The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Fourier Feature Approximations for Periodic Kernels in Time-Series Modelling

Anthony Tompkins, Fabio Ramos
School of Information Technologies
The University of Sydney, New South Wales, 2006, Australia
atom7176@uni.sydney.edu.au, fabio.ramos@sydney.edu.au

Abstract

Gaussian Processes (GPs) provide an extremely powerful mechanism to model a variety of problems but incur an $O(N^3)$ complexity in the number of data samples. Common approximation methods rely on what are often termed *inducing points* but still typically incur an $O(NM^2)$ complexity in the data and corresponding inducing points. Using Random Fourier Feature (RFF) maps, we overcome this by transforming the problem into a Bayesian Linear Regression formulation upon which we apply a Bayesian Variational treatment that also allows learning the corresponding kernel hyperparameters, likelihood and noise parameters. In this paper we introduce an alternative method using Fourier series to obtain spectral representations of common kernels, in particular for periodic warpings, which surprisingly have a convergent, non-random form using special functions, requiring fewer spectral features to approximate their corresponding kernel to high accuracy. Using this, we can fuse the Random Fourier Feature spectral representations of common kernels with their periodic counterparts to show how they can more effectively and expressively learn patterns in time-series for both interpolation and extrapolation. This method combines robustness, scalability and equally importantly, interpretability through a symbolic declarative grammar that is both functionally and humanly intuitive - a property that is crucial for explainable decision making. Using probabilistic programming and Variational Inference we are able to efficiently optimise over these rich functional representations. We show significantly improved Gram matrix approximation errors, and also demonstrate the method in several time-series problems comparing other commonly used approaches such as recurrent neural networks.

1 Introduction

Non-parametric modelling methods (Ghahramani 2005) such as GPs (Rasmussen and Williams 2006), infinite Hidden Markov Models, infinite latent factor models, and Dirichlet process mixtures are flexible modelling methods that, in contrast to parametric models, assume the data distribution cannot be defined by a finite set of parameters θ. A commonly used technique is GP regression which defines a distribution over *functions*: $p(f)$. Such methods can give useful probabilistic inference capabilities and are highly applicable for decision making processes - such problems include environmental or disease modelling, robotics, and control systems, and these can in turn influence planning decisions by humans or automated systems. Regarding temporal modelling, the ability to accurately model long term forecasts in the form of multi-step ahead predictions is often critical to making informed decisions and indeed these processes often contain fully or quasi-periodic trends which are difficult or impossible to model by only using individual kernels.

Although GPs in their original formulation are an excellent method for modelling data, they are often unable to take into account all the information present in large datasets. Recent developments in stochastic gradient optimisation which take advantage of automatic-differentiation and Variational Inference methods (Kingma and Welling 2014; Tran et al. 2016) have extended the applicability of machine learning to huge datasets. Furthermore, Bayesian machine learning algorithms have a number of well defined methods for learning model parameters and *hyperparameters* from training data, that do not involve cross validation. When used in combination, stochastically optimized Bayesian machine learning algorithms allow practitioners to learn probabilistic predictors from large data sets with minimal tuning and retraining.

Previous works investigating periodicity in the context of GPs often assume pre-existing structures within the data and set the periodicity to be reasonable based on expert knowledge of the problem (Senanayake, Simon Timothy, and Ramos 2016) (e.g. yearly), constrained random initialisation based with or without random restarts (Solin and Särkkä 2014), heuristic re-optimisation (Klenske et al. 2016). We demonstrate that one can use a natural formalisation using the Fast Fourier Transform (FFT) for seeding kernel hyperparameter periodicities and then optimise within a full variational model. This step further demonstrates a reduction in computational burden of either relying on a good random initialisation for selecting periodic hyperparameters or spawning large numbers of periodic kernels across a sweep of frequencies.

In this paper we make use of some of these recent developments in stochastic gradient methods and stochastic variational inference for supervised regression tasks. This paper thus presents a novel compositional model and methodology

for capturing short and long term temporal trends in signals. Our presented contributions (i) show that GPs are indeed scalable to large datasets in their dual form, (ii) demonstrate the general methodology for deriving convergent Fourier Series Features (FSF) representations of periodic analogues of stationary kernels, as well as empirically evaluate their Gram matrix reconstruction error, (iii) show FSF may be integrated into compositional kernel learning, (iv) show how FSF can be formulated in a Bayesian Linear Regression model with variational inference using stochastic marginalisation over standard length-scale hyperparameters as well as periodic ones, and (v) demonstrate Fourier Series Features in a compositional feature-space framework and evaluate their predictive performance on four real world time-series datasets.

2 Preliminaries

Kernel methods (Schölkopf and Smola 2002) are perhaps one of the most widespread examples of non-parametric modelling methods in which $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is some kernel on an input domain $\mathcal{X} \subset \mathbb{R}^D$. This kernel k may correspond to an embedding in a high-dimensional Hilbert space $\mathcal{H}$ through a feature map $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ via an inner product between points from the feature map with $\kappa(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{H}}$.

GP regression is a method of learning some probability distribution over functions $f(\mathbf{x}_*)$ given inputs $\mathbf{x} \in \mathbb{R}^D$ given training data $\mathcal{D} = \{\mathbf{X}_n, y_n\}$ where $n = 1, 2, ..., N$. The model $f \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'; \theta))$ is a collection of Gaussian random process priors with Gaussian noise: $y_n = f(\mathbf{x}_n) + \varepsilon_n$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. This gives the predictive form $p(f(\mathbf{x}_*)|\mathbf{x}_*, \mathcal{D}) = \mathcal{N}(\mathbb{E}[f(\mathbf{x}_*)], \mathbb{V}[f(\mathbf{x}_*)])$ which has a closed form solution for the mean $\mathbb{E}[f(\mathbf{x}_*)] = \kappa_*^T (K + \sigma^2 I)^{-1} y$ and variance $\mathbb{V}[f[\mathbf{x}_*] = \kappa(\mathbf{x}_*, \mathbf{x}_*) - \kappa_*^T (K + \sigma^2 I)^{-1} \kappa_*$ where $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ is the Gram matrix of the kernel function, k_* is an N-dimensional vector with the i^{th} entry being $\kappa(\mathbf{x}_*, \mathbf{x}_i)$ and y a vector of the N observations. Typically this solution involves an $O(N^3)$ complexity arising from an $N \times N$ matrix inversion which significantly hinders scalability to massive datasets.

