Shen, Zheyang; Heinonen, Markus; Kaski, Samuel

Learning spectrograms with convolutional spectral kernels

Published in:
The 23rd International Conference on Artificial Intelligence and Statistics

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Shen, Z., Heinonen, M., & Kaski, S. (2020). Learning spectrograms with convolutional spectral kernels. In S. Chiappa, & R. Calandra (Eds.), *The 23rd International Conference on Artificial Intelligence and Statistics* (pp. 3826-3836). (Proceedings of Machine Learning Research; Vol. 108). JMLR.
http://proceedings.mlr.press/v108/shen20a.html

Learning spectrograms with convolutional spectral kernels

Zheyang Shen **Markus Heinonen** **Samuel Kaski**
Helsinki Institute for Information Technology, HIIT,
Department of Computer Science, Aalto University

Abstract

We introduce the *convolutional spectral kernel* (CSK), a novel family of non-stationary, nonparametric covariance kernels for Gaussian process (GP) models, derived from the convolution between two imaginary radial basis functions. We present a principled framework to interpret CSK, as well as other deep probabilistic models, using approximated Fourier transform, yielding a concise representation of input-frequency *spectrogram*. Observing through the lens of the spectrogram, we provide insight on the interpretability of deep models. We then infer the functional hyperparameters using scalable variational and MCMC methods. On small- and medium-sized spatiotemporal datasets, we demonstrate improved generalization of GP models when equipped with CSK, and their capability to extract non-stationary periodic patterns.

1 Introduction

Gaussian processes (GP), as rich distributions over functions, are a cornerstone of a wide array of probabilistic modeling paradigms, thanks largely to their tractability, flexibility, robustness to overfitting and principled quantification of uncertainty (Rasmussen and Williams, 2006). At the helm of every GP model lies the *covariance kernel*, a function depicting its covariance structure and encoding prior knowledge.

Despite their affinity to neural networks (Williams, 1997; Lee et al., 2018), GP models seldom exhibit the generalization of the former due to the innate rigidity of the widely used squared exponential (SE) kernel, rendering them insufficient for genuine pattern recognition. While myriad studies (Paciorek and Schervish, 2004; Alvarez et al., 2009; Wilson and Adams, 2013; Duvenaud et al., 2013; Tobar et al., 2015; Wilson et al., 2016; Remes et al., 2017; Tobar, 2018; Shen et al., 2019) have sought more expressive kernel choices, their efforts fall notably short on (i) flexibility, (ii) interpretability or (iii) scalability, all of which are essential to large-scale data analysis. In this work, we propose and analyze a novel kernel family satisfying the above properties.

We propose the *convolutional spectral kernel* (CSK), a novel kernel family with both spatially varying lengthscales and frequencies, derived from the convolution of two complex radial basis functions. We demonstrate that CSK possesses superior flexibility, unifying the monotonic non-stationary quadratic (NSQ) kernel (Paciorek and Schervish, 2004) and the stationary spectral mixture (SM) kernel (Wilson and Adams, 2013).

We introduce the notion of the *spectrogram* as a new, principled framework to interpret nonparametric kernels. The spectrogram is a joint distribution of input and frequency, conveniently displaying local covariance patterns. Our analysis shows that CSK retains an unbiased description of the instantaneous frequency, as opposed to the similar generalized spectral mixture (GSM) kernel (Remes et al., 2017). Meanwhile, our analysis sheds light on previously un-interpretable state-of-the-art deep probabilistic models, namely deep GPs (Damianou and Lawrence, 2013; Salimbeni and Deisenroth, 2017; Havasi et al., 2018) and the deep kernel learning (Wilson et al., 2016), and justifies the adoption of certain heuristics in the said models.

We introduce scalable inference schemes for GP models equipped with CSK, which combine sparse GPs (Snelson and Ghahramani, 2006; Titsias, 2009; Hensman et al., 2017), stochastic gradient Hamiltonian Monte Carlo (Neal, 1993; Chen et al., 2014), and moving window MCEM (Havasi et al., 2018). Our method can be extended to covariance function deep GPs (Dunlop et al., 2018), a hierarchical generalization of our current model.

