol{b} = \mathrm{trace}(\boldsymbol{b}\boldsymbol{a}^\top)$. Since the Gaussian kernel k_{SE} is stationary, the integral in (27) is simply given by the signal variance α^2. The integral in (28) results in a matrix $\boldsymbol{Q}$, whose entries are

$$\begin{aligned} Q_{ij} = {} & |2\boldsymbol{\Lambda}^{-1}\tilde{\boldsymbol{\Sigma}}_t + \boldsymbol{I}|^{-1/2} \\ & \times k_{\mathrm{SE}}(\boldsymbol{u}_i, \tilde{\boldsymbol{\mu}}_t)k_{\mathrm{SE}}(\boldsymbol{u}_j, \tilde{\boldsymbol{\mu}}_t) \\ & \times \exp(-\tfrac{1}{2}(\boldsymbol{\nu}-\tilde{\boldsymbol{\mu}}_t)^\top(\tfrac{1}{2}\boldsymbol{\Lambda}+\tilde{\boldsymbol{\Sigma}}_t)^{-1}(\boldsymbol{\nu}-\tilde{\boldsymbol{\mu}}_t)) \end{aligned}$$

for $i,j = 1,\ldots,n$ and with $\boldsymbol{\nu} = (\boldsymbol{u}_i + \boldsymbol{u}_j)/2$.

These results allow us to analytically compute approximations to the mean $\boldsymbol{\mu}_{t+1}$ and the covariance $\boldsymbol{\Sigma}_{t+1}$ of the successor state distribution $p(\boldsymbol{x}_{t+1})$ for GPs with periodic kernels. Although all computations can be performed analytically, the additional Gaussian approximation of the trigonometrically transformed state variable $\boldsymbol{u}_t$ (Step 1) makes the computation of $\boldsymbol{\mu}_{t+1}$ and $\boldsymbol{\Sigma}_{t+1}$ only approximate. In Section 4 we will shed some light on the quality of this approximation.

4 Experiments

In this section, we assess the quality of the proposed double approximation for a single time step and for long-term forecasting of periodic signals. More specifically, in Section 4.1, we quantify the error introduced by the double approximation compared to an optimal Gaussian approximation and a kernel density estimator of the true predictive distribution. In Section 4.2, we demonstrate the advantage of our model when predicting the long-term evolution of a periodic pattern.

4.1 Quality of the Double Approximation

To evaluate the quality of our proposed double approximation, we considered the periodic signal $y = \sin(x/2)+\cos(x+0.35)+\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1.6\times10^{-3})$. We trained our GP model on a data set of size 400, where the training inputs x_i were linearly spaced between -17 and 17. The test data were in the range $[-11\pi, 11\pi]$. The function and the range of the training data are visualized in Figure 1b in blue and red, respectively.

We defined test input distributions $p(x_0^{ij}) = \mathcal{N}(\mu_i, \sigma_j^2)$ from which we sampled 100 inputs x_* and mapped them through the periodic GP. The mean values μ_i of the test input distributions $p(x_0^{ij})$ were selected on a linear grid from -11π to 11π. The corresponding variances σ_j^2 were set to 10^{-j}, $j = 0,\ldots,4$. Moreover, we tested the approximation for $\sigma_0^2 = 0$, which corresponds to a deterministic input.

We computed the root-mean-squared error (RMSE) and the negative log predictive density (NLPD) of the true function values under three models for $p(y)$: our proposed analytic double approximation (DA), an optimal Gaussian approximation of the true predictive density (SA), which was determined from the sample mean and variance, and a ground-truth baseline (BASE) using a kernel density estimator.

Table 1 shows the average performance of the double approximation, where we averaged the NLPD and RMSE values over all means μ_i of the test input distributions. The RMSE values for DA, SA, and BASE are basically identical across varying input variances σ_j^2. This means that the mean estimate by the double approximation is relatively robust. The NLPD values on the other hand indicate that the coherence of the predictive distribution suffers to from increasing uncertainty in the input distribution. However, the sampling-based optimal Gaussian approximation (SA) and the double approximation perform equally well.