There have been several past works investigating long term trends in temporal problems: (Ghassemi and Deisenroth 2014; Senanayake, Simon Timothy, and Ramos 2016; Solin and Särkkä 2014; Roberts et al. 2013).

2.1 Random Fourier Features

The work by (Rahimi and Recht 2007), broadly termed *Random Fourier Features*, motivates using a randomized lower-dimensional feature mapping that allow scalability. Additionally, there exist various works that focus on either additionally learning or alternatively representing these spectral fequencies such as FastFood (Le, Sarlós, and Smola 2013), A la Carte (Yang et al. 2015) and Quasi-Monte Carlo feature maps (Avron et al. 2016) and in fact these extensions may be directly applied within our presented framework. RFF constructions involve approximating the feature map $\hat{\Phi} : \mathcal{X} \rightarrow \mathbb{C}^C$ where $\mathbb{C}^C$ is the space of C-dimensional complex numbers. This gives the definition of the approximate feature map

$$\kappa(\mathbf{x}, \mathbf{x}') \approx \langle \hat{\Phi}(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathbb{C}^C}. \tag{1}$$

The key result from (Rahimi and Recht 2007) from which Fourier Features can reconstruct positive definite kernels is summarised below following Theorem 1:

Theorem 1 *(Bochner 1933) A complex-valued function $g : \mathbb{R}^D \rightarrow \mathbb{C}$ is positive definite if and only if it is the Fourier Transform of a finite non-negative Borel measure μ on $\mathbb{R}^D$:*

$$g(\mathbf{x}) = \hat{\mu}(\mathbf{x}) = \int_{\mathbb{R}^D} e^{-i\mathbf{x}^T \boldsymbol{\omega}} d\mu(\boldsymbol{\omega}), \quad \forall \mathbf{x} \in \mathbb{R}^D \tag{2}$$

Without loss of generality assuming the measure μ has an associated probability density function p, we have

$$\kappa(\mathbf{x}, \mathbf{x}') = g(\mathbf{x} - \mathbf{x}') = \int_{\mathbb{R}^D} e^{-i(\mathbf{x}-\mathbf{x}')^T \boldsymbol{\omega}} p(\omega) d\boldsymbol{\omega}. \tag{3}$$

allowing a shift-invariant kernel to be approximated as

$$\begin{aligned} \kappa(\mathbf{x}, \mathbf{x}') &= \int_{\mathbb{R}^D} e^{-i(\mathbf{x}-\mathbf{x}')^T \boldsymbol{\omega}} p(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &\approx \frac{1}{C} \sum_{c=1}^{C} e^{-i(\mathbf{x}-\mathbf{x}')^T \boldsymbol{\omega}_C} \\ &= \langle \hat{\Phi}(\mathbf{x}), \hat{\Phi}(\mathbf{x}') \rangle_{\mathbb{C}^C}, \end{aligned} \tag{4}$$

yielding the kernel approximation of $\kappa(\mathbf{x}, \mathbf{x}')$ as:

$$\kappa(\mathbf{x}, \mathbf{x}') \approx \frac{1}{C} \sum_{c=1}^{C} \big[\cos(\omega_c^T (\mathbf{x} - \mathbf{x}')) \big]. \tag{5}$$

where C is the number of spectral samples from the density p. This is in fact a *Monte Carlo* (MC) approximation to the integral. Through the standard trigonometric identity $\cos(u - v) = \cos(u)\cos(v) + \sin(u)\sin(v)$ we arrive at the $2C$-dimensional mapping $\Phi : \mathcal{X} \rightarrow \mathbb{R}^{2C}$ the final representation:

$$\begin{aligned} \hat{\Phi}(\mathbf{x}) = \frac{1}{\sqrt{C}} \Big[&\cos(\mathbf{x}^T \omega_1), ..., \cos(\mathbf{x}^T \omega_C), \\ &\sin(\mathbf{x}^T \omega_1), ..., \sin(\mathbf{x}^T \omega_C) \Big]. \end{aligned} \tag{6}$$

2.2 Kernel Compositions

While standard kernels such as the Squared Exponential (SE) provide suitable expressiblity for modelling many problems, alone, they are incapable of identifying and containing more realistic complexities present within time-series. While there exists Multiple Kernel Learning for GPs, this is largely concerned with a weighted sum of standard kernels. More recently, a promising body of work termed Compositional Kernel Learning (CKL) and Structure Discovery has appeared in the Automatic Machine Learning (AutoML) literature (Klenske et al. 2016; Duvenaud et al. 2013). While these methods use the full kernel, we present a scalable method in the dual space using Fourier decompositions in terms of basis functions. One of the most useful features of CKL is interpretability by humans in which

each kernel and their compositions have an intuitive real-world interpretation; and indeed recent work by (Schulz et al. 2016) demonstrates parallels between human thought processes and favourable interpretation by humans with compositional kernel rather than functionally similar non-compositional alternatives.