Empirically, we provide evidence from synthetic and

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

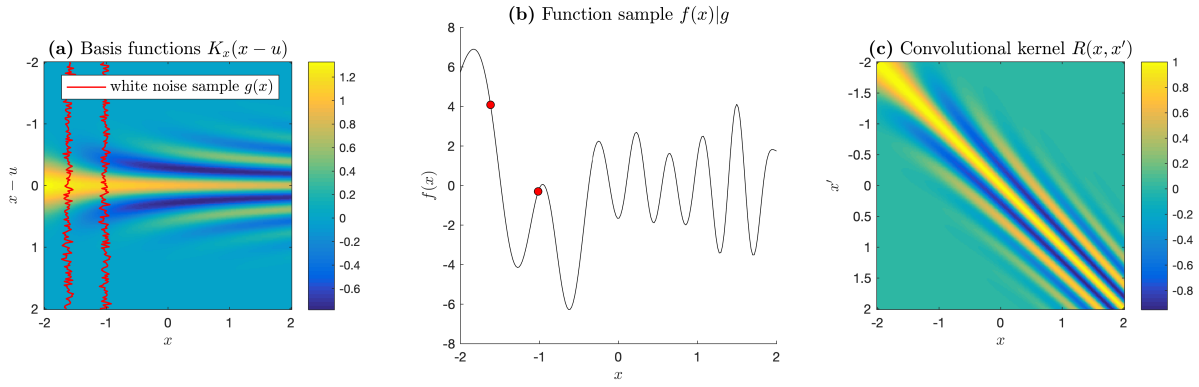

Figure 1: Visualization of the **(a)** basis function and convolution with a white noise process; **(b)** a sample from the resulting GP, where the highlighted points are computed from a convolution between the moving white noise process, denoted by the jagged lines in **(a)**; **(c)** the kernel matrix of the resulting CSK.

real-world spatiotempral datasets to support our theoretical claims. Our experiments visually and numerically demonstrate interpretable pattern extraction and superior predictive performance of the CSK-GP model.

2 Convolutional spectral kernel (CSK)

In this section, we derive the *convolutional spectral kernel*, a non-stationary, nonparametric kernel, interpretable through its local lengthscale, frequency and variance functions. Throughout the discussion of this paper, we assume a simple regression task: the objective is to infer a scalar function $f(\mathbf{x}) \in \mathbb{R}$ with D-dimensional inputs $\mathbf{x} \in \mathbb{R}^D$, with a finite supply of N observed data points as a data matrix $\mathbf{X} \in \mathbb{R}^{N\times D}$, and a set of noisy observations $\mathbf{y} \in \mathbb{R}^N$. We assume the function f is a realization of some underlying zero-mean Gaussian process, with homoskedastic observation noise of precision β,

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), \tag{1}$$

$$y = f(\mathbf{x}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \beta^{-1}). \tag{2}$$

Our construction of CSK is inspired by the construction of non-stationary kernels with spatially-varying lengthscales (Gibbs, 1998; Higdon et al., 1998; Paciorek and Schervish, 2004). We propose a novel feature map $K_{\mathbf{x}}(\mathbf{u}) \in \mathbb{C}$ of complex-valued radial bases:

$$K_{\mathbf{x}_i}(\mathbf{u}) = e^{-\frac{\left\|\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{u}-\mathbf{x}_i)\right\|^2}{2}} \exp(\imath\langle\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i, \mathbf{u}-\mathbf{x}_i\rangle)$$
$$\propto \mathcal{N}\left(\mathbf{u}|\mathbf{x}_i + \imath\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right). \tag{3}$$