Figure 3 shows examples of the quality of the approximation of the predictive distribution $p(y)$ for all three

Table 1: Average NLPD and RMSE performance of the double approximation (DA), an optimal single approximation (SA) by sampling-based moment matching, and a ground-truth baseline (BASE).

	NLPD						RMSE ($\times 10^{-2}$)					
σ_j^2	0	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	0	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1
DA	−1.78	−1.77	−1.65	−1.18	−0.24	0.85	4.1	4.1	4.7	7.9	21.1	57.4
SA	−1.78	−1.77	−1.65	−1.18	−0.24	0.84	4.1	4.1	4.7	7.9	21.1	57.3
BASE	−1.78	−1.76	−1.65	−1.17	−0.31	0.58	4.1	4.1	4.7	7.9	21.1	57.4

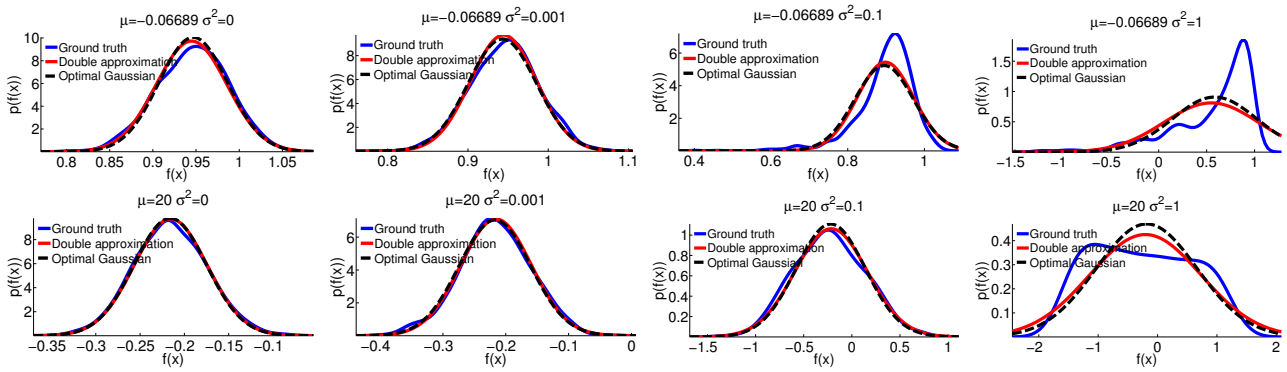


Figure 3: Predictive distributions when mapping a Gaussian (with different variances (columns) and at two different means (rows)) through a periodic GP. The Gaussian computed by means of the proposed double approximation is shown in red, an optimal Gaussian approximation (moment matching) based on sampling is shown in black. A kernel density estimate of the true predictive density is shown in blue. We observe that the double approximation is close to the optimal Gaussian across all experiments and that the kernel density estimate is close to these Gaussians for small variances ($\sigma^2 \leq 0.1$), which is also confirmed by Table 1.

models. The double approximation is close to the optimal Gaussian and that the kernel density estimate is close to these Gaussians for small variances ($\sigma^2 \leq 0.1$).

4.2 Long-Term Forecasting of Limit-Cycle Behavior

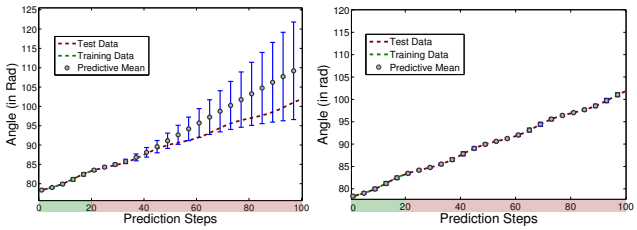
To evaluate the long-term performance of our double approximation, we considered limit-cycle behavior of a pendulum motion. The state $\boldsymbol{x}$ of the pendulum system was given by the angle φ and the angular velocity $\dot{\varphi}$. The angle φ measures the deviation of the pendulum from the vertical, measured anti-clockwise in radians. A constant torque was applied to the pendulum, such that it reached a limit-cycle behavior after about 2 s, in which both the angle and the angular velocity followed periodic patterns. Once the pendulum was in the limit cycle, we trained a GP on 300 data points, where the measurement noise variance was $10^{-2}\boldsymbol{I}$.