Multiple Kernel Learning methods (Gönen and Alpaydın 2011), also termed Kernel Compositions, have appeared in the kernel and in particular GP literature in recent years. Originally focusing on standard kernels and weighted sums but expanding towards more complex compositions, they provide a potentially more expressive way of modelling. In the last few years there has been a surge of work delving into such compositions and we are inspired by these works as a motivation for attempting to make them even more scalable through recent advancements in variational inferences and spectral methods for kernels. Indeed our system easily lends itself to automatic probabilistic search over the symbolic compositions through methods in the kernel search literature such as in Automatic Bayesian Covariance Discovery (ABCD) (Lloyd et al. 2014), and Bayesian Optimisation over Models (Malkomes, Schaff, and Garnett 2016; Rainforth et al. 2016).

3 Fourier Features for Periodic Kernels

In most applications regarding GPs one encounters the Squared Exponential (SE) (MacKay 1998) kernel most often. For demonstration we explore the SE, however in our implementation we draw from a variety of kernels including the SE, Matèrn $1/2$, Matèrn $3/2$. Following (MacKay 1998) and using the warping $\mathbf{u}(t) = [\sin(t), \cos(t)]^T$ one can construct an isotropic stationary periodic kernel within an existing non-periodic stationary kernel. First consider the general distance metric appearing in many stationary kernels:

$$\begin{aligned}\|\mathbf{u}(t) - \mathbf{u}(t')\|^2 \\ &= (\sin(t) - \sin(t'))^2 + (\cos(t) - \cos(t'))^2 \\ &= 4\sin^2(\frac{t-t'}{2}) = 2(1 - \cos(\tau)),\end{aligned} \tag{7}$$

where $\tau = t - t'$. The SE kernel defined as:

$$\kappa_{SE}(\mathbf{x} - \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right). \tag{8}$$

which has hyperparameter length-scale l and corresponding spectral density $\boldsymbol{\omega} \sim \mathcal{N}\big(\mathbf{0}, l^{-1}\mathbf{I}_C\big)$.

Following the definition of (8) it is possible to construct a warping of an input t with the function $\mathbf{u}(t)$. Using this formulation we can substitute (7) into (8) to obtain the standard Periodic SE kernel:

$$\kappa_{perSE}(t, t') = \kappa_{perSE}(\tau) = \exp\left(-\frac{\cos(w_0\tau) - 1}{l^2}\right). \tag{9}$$

Following a process analogous to deriving Random Fourier Features for standard stationary kernels, we demonstrate an alternative method for arriving at the periodic kernel and show a direct link to the Taylor series expansion from (Solin and Särkkä 2014) while also formalising it within the Fourier Features framework.

To begin, note that κ_{perSE} is both periodic and symmetric over τ. Due to periodicity we can represent the kernel as a Fourier Series over the interval $[-L, L]$ where L is the half period and fundamental frequency $\omega_0 = \frac{\pi}{L}$. We first state the Fourier Series representation of some time-domain function:

$$f(t) \approx F_k[f(t)] = \sum_{k=-\infty}^{\infty} \mathbf{c}_k e^{ik\omega_0 t}, \tag{10}$$

with coefficients

$$\mathbf{c}_0 = \frac{1}{2L}\int_{-L}^{L} f(t)dt, \tag{11}$$

$$\mathbf{c}_k = \frac{1}{2L}\int_{-L}^{L} f(t)e^{-ik\omega_0 t}dt, \quad \forall k \in \mathbb{N}^+. \tag{12}$$

For even functions, such as stationary periodic kernels, the the Fourier Series only exists at integer multiples of the fundamental periodic frequency $\omega = k\omega_0$ where $k \in \mathbb{N}^+$ and exists only in terms of the cosine-only series from (10). We then evaluate the integral to find the k^{th} coefficient $\mathbf{c}_k$:

$$\begin{aligned}\mathbf{c}_k &= \frac{1}{2L}\int_{-L}^{L} e^{l^{-2}(\cos(\omega_0\tau)-1)}e^{-ik\omega_0\tau}d\tau \\ &= \frac{e^{-l^{-2}}}{2L}\int_{-L}^{L} e^{l^{-2}(\cos(\omega_0\tau))}\cos(k\omega_0\tau)d\tau \\ &= \frac{2\pi I_k(l^{-2})}{e^{l^{-2}}},\end{aligned} \tag{13}$$

where we have used the substitution $\omega_0 = \frac{\pi}{L}$, $L = \pi$, and $I_n(z)$ is the Modified Bessel function of the first kind of integer order n and argument z. The solution is found after noting the special function identity $I_n(z) = \frac{1}{\pi}\int_0^{\pi} e^{z\cos(\theta)}\cos(n\theta)d\theta$ (Abramowitz and Stegun 1972) which allows one to collapse the oscillatory integral into a convergent form.

We now have an approximate representation of the kernel as an infinite Fourier series $\kappa(\tau) \approx F_k[\kappa(\tau)]$:

$$\begin{aligned}\kappa_{perSE}(\tau) &\approx F_k[\kappa(\tau)] \\ &= \sum_{k=-\infty}^{\infty} \frac{I_k(l^{-2})}{\exp(l^{-2})}\cos(k\omega_0\tau).\end{aligned} \tag{14}$$

This can thus be decomposed into a truncated sum $k = \pm 1, \pm 2, ... \pm K$ which admits a decomposable form in the same way as standard Random Fourier Features from (5) into (6). Thus we have a "convergent" Fourier Feature representation in the sense that there is no randomness in the frequency domain but instead an exponentially converging series.