Here we abuse the notation of a multivariate normal density, where $\imath$ denotes the imaginary unit, $\mathbf{u} \in \mathbb{R}^D$ denotes a point in input space, and the $\boldsymbol{\mu}_i := \boldsymbol{\mu}(\mathbf{x}_i) \in \mathbb{R}^D$, $\boldsymbol{\Sigma}_i := \boldsymbol{\Sigma}(\mathbf{x}_i) \in \mathbb{R}^{D\times D}_{\succeq \mathbf{0}}$ are vector- and positive-semidefinite matrix-valued functions of $\mathbf{x}_i$ denoting the frequency and covariance parameters of the input space, from which we can construct the *frequency* as an inverse product $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Viewing GPs as continuously-indexed moving average processes, the feature map (3) denotes a potentially infinite window (Tobar et al., 2015). We can henceforth represent a GP as a convolution between $K_{\mathbf{x}_i}$ and a white noise process $g(\mathbf{x}) \sim \mathcal{GP}(0, \delta_{\mathbf{x}=\mathbf{x}'})$ (Higdon et al., 1998):

$$f(\mathbf{x}_i) = \int K_{\mathbf{x}_i}(\mathbf{u}) g(\mathbf{u})\, \mathrm{d}\mathbf{u}. \tag{4}$$

The kernel of f is the Hermitian inner product between $K_{\mathbf{x}_i}(\mathbf{u})$ and $K_{\mathbf{x}_j}(\mathbf{u})$, which is solved analytically:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \int K_{\mathbf{x}_i}(\mathbf{u})\overline{K_{\mathbf{x}_j}(\mathbf{u})}\, \mathrm{d}\mathbf{u}$$
$$\propto \mathcal{N}\left(\mathbf{x}_i - \mathbf{x}_j|\imath(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j), \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right). \tag{5}$$

Here $\overline{K_{\mathbf{x}_i}(\cdot)}$ denotes the complex conjugate. The solution to this integral is detailed in Section 1 of the appendix. We obtain a non-stationary correlation function after normalization:

$$R(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathrm{Re}\left(k(\mathbf{x}_i, \mathbf{x}_j)\right)}{\sqrt{k(\mathbf{x}_i, \mathbf{x}_i)k(\mathbf{x}_j, \mathbf{x}_j)}}$$
$$= \sigma_{ij}\ e^{-\frac{Q_{ij}+S_{ij}}{2}} \cos\langle\boldsymbol{\omega}_{ij}, \mathbf{x}_i - \mathbf{x}_j\rangle, \tag{6}$$

$$\sigma_{ij} = \frac{|\boldsymbol{\Sigma}_i|^{1/4}|\boldsymbol{\Sigma}_j|^{1/4}}{|(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)/2|^{1/2}}, \tag{7}$$

$$Q_{ij} = \left\|\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)^{-1/2}(\mathbf{x}_i - \mathbf{x}_j)\right\|^2, \tag{8}$$

$$\boldsymbol{\omega}_{ij} = \left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)^{-1}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j), \tag{9}$$

$$S_{ij} = \left\|\left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\right)^{-1/2}(\boldsymbol{\omega}_{ii} - \boldsymbol{\omega}_{jj})\right\|^2. \tag{10}$$

We can see from the cosine term in (6) that CSK computes *pairwise frequencies* $\boldsymbol{\omega}_{ij}$ for each pair of data

points, and the exponential terms include squared Mahalanobis distances between $\mathbf{x}_i$ and $\mathbf{x}_j$ (8), and between local frequencies $\boldsymbol{\omega}_{ii}$ and $\boldsymbol{\omega}_{jj}$ (10).