For model learning, we trained the hyper-parameters of the periodic GP (periodicity a, length-scales l_i, signal variance α^2, and noise variance σ_ε^2). Moreover, we trained a GP with a Gaussian kernel, where the hyper-parameters were the length-scales l_i, the signal variance α^2, and the variance σ_ε^2. The training targets for both GP models were the differences between consecutive states, i.e., $\boldsymbol{y}_i = \boldsymbol{x}_i - \boldsymbol{x}_{i-1}$, which effectively encodes a constant prior mean function.

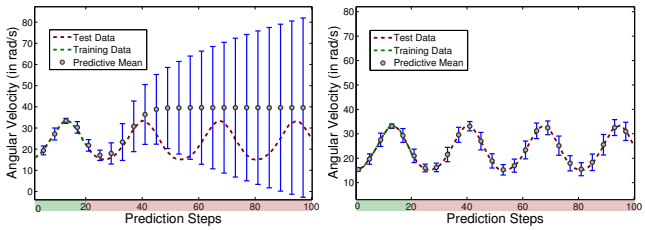
To evaluate the performance of the models for long-term forecasting, the models were used to predict the pendulum's state evolution for 100 time steps ahead. We took the last point in our training set as the mean $\boldsymbol{\mu}_0$. We set the initial covariance to $0.01\boldsymbol{I}$, i.e., the 95% confidence bound allowed for a deviation of about 12° in the angle φ. For long-term forecasting with the Gaussian kernel, we used the moment-matching technique proposed by Quiñonero-Candela et al. (2003).

Figure 4 shows the long-term predictive performance of both the GP with a Gaussian kernel and the GP with a periodic kernel. When the predictive distributions are no longer "covered" by the training data, the GP with Gaussian kernel suffered from the stationarity assumption and lost track of the state. On the other hand, the periodic GP could predict the long-term limit-cycle behavior of the pendulum's state better. Despite an 80-step ahead prediction, the uncertainty is fairly small in both the angle and angular velocity (right graphs in Figures 4a and 4b).

Table 2 summarizes the average long-term predictive performance of the periodic GP using the proposed double approximation and the GP with a Gaussian



(a) Long-term prediction of the angle with GPs. Left: Gaussian kernel, right: periodic kernel.



(b) Long-term prediction of the angular velocity prediction with GPs. Left: Gaussian kernel, right: periodic kernel.

Figure 4: GP long-term predictions of the limit-cycle behavior of the pendulum system. The training and test data are shown in green and red, respectively. The mean predictions of the model are indicated by the circles. The corresponding 2σ confidence bounds are represented by the error bars. Once the test inputs are sufficiently far from the training data, the GP with the Gaussian kernel loses track of the state, while the periodic GP predicts the state with high confidence, even 80 time steps ahead. a: Multiple-step ahead prediction of the pendulum's angle, b: multiple-step ahead prediction of the angular velocity.

Table 2: Average NLPD and RMSE (pendulum tracking) for GPs with Gaussian and periodic kernels.

	NLPD					RMSE (angle)					RMSE (angular velocity)				
	k-Step Ahead Prediction					k-Step Ahead Prediction					k-Step Ahead Prediction				
Kernel	20	40	60	80	100	20	40	60	80	100	20	40	60	80	100
Gaussian	5.89	5.67	6.52	7.11	7.38	1.31	2.18	3.01	4.32	5.50	8.59	10.76	7.80	2.58	9.40
Periodic	0.80	1.19	1.49	1.75	2.29	0.15	0.15	0.13	0.16	0.22	0.89	0.94	0.82	0.92	1.08

kernel for predicting the pendulum's limit-cycle behavior. We averaged the performance for k-step ahead predictions over 100 different initial states and 100 trajectories per state. The relatively small RMSE (Gaussian kernel) of predicting the angular velocity for $k = 80$ steps ahead occurs since at this time step the true angular velocity is fairly close to zero, which corresponds to the posterior predictive mean of the GP with Gaussian kernel. While the GP with a Gaussian kernel immediately loses track of the state (RMSE is measured in radians) when predicting the limit-cycle behavior, the periodic GP predicts the true periodic behavior with high confidence. The slight performance loss in NLPD and RMSE is due to the repeated propagation of uncertainty over time without obtaining a single new measurement.