This leads to the corresponding Fourier Series Features

$$\begin{aligned}\hat{\Phi}(\mathbf{x}) = \Big[&q_k\cos(\mathbf{x}^T k\omega_0), ..., q_K\cos(\mathbf{x}^T K\omega_0), \\ &q_k\sin(\mathbf{x}^T k\omega_0), ..., q_K\sin(\mathbf{x}^T K\omega_0)\Big],\end{aligned} \tag{15}$$

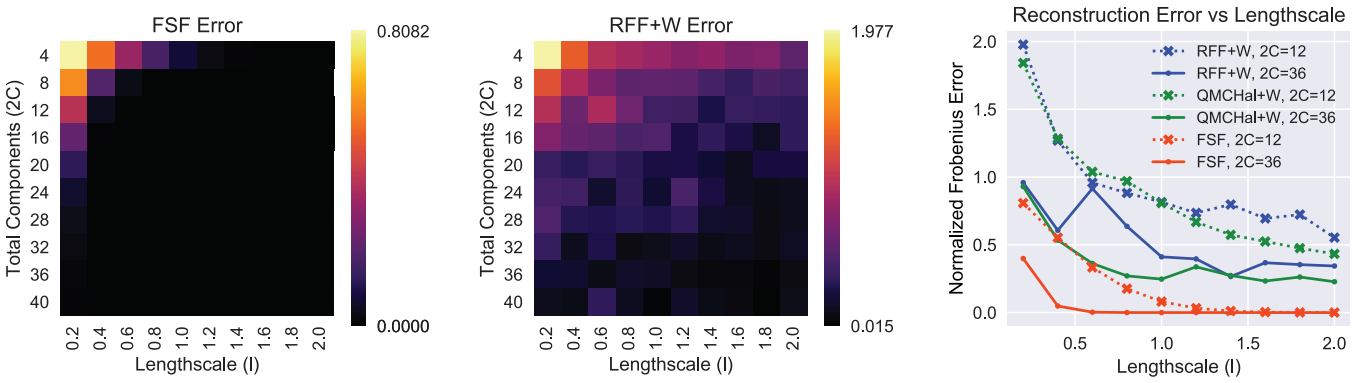

Figure 1: Gram matrix approximation errors for Fourier Series Features (FSF), periodically warped Random Fourier Features (RFF+W) and Quasi-Monte Carlo Fourier Features with Halton sequence (QMCHal+W), as a function of length-scale l and total features $2C$. The rightmost plot depicts comparative slices from the preceding heatmaps. Normalized Frobenius error calculated from $\frac{\|\tilde{\mathbf{K}}-\mathbf{K}\|_F}{\|\mathbf{K}\|_F}$ where $\tilde{\mathbf{K}}$ and $\mathbf{K}$ are respectively the approximate and full Gram matrices. Note we have omitted the heatmap for QMCHal+W due to space.

where

$$q_k^2 = \begin{cases} \frac{I_k(l^{-2})}{(l^{-2})} & \text{if } k = 0, \\ \frac{2I_k(l^{-2})}{(l^{-2})} & \text{if } k = 1, 2, ..., K. \end{cases}$$

This method of constructing a periodic kernel is notable as it demonstrates that periodically warped representations of stationary kernels permit an integral convergence in terms of truncated series of special functions by removing any randomness that was required by the MC sampling for the original stationary kernel. The general process is similarly applicable to periodically warping other isotropic stationary kernels.

Extending the analysis, we note a contrast with (Solin and Särkkä 2014) which approaches the problem with a Taylor series expansion, which is only *locally* convergent. By instead using a Fourier series expansion we benefit from *global* convergence. Following (Stein and Shakarchi 2011), let $S_N(f)(x) = \sum_{-N}^{N} \hat{f}(n) e^{2\pi i n x/L}$ be the N^{th} partial sum of the Fourier series of f, for a positive integer N. Thus we have from the theorem of mean square convergence in (Stein and Shakarchi 2011), the given Lemma 1.2: *If f is integrable on the circle with Fourier coefficients a_n, then $\|f - S_N(f)\| \leq \|f - \sum_{|n|\leq N} c_n e_n\|$ for any complex number c_n.*

3.1 Quality of Kernel Approximation

The clearest way to demonstrate the quality of the kernel approximation is by measuring against the full Gram matrix $\mathbf{K}$ generated by the analytic solution of the kernel. We show the error between $\mathbf{K}$ and the approximated Gram matrix $\tilde{\mathbf{K}}$ with $\tilde{\mathbf{K}}_{ij} = \hat{\kappa}(x_i, x_j)$. Figure 1 demonstrates the normalized Frobenius norm for the approximated kernel against the full SE kernel across multiple length-scales and total number of components. We used $N = 2000$ random values drawn uniformly on the interval $[-2, 2]$ with kernel periodicity $T = 2$, noting the result is representative over larger ranges and higher samples. It is clear that both periodically warped RFFs and state of the art performant low-discrepancy Halton QMC Features (Avron et al. 2016), even with $D = 1$, require significantly more random samples to achieve a similar approximating error norm.

4 Feature Space Compositions

Formulating the periodic feature as (15) permits the kernel to be naturally consolidated into a Compositional Fourier Feature (CFF) architecture allowing one to easily express periodic compositions alongside standard RFF kernel approximations.

The feature space operations *sum* and *cartesian product* are the applicable kernel compositions for our regression framework. In operator notation these are as defined as follows:

$$\begin{aligned}(\kappa_1 + \kappa_2)(\mathbf{x}, \mathbf{x}') &= \kappa_1(\mathbf{x}, \mathbf{x}') + \kappa_2(\mathbf{x}, \mathbf{x}') \\ &= [\hat{\Phi}_1(\mathbf{x}) \hat{\Phi}_2(\mathbf{x})][\hat{\Phi}_1(\mathbf{x}') \hat{\Phi}_2(\mathbf{x}')]^T, \quad (16)\end{aligned}$$