CSK unifies two generalizations of the SE kernel. Paciorek and Schervish (2004) generalize the SE kernel by allowing lengthscales to spatially vary, which is a special case of CSK when $\boldsymbol{\mu}_i \equiv \mathbf{0}$. Formally,

$$k_{NS}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_{ij} e^{-\frac{Q_{ij}}{2}} \propto \mathcal{N}\left(\mathbf{x}_i | \mathbf{x}_j, \boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right). \quad (11)$$

The spectral mixture kernel (Wilson and Adams, 2013) generalizes the SE kernel by allowing for non-zero frequency mean, which is a special case of CSK when functions $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are kept constant:

$$k_{SM}(\mathbf{x}_i, \mathbf{x}_j) = k_{SE}(\mathbf{x}_i, \mathbf{x}_j) \cos\langle \boldsymbol{\mu}, \mathbf{x}_i - \mathbf{x}_j \rangle. \quad (12)$$

With the reasonable assumption of a smooth and slow-varying frequency mean $\boldsymbol{\mu}_i$, CSK (6) identifies one *mostly stationary covariance substructure* of the data. While a dataset might exhibit behaviors such as multiple frequencies or spatially varying variances, such behaviors can be accounted for by stacking multiple CSKs multiplied by a standard deviation function $\sigma_p(\cdot) \in \mathbb{R}_{\geq 0}$:

$$k_{\mathrm{CS}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{p=1}^{P} \sigma_p(\mathbf{x}_i) \sigma_p(\mathbf{x}_j) R_p(\mathbf{x}_i, \mathbf{x}_j). \quad (13)$$

CSK is defined through the component functions $\sigma_p(\cdot)$, $\boldsymbol{\Sigma}_p(\cdot)$, and $\boldsymbol{\mu}_p(\cdot)$. We denote the vector of functional parameters as $\boldsymbol{\theta}$, and each functional parameter $\theta_d(\mathbf{x})$ has a warped GP (Snelson et al., 2004) prior:

$$\theta_d(\mathbf{x}_i) = F_d(h_d(\mathbf{x}_i)), \quad (14)$$

$$h_d(\mathbf{x}_i) \sim \mathcal{GP}(0, k_d(\mathbf{x}_i, \mathbf{x}_j)), \quad (15)$$

For simplification, we assume diagonal covariances: $\boldsymbol{\Sigma}_p = \mathrm{diag}(\ell_{p1}^2, \cdots \ell_{pD}^2)$. The warping function F_d ensures the CSK to be positive definite.

3 The spectrogram

This section coins the notion of *spectrogram*, a joint input-frequency distribution, as a principled framework to interpret typical nonparametric kernels encountered in GP models regardless of input dimensions, which often lacks interpretability.

In signal processing and time-series analysis, the Wigner transform (Flandrin, 1998) converts covariance functions into quasi-probability distributions between input and frequency via a Fourier transform:

$$W(\mathbf{x}, \boldsymbol{\omega}) = \int_{\mathbb{R}^D} k\left(\mathbf{x} + \frac{\boldsymbol{\tau}}{2}, \mathbf{x} - \frac{\boldsymbol{\tau}}{2}\right) e^{-2\imath\pi\boldsymbol{\omega}^\top \boldsymbol{\tau}} \mathrm{d}\boldsymbol{\tau}. \quad (16)$$

The Wigner distribution function (WDF) $W(\mathbf{x}, \boldsymbol{\omega})$ is a generalized probability distribution that retains instantaneous spectral density on all inputs. Despite their potential in interpretation, few machine learning models (Shen et al., 2019) apply the transform (16) due to its intractability.

In this work, we consider all kernels taking the form of a generalized mixture of Gaussian characteristic functions, as indexed by $\mathscr{A}$:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{a \in \mathscr{A}} \sigma_{ij}^{(a)} e^{-\frac{D_{ij}^{(a)}}{2}} \exp(\imath U_{ij}). \quad (17)$$