5 Conclusion

In this paper, we proposed an algorithm for long-term forecasting with periodic Gaussian processes. For long-term forecasting, it is necessary to iteratively map probability distributions through a Gaussian process. If these probability distributions are Gaussians, the moments of the predictive distributions can only be computed for Gaussian or polynomial kernels, but not for periodic kernels. We exploited a re-parametrization of a commonly used stationary periodic kernel, which is equivalent to applying a standard Gaussian kernel to a trigonometrically transformed input. This allowed us to employ an analytic double-approximation strategy to compute the moments of the predictive distribution. We evaluated the quality of the double approximation on a periodic example system and applied our methodology to long-term forecasting of the limit-cycle behavior of a pendulum system.

Currently, our approach is limited to data sets, which are exactly periodic, e.g., angular relationships. In future, we will generalize our inference method to data sets, which have a periodic trend but where the periods are not exactly identical. This can be achieved by multiplying the periodic kernel with a Gaussian kernel, for instance, which has also been suggested by Roberts et al. (2013). We will investigate the extension of these models to long-term forecasting. Moreover, we believe that spatio-temporal models could greatly profit from long-term forecasting with periodic kernels.

Acknowledgements

NHG acknowledges financial support from the Department of Computing, Imperial College London. We thank the anonymous reviewers for valuable paper feedback. Large parts of this research have been conducted at TU Darmstadt and with support from Jan Peters. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement #270327 and the Department of Computing, Imperial College London.

References

Deisenroth, M. P., Fox, D., and Rasmussen, C. E. (2014). Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Deisenroth, M. P., Turner, R., Huber, M., Hanebeck, U. D., and Rasmussen, C. E. (2012). Robust filtering and smoothing with Gaussian processes. *IEEE Transactions on Automatic Control*, 57(7):1865–1871.

Durrande, N., Hensman, J., Rattray, M., and Lawrence, N. D. (2013). Gaussian process models for periodicity detection. http://arxiv.org/abs/1303.7090.

Girard, A., Rasmussen, C. E., Quiñonero Candela, J., and Murray-Smith, R. (2003). Gaussian process priors with uncertain inputs—application to multiple-step ahead time series forecasting. In *Advances in Neural Information Processing Systems*, pages 529–536. MIT Press.

Gradshteyn, I. S. and Ryzhik, I. M. (2000). *Table of Integrals, Series, and Products*. Academic Press, 6th edition edition.

Haji Ghassemi, N. (2013). Analytic long-term forecasting with periodic Gaussian processes. Master's thesis, Blekinge Institute of Technology.

MacKay, D. J. C. (1998). Introduction to Gaussian processes. *Neural Networks and Machine Learning*.

Neal, R. M. (1997). Probabilistic inference using Markov chain Monte Carlo methods. Technical report, Technical Report CRG-TR-92-1, Department of Computer Science, University of Toronto.

O'Hagan, A. (1978). On curve fitting and optimal design for regression. *Journal of the Royal Statistical Society B*, 40(1):142.

Quiñonero-Candela, J., Girard, A., Larsen, J., and Rasmussen, C. E. (2003). Propagation of uncertainty in Bayesian kernel models. In *International Conference on Acoustics, Speach, and Signal Processing*, pages 701–704.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Reece, S. and Roberts, S. (2010). The near constant acceleration Gaussian process kernel for tracking. *IEEE Signal Processing Letters*, 17(8):707–710.

Roberts, S., Osborne, M. A., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. (2013). Gaussian processes for time series modelling. *Philosophical Transactions of the Royal Society (Part A)*, 371(1984).

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels—Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. *Advanves in Neural Information Processing Systems*.