defines the sum or concatenation of the feature maps (6) and,

$$\begin{aligned}(\kappa_1 \times \kappa_2)(\mathbf{x}, \mathbf{x}') &= \kappa_1(\mathbf{x}, \mathbf{x}') \times \kappa_2(\mathbf{x}, \mathbf{x}') \\ &= \sum_i^{n,m} \hat{\Phi}_{1,2}^{(i)}(\mathbf{x}) \hat{\Phi}_{1,2}^{(i)}(\mathbf{x}'). \quad (17)\end{aligned}$$

defines the feature space product of the feature maps (6), where $\hat{\Phi}_{1,2}(\mathbf{x}) = \hat{\Phi}_1 \times \hat{\Phi}_2$ is the Cartesian product. Using compositional Fourier Features, depending on the composition, our data is transformed into a compositional feature map with resulting dimensionality $\mathscr{C}$: $\hat{\mathbf{\Phi}}(\mathbf{x}) \in \mathbb{R}^{N \times \mathscr{C}}$. For instance, it is possible to create the compositional structure as an interpretable string literal: $composition =$ "$(LINEAR + SE) \times PER_1$" which provides an abstract modelling structure which is crucial for human interpretability. By using a custom LALR(1) (DeRemer 1971) grammar, these compositional operations upon the feature space are automatically created and executed at runtime by generating

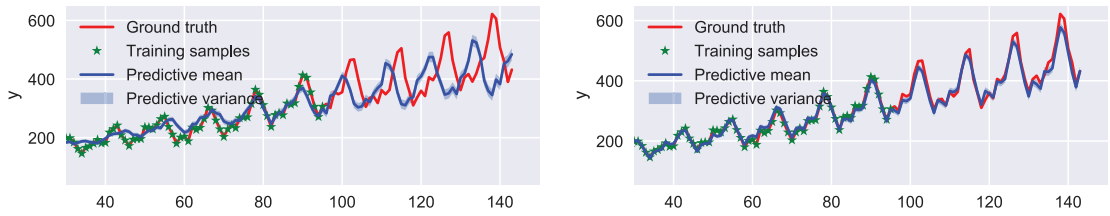

Figure 2: Models learned with constant, misspecified periodic hyperparameter (left), and constant with FFT (right).

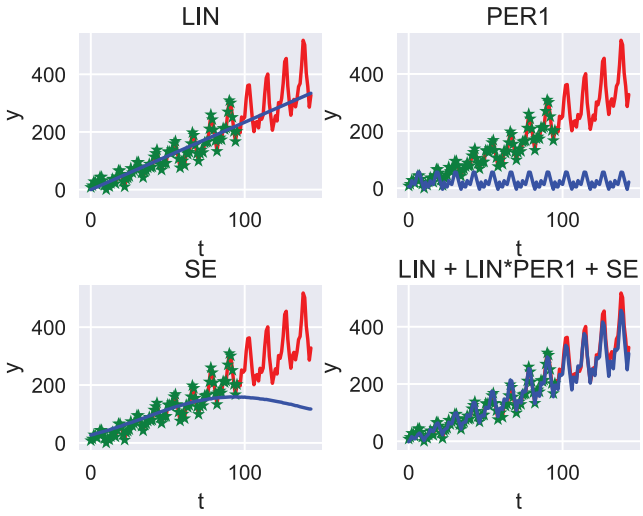

Figure 3: The effect of kernel compositions using Linear, SE, Periodic SE. Training data is marked green, truth with red, and prediction with blue over seen and unseen time-stamps. The top row depicts two separately learned models using just a Linear and Periodic kernel. These are able to individually model the data however can only capture a limited structure. Similarly, the SE at the bottom left can approximately model previously seen time-series data, however when extrapolating, diverges because time is a continuously increasing variable. The bottom right consists of a more elaborate composition which shows how one can capture far more information and then extrapolate more accurately using compositions.

a parse tree and evaluating each operator node. Figure 3 depicts Fourier Series Features alongside conventional features to demonstrate the additional expressiveness that compositions allow.

5 Variational Bayesian Linear Regression for Fourier Features

One of the primary advantages of using Fourier Feature approximation is it allows us to estimate a function in the RKHS as a linear function in the dual space of $\hat{\Phi}(\mathbf{x})$ instead of the primal space of $\kappa(\cdot,\cdot)$ which is typically restricted by large matrix inversions. We thus utilise Bayesian Linear Regression (Murphy 2012) and factorise over latent variables

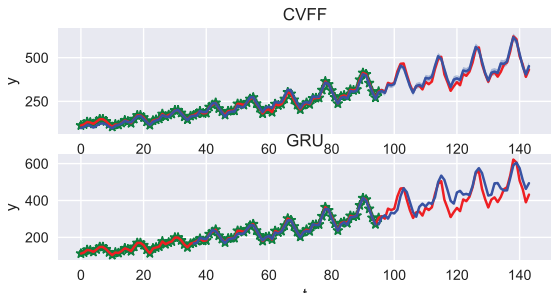

Figure 4: Comparison between our method and GRUs for long term forecasting.

including the linear model's weights $\mathbf{w}$, intercept b and approximating kernel hyperparameters.