Here the σ, D and U are real-valued functions of $\mathbf{x}_i$ and $\mathbf{x}_j$. A linearized approximation of the above functions gives an estimate of the kernel value, where the variables $\mathbf{x}$ and $\boldsymbol{\tau}$ are separated:

$$k^{(a)}\left(\mathbf{x} + \frac{\boldsymbol{\tau}}{2}, \mathbf{x} - \frac{\boldsymbol{\tau}}{2}\right) \approx \sigma_{\mathbf{xx}}^{(a)} e^{-\frac{1}{2}\boldsymbol{\tau}^\top \boldsymbol{\Lambda}_{\mathbf{x}}^{(a)} \boldsymbol{\tau}} \exp(\imath \langle \boldsymbol{\xi}_{\mathbf{x}}^{(a)}, \boldsymbol{\tau} \rangle), \quad (18)$$

$$\boldsymbol{\Lambda}_{\mathbf{x}}^{(a)} = \mathcal{H}_{\mathbf{t}}\ D_{\mathbf{x}+\mathbf{t}/2, \mathbf{x}-\mathbf{t}/2}\big|_{\mathbf{t}=\mathbf{0}}, \quad (19)$$

$$\boldsymbol{\xi}_{\mathbf{x}}^{(a)} = \mathcal{J}_{\mathbf{t}}\ U_{\mathbf{x}+\mathbf{t}/2, \mathbf{x}-\mathbf{t}/2}\big|_{\mathbf{t}=\mathbf{0}}. \quad (20)$$

$\mathcal{H}_{\mathbf{t}}$ and $\mathcal{J}_{\mathbf{t}}$ in (19) and (20), respectively, denote the Hessian and Jacobian operators with respect to the variable $\mathbf{t}$. We can henceforth approximate the WDF with the kernel estimate (18):

$$\widehat{W}(\mathbf{x}, \boldsymbol{\omega}) = \sum_{a \in \mathscr{A}} \sigma_{\mathbf{xx}}^{(a)} \mathcal{N}\left(\boldsymbol{\omega} \,\middle|\, \frac{\boldsymbol{\xi}_{\mathbf{x}}^{(a)}}{2\pi}, \frac{\boldsymbol{\Lambda}_{\mathbf{x}}^{(a)}}{2\pi^2}\right). \quad (21)$$

Our approximation is exact for both stationary and harmonizable spectral kernels (Wilson and Adams, 2013; Shen et al., 2019). The rest of the section delineates the significance of our method in interpreting CSK

Kernel	form	$\boldsymbol{\xi}_{\mathbf{x}}$	$\boldsymbol{\Lambda}_{\mathbf{x}}$	reference
Spectral mixture	(12)	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}^{-1}$	Wilson et al. (2015)
Non-stationary quadratic	(11)	$\mathbf{0}$	$\frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$	Paciorek and Schervish (2004)
Generalized spectral mixture	(23)	$\boldsymbol{\mu}_{\mathbf{x}} + (\mathcal{J}_{\mathbf{x}}\boldsymbol{\mu}_{\mathbf{x}})\,\mathbf{x}$	$\frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$	Remes et al. (2017)
Convolutional spectral kernel	(6)	$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}$	$\frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \mathcal{J}_{\mathbf{x}}\left(\boldsymbol{\omega}_{\mathbf{xx}}\right)^\top \boldsymbol{\Sigma}_i \mathcal{J}_{\mathbf{x}}\left(\boldsymbol{\omega}_{\mathbf{xx}}\right)\right)$	current work
DGP-SE	(26)	$\mathbf{0}$	$\mathcal{J}_{\mathbf{x}}\mathbf{f}_{L-1}^\top \boldsymbol{\Sigma}^{-1} \mathcal{J}_{\mathbf{x}}\mathbf{f}_{L-1}$	Damianou and Lawrence (2013)

Table 1: Spectrogram function parameters for various non-stationary covariance function and compositional DGP.