5.1 Variational Inference

Variational Inference (VI) (Blei, Kucukelbir, and McAuliffe 2017) is a term that describes methodologies for determining probabilistic posterior inference through tractable optimisation. Fundamentally, it consists of two parts: 1. Assume an approximating distribution $q(\mathbf{z};\lambda)$ over latent variables, and 2. Optimise over the parameters λ to bring the variational distribution $q(\mathbf{z};\lambda)$ closer to the true posterior $p(\mathbf{z}|\mathbf{x})$. Thus the posterior is approximated through minimizing some divergence measure:

$$\lambda^* = \underset{\lambda}{\operatorname{argmin}}\ Div(p(\mathbf{z}|\mathbf{x}), q(\mathbf{z};\lambda)). \tag{18}$$

Typically the posterior is intractable and so VI aims to learn the approximate generating model instead. One of the ways to minimize divergence is by using the Kullback-Leibler (KL) from $q(\mathbf{z};\lambda)$ to $p(\mathbf{z} \mid \mathbf{x})$,

$$\lambda^* = \arg\min_{\lambda} \mathrm{KL}(q(\mathbf{z};\lambda) \parallel p(\mathbf{z} \mid \mathbf{x})) \tag{19}$$

$$= \arg\min_{\lambda}\ \mathbb{E}_{q(\mathbf{z};\lambda)}\big[\log q(\mathbf{z};\lambda) - \log p(\mathbf{z} \mid \mathbf{x})\big]. \tag{20}$$

The form of the problem in (20) depends on the posterior and is therefore intractable, however one can instead take advantage of the property

$$\log p(\mathbf{x}) = \mathrm{KL}(q(\mathbf{z};\lambda) \parallel p(\mathbf{z} \mid \mathbf{x})) + \mathbb{E}_{q(\mathbf{z};\lambda)}\big[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\lambda)\big],$$

where the left hand side is the logarithm of the marginal likelihood and $p(\mathbf{x}) = \int p(\mathbf{x},\mathbf{z})\mathrm{d}\mathbf{z}$ is termed the model evidence.

This evidence is a constant relative to the variational parameters λ, allowing one to minimize $\mathrm{KL}(q\|p)$ by maximizing the *Evidence Lower Bound* (ELBO),

$$\mathrm{ELBO}(\lambda) = \mathbb{E}_{q(\mathbf{z};\lambda)}\big[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z};\lambda)\big].$$

Both $p(\mathbf{x},\mathbf{z})$ and $q(\mathbf{z};\lambda)$ are tractable within the ELBO and thus we have an optimisable objective:

$$\lambda^* = \arg\max_{\lambda} \mathrm{ELBO}(\lambda).$$

There are various ways to perform this optimization, however we adopt the approach from (Kingma and Welling 2014) which allows convenient reparameterisations of distributions, allowing automatic differentiation approaches to follow the variational distributions' gradients:

$$\nabla_\lambda \ \mathrm{ELBO}(\lambda) =$$
$$\mathbb{E}_{q(\epsilon)}\big[\nabla_\lambda\big(\log p(\mathbf{x},\mathbf{z}(\epsilon;\lambda)) - \log q(\mathbf{z}(\epsilon;\lambda);\lambda)\big)\big],$$

in which the gradient of the ELBO is an expectation over some base distribution $q(\epsilon)$ which does not rely on the variational parameters.

5.2 Bayesian Linear Regression Model

We posit the model as a fully Bayesian linear regression model with Automatic Relevance Determination (ARD) prior following (Drugowitsch 2013). Our data consists of N samples each of dimensionality D: $\mathbf{x} \in \mathbb{R}^{N\times D}$ and corresponding outputs $\mathbf{y} \in \mathbb{R}^N$. With compositional dimension defined from 4 we have the model:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\alpha}^{-1}\mathbf{I})$$
$$p(\mathbf{y}|\mathbf{w}, \hat{\boldsymbol{\Phi}}(\mathbf{x})) = \prod_{n=1}^{N} \mathcal{N}(y_n|\hat{\boldsymbol{\Phi}}(\mathbf{x})_n^T\mathbf{w}, \boldsymbol{\beta}^{-1}). \qquad (21)$$

We place Gamma hyper-priors α and β with Log-Normal variational posteriors on the likelihood and weight inverse variances respectively. Normal priors with Log-Normal variational posteriors over all standard hyperparameters $\boldsymbol{\theta}$ are used for the approximating kernel hyperparameters which are globally termed $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w}, \mathbf{T}, \mathbf{l}]$ containing likelihood and weight precisions $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, regression weights $\mathbf{w}$, periodic and lengthscale hyperparameters $\mathbf{T}$, $\mathbf{l}$ respectively. Inference is then performed within a probabilistic programming framework (Tran et al. 2017).

5.3 Periodic Hyperparameter Learning

In related works there is often an ad-hoc method for choosing (as constant) or initialising (random seed) the periodic hyperparameter T. We propose it is justified and straightforward to apply the Fast Fourier Transform (FFT) (Weisstein 2004) to extract the most significant frequencies from available training data. If the goal of modelling and inference is to capture important periodicites then the FFT is the natural method to easily expose important fundamental frequencies within the time-series. Figure 5 shows these extracted periods on the Airline and Melbourne Daily Temps datasets and we demonstrate empirically in Table 2 the benefit of initialising the periodic hyperparameters with the FFT in contrast to random sampling. These values are then learned via the aforementioned system as another parameter within $\boldsymbol{\theta}$

6 Experiments

In this section we present various model evaluations first focusing on the periodic hyperparameters, performance in general, and then contrast to classical methods and recent recurrent neural networks. The datasets vary in size from 144 to 39432 samples. With the Airline dataset we train on the first 8 years and predict the last 4. For the remaining datasets we train on the first 80% of the data and test on the remaining 20%. Running times in Table 1 demonstrate how GP methods break down with larger samples while our proposed method scales tractably.

6.1 Periodic Hyperparameter Evaluation

We investigate here the significance of misspecification of the periodic kernel hyperparameter, i.e. random vs FFT. We demonstrate a more suitable initialization procedure by taking the real FFT of the (training) signal and using the top P periods to seed kernel hyperparameter T before variational learning. We posit that carrying out an initial FFT is a natural method to expose the underlying data's latent periodicity. This analysis highlights how crucial it is for the correct to seed the latent periods well. Figure 2 demonstrates how an accurately seeded hyperparameter allows the model to capture periodicity in the data than blind initialisation.

While Figure 2 demonstrates it may be sufficient to approximately specify the periodicity as constant using methods like the FFT, this brings other pitfalls such as sampling artifacts. It is important to recognise the FFT will always produce sub-optimal results due to the nature of data sampling never being truly perfect and hyperparameter optimisation allows one to overcome such misspecifications.

6.2 Carbon Dioxide Levels

We test on the classic Mauna Loa from 1965 to more recent readings in the beginning of 2017 (Keeling et al. 2017). This dataset has been examined in great detail in the past (Rasmussen and Williams 2006; Duvenaud et al. 2013) and so provides a good baseline for validating our methodology.

6.3 Airline Passengers

Consisting of 144 samples this data depicts airline passenger numbers from 1949 to 1961 (Box, Jenkins, and Reinsel 1976). The time-series here exhibits an increasing trend over time with observably constant periodicity - traits which can intuitively be expressed in a compositional grammar of kernel functions: $LINEAR + LINEAR * (PER_1 + PER_2 + PER_3) + (SE_2) * (PER_1 + PER_2 + PER_3)$. This symbolic form expresses a broad structural belief over the signal and following humanly interpretable natural language description from (Lloyd et al. 2014) the composition may be interpreted as: "a linearly trending function, with linearly increasing periodic amplitude and locally periodic components". The CVFF regression model is able to learn this series for an accurate extrapolation while The RNN models appear to capture local patterns but are unable to learn longer term structures and degenerate quickly outside of the forecast window due to the propagation of errors over time without having access to new data.

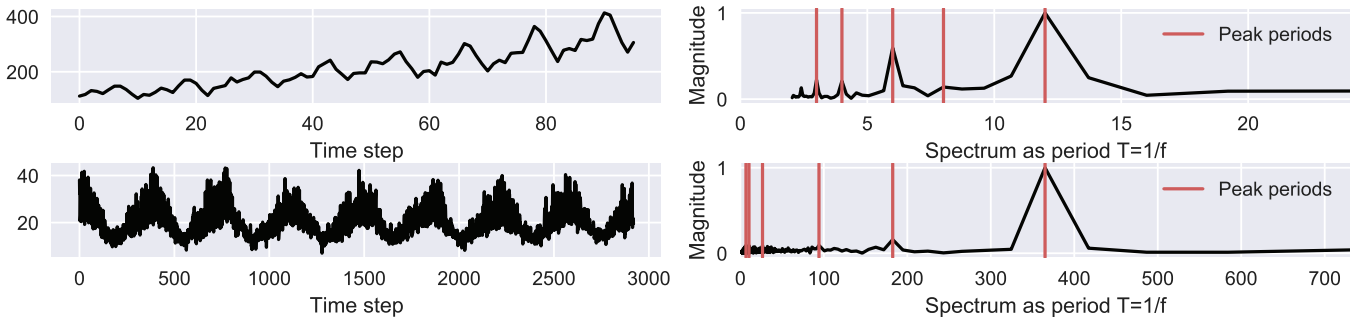

Figure 5: Time-series (left) with corresponding normalized spectrums (right) with primary periodicities by magnitude.

	RMSE					SMSE				
Dataset	CFGP	VRFF	CVFF	LSTM	GRU	CFGP	VRFF	CVFF	LSTM	GRU
Airline	17.726 (0m 10s)	251.651	**16.984** (0m 30s)	373.031	50.715	0.052	10.496	**0.048**	22.612	0.421
CO2	**2.396** (1m 19s)	3.6252	2.433 (1m 1s)	4.421	6.36	**0.111**	0.253	0.114	0.392	0.812
Melbourne	4.681 (11m 58s)	6.064	**4.201** (2m 12s)	11.612	6.347	0.607	0.977	**0.469**	3.581	1.071
Zone Temps	NA	16.853	**12.082** (4m 20s)	23.319	17.864	NA	0.988	**0.788**	1.894	1.117

Table 1: Average performance in RMSE and SMSE for standard Random Fourier Features using Compositional Full GPs (CFGP), Variational RFF with RBF (VRFF), our Compositional Variational Fourier Features (CVFF), and LSTM and GRU recurrent neural networks. Running times for equivalent compositions with full GP and FF methods are provided.

Initialisation	P	Airline	Melbourne
Uniform	1	135.26	6.21
	2	225.4	8.28
	3	65.85	7.04
FFT	1	**55.52**	**4.16**
	2	**33.56**	**4.2**
	3	**24.88**	**4.21**

Table 2: Comparison of RMSE with our method for an increasing number of latent periodicities. Each P^{th} FFT component is selected in decreasing magnitude. We observe that when the periodicity hyperparameter is seeded randomly, there is no apparent improvement in performance. Conversely, even a single primary periodicity seeded by the FFT can significantly improve performance.

6.4 Melbourne Daily Average Temperatures

The dataset (BOM 2014) contains daily temperatures from Melbourne, Australia and represents a more challenging problem than the previous two dataset in two ways. First, it has many more samples at around 3000 instead of in the hundreds, and secondly contains a lot of high frequency information. We show that by combining Periodic Fourier Features with standard features we are able to plausibly model the data into the future and further discover long and short term periodicities very quickly using the FFT.

6.5 Smart Grid Hourly Temperatures

This dataset is from a 2012 Kaggle competition (GEFCom 2012) and consists of an 11-dimensional time-series with 39432 samples. We focus on a single zone with identification number 3. Similar to the Melbourne Temps dataset, this data exhibits extremely high noise and both short and long term periodicities. While our model is able to capture the general underlying trend of the data and extrapolate into the future, short term patterns are not captured effectively. This can be explained by the FFT selecting the stronger *low* frequency components and ignoring higher frequency oscillatory behaviour. Comparing against the RNN methods one can see the model does not degenerate for longer term extrapolations as it does not rely on more recent observations.