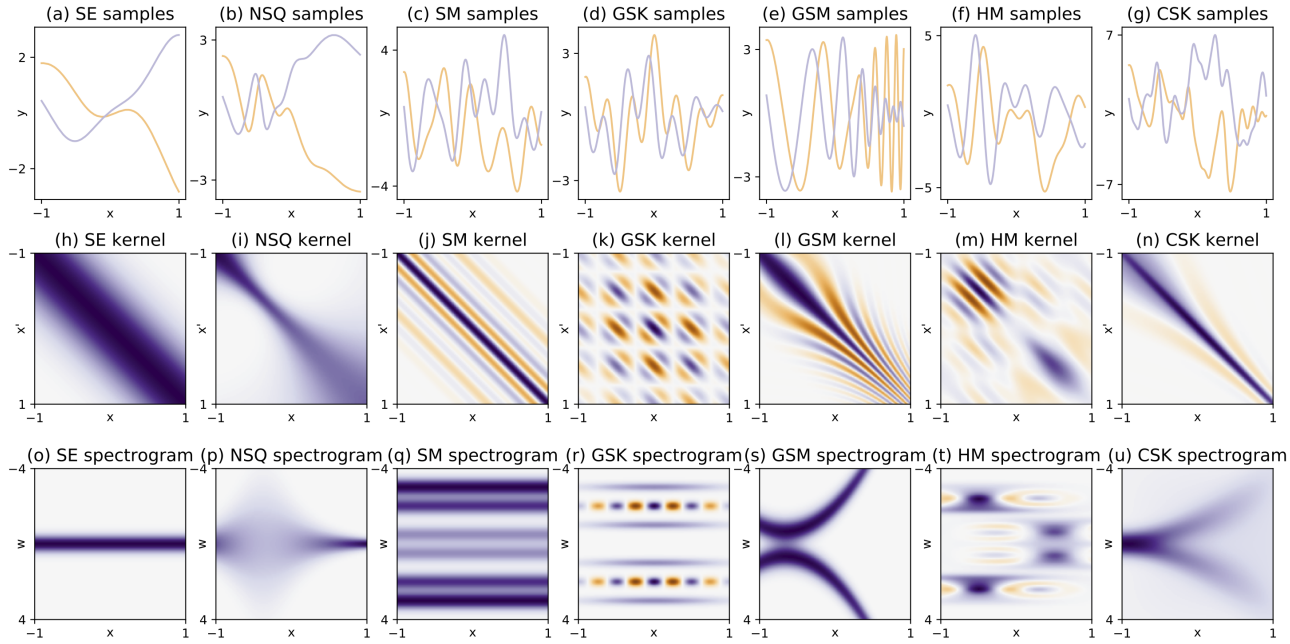

Figure 2: Overview of some kernels used in GP models, with sample paths (first row), kernel matrices (second row), and spectrogram (third row). Examples include SE (Rasmussen and Williams, 2006), NSQ (Paciorek and Schervish, 2004), SM (Wilson and Adams, 2013), GSK (Samo, 2017), GSM (Remes et al., 2017), HM (Shen et al., 2019) and CSK (current work).

and deep GP (DGP) models. The models in interest and their spectrograms are listed in Table 1, and their derivation summarized in Section 2 of the appendix.

3.1 Spectrograms of non-stationary kernels

The spectrogram applies to GP models equipped with nonparametric kernels (Paciorek and Schervish, 2004; Damianou and Lawrence, 2013; Remes et al., 2017), as demonstrated in Table 1. In particular, we investigate the notable similarity between CSK and generalized spectral mixture (GSM) (Remes et al., 2017) kernel.

The GSM kernel is a nonparametric, quasi-periodic kernel with an intuitively defined spectrogram. The correlation of GSM is a parametrization of (17)

$$D_{ij} = \left\| \left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)^{-1/2} \left(\mathbf{x}_i - \mathbf{x}_j\right) \right\|^2, \tag{22}$$

$$U_{ij} = \langle \boldsymbol{\mu}_i, \mathbf{x}_i \rangle - \langle \boldsymbol{\mu}_j, \mathbf{x}_j \rangle. \tag{23}$$

While it is tempting to equate the frequency mean $\boldsymbol{\xi}_{\mathbf{x}}$ (20) with $\boldsymbol{\mu}_i$, our analysis yields contradictory evidence, with $\boldsymbol{\xi}_{\mathbf{x}_i} = \boldsymbol{\mu}_i + (\mathcal{J}_{\mathbf{x}}\boldsymbol{\mu})\,|_{\mathbf{x}=\mathbf{x}_i}\mathbf{x}_i$, rendering the GSM kernel inherently biased in instantaneous frequency, which leads to an erroneously defined intuitive spectrogram (Remes et al., 2017).