7 Conclusion

In this paper we have described how to integrate periodic transformations of the standard SE kernel into the Fourier Feature framework while also showing that it requires very few features in practice to achieve convergent downstream behaviour. We have further shown how it is possible to integrate these periodic Fourier Features into a composition framework defined by an interpretable grammar which has the added benefit of being far more humanly intuitive and interpretable than alternative methods for modelling temporal patterns such as RNNs. By adopting a Bayesian parameterization of the kernel hyperparameters, as well as a more principled way of initialising periodic hyperparameters and optimizing them jointly over their variational distributions, we can simultaneously scale learning with stochastic optimization while avoiding overfitting and providing predictive uncertainty for extrapolations. Avenues for extending this work include defining the compositional *grammar* itself in a fully probabilistic manner such as in (Malkomes, Schaff, and Garnett 2016) and optimizing these jointly in fully or partially hierarchical manner.

References

Abramowitz, M., and Stegun, I. A. 1972. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. book section "Modified Bessel Functions I and K." 9.6, 374–377.

Avron, H.; Sindhwani, V.; Yang, J.; and Mahoney, M. W. 2016. Quasi-monte carlo feature maps for shift-invariant kernels. *Journal of Machine Learning Research* 17(120):1–38.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*.

Bochner, S. 1933. Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen* 108:378–410.

BOM. 2014. Daily maximum temperatures in melbourne, australia, 1981-1990. Available from https://datamarket.com/data/set/2323/daily-maximum-temperatures-in-melbourne-australia-1981-1990.

Box, G. E. P.; Jenkins, G. M.; and Reinsel, G. C. 1976. *Time Series Analysis, Forecasting and Control*. Third edition edition.

DeRemer, F. L. 1971. Simple lr (k) grammars. *Communications of the ACM* 14(7):453–460.

Drugowitsch, J. 2013. Variational bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*.

Duvenaud, D.; Lloyd, J. R.; Grosse, R.; Tenenbaum, J. B.; and Ghahramani, Z. 2013. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*.

GEFCom. 2012. Global energy forecasting competition 2012 - load forecasting. Available from https://www.kaggle.com/c/global-energy-forecasting-competition-2012-load-forecasting.

Ghahramani, Z. 2005. Non-parametric Bayesian methods. Available from mlg.eng.cam.ac.uk/zoubin/talks/uai05tutorial-b.pdf.

Ghassemi, N. H., and Deisenroth, M. 2014. Analytic long-term forecasting with periodic gaussian processes. In *Proc. of AISTATS*, 303–311.

Gönen, M., and Alpaydın, E. 2011. Multiple kernel learning algorithms. *Journal of machine learning research* 12(Jul):2211–2268.

Keeling, R. F.; Walker, S. J.; Piper, S. C.; and Bollenbacher, A. F. 2017. Atmospheric CO2 concentrations (ppm) derived from in situ air measurements at mauna loa. Available from http://scrippsco2.ucsd.edu.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational Bayes.

Klenske, E. D.; Zeilinger, M. N.; Schölkopf, B.; and Hennig, P. 2016. Gaussian process-based predictive control for periodic error correction. *IEEE Transactions on Control Systems Technology* 24(1):110–121.

Le, Q.; Sarlós, T.; and Smola, A. 2013. Fastfood-approximating kernel expansions in loglinear time. In *Proceedings of the international conference on machine learning*, volume 85.

Lloyd, J. R.; Duvenaud, D. K.; Grosse, R. B.; Tenenbaum, J. B.; and Ghahramani, Z. 2014. Automatic construction and natural-language description of nonparametric regression models. In *AAAI*, 1242–1250.

MacKay, D. J. 1998. Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences* 168:133–166.

Malkomes, G.; Schaff, C.; and Garnett, R. 2016. Bayesian optimization for automated model selection. In *Advances in Neural Information Processing Systems*, 2900–2908.

Murphy, K. P. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Rahimi, A., and Recht, B. 2007. Random features for large-scale kernel machines. *Conference on Neural Information Processing Systems*.

Rainforth, T.; Le, T. A.; van de Meent, J.-W.; Osborne, M. A.; and Wood, F. 2016. Bayesian optimization for probabilistic programs. In *Advances in Neural Information Processing Systems*, 280–288.

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

Roberts, S.; Osborne, M.; Ebden, M.; Reece, S.; Gibson, N.; and Aigrain, S. 2013. Gaussian processes for time-series modelling. *Phil. Trans. R. Soc. A* 371(1984):20110550.

Schölkopf, B., and Smola, A. J. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. the MIT Press.

Schulz, E.; Tenenbaum, J.; Duvenaud, D. K.; Speekenbrink, M.; and Gershman, S. J. 2016. Probing the compositionality of intuitive functions. In *Advances in neural information processing systems*, 3729–3737.

Senanayake, R.; Simon Timothy, O.; and Ramos, F. 2016. Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression. In *AAAI*, 3901–3907.

Solin, A., and Särkkä, S. 2014. Explicit link between periodic covariance functions and state space models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, 904–912.

Stein, E. M., and Shakarchi, R. 2011. *Fourier analysis: an introduction*, volume 1. Princeton University Press.

Tran, D.; Kucukelbir, A.; Dieng, A. B.; Rudolph, M.; Liang, D.; and Blei, D. M. 2016. Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

Tran, D.; Hoffman, M. D.; Saurous, R. A.; Brevdo, E.; Murphy, K.; and Blei, D. M. 2017. Deep probabilistic programming. *arXiv preprint arXiv:1701.03757*.

Weisstein, E. W. 2004. Fourier transform. Available from http://mathworld.wolfram.com/FourierTransform.html.

Yang, Z.; Wilson, A.; Smola, A.; and Song, L. 2015. A la carte–learning fast kernels. In *Artificial Intelligence and Statistics*, 1098–1106.