The CSK records unbiased frequencies in contrast, albeit suffers a bias in the lengthscales (as shown in Table 1). We posit that it is essential that we adopt unbiased frequencies, so that the kernel extracts correct periodicities.

3.2 Standard DGPs are equivalent to GPs with NSQ kernels

Through spectrogram analysis, we uncover an equivalence between two classes of DGPs, namely the ones constructed with compositions and the ones with nonparametric covariance functions (Dunlop et al., 2018).

The compositional DGP (Damianou and Lawrence, 2013; Salimbeni and Deisenroth, 2017; Dunlop et al., 2018) generalizes standard GP with recursive functional composition:

$$f = f_L \circ f_{L-1} \circ \cdots \circ f_0, \tag{24}$$

$$f_l \sim \mathcal{GP}(m_l, k_l),\ l = 0, \cdots L. \tag{25}$$

The conditional kernel $k_l|f_{l-1}$ can be seen as one with varying lengthscales when k_L is the default SE kernel:

$$k_l(x, x'|f_{l-1}) = w^2 e^{-\frac{(f_{l-1}(x) - f_{l-1}(x'))^2}{2\ell^2}} \tag{26}$$

$$\approx w^2 \sigma_{ij} e^{-\frac{Q_{x,x'}}{2}} = k_{NS}(x, x'|\Sigma_i), \tag{27}$$

$$\Sigma_i = \frac{2\ell^2}{f_{l-1}'^2(x_i)}. \tag{28}$$

Here the $Q_{x,x'}$ corresponds to the squared distance

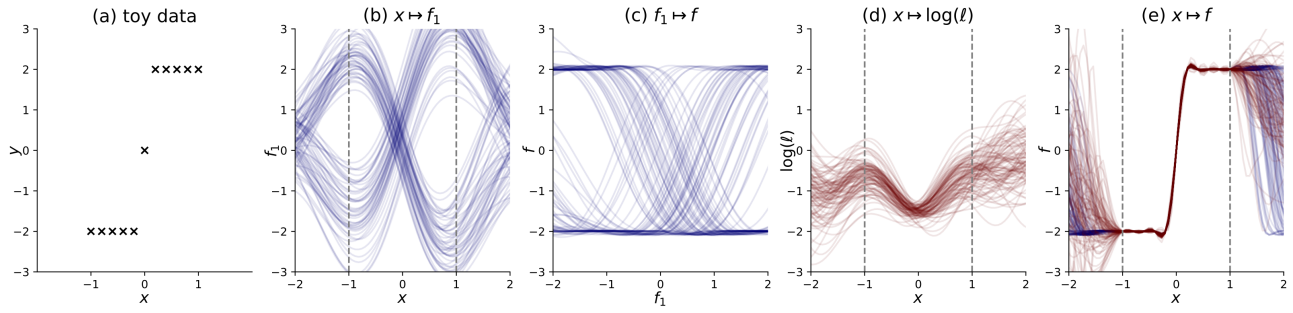

Figure 3: Posterior draws on two DGPs on toy dataset (see **(a)**). **(b)** and **(c)** demonstrate a 2-layer compositional DGP $f_2 \circ f_1$, where **(b)** maps from input $\mathbf{x}$ to the first layer f_1, and **(c)** from f_1 to the second layer f_2. **(d)** shows draws from an NSQ-DGP $f \sim \mathcal{GP}(0, k_{NS}(\cdot,\cdot;\ell))$, with log-lengthscale following prior $\log \ell \sim \mathcal{GP}(0, k)$. **(e)** shows direct mapping from input x to the latent function f, with consistent color markings.

defined in (8). According to our analysis (Table 1), the DGP kernel (26) and an NSQ kernel (11) with a lengthscale parametrization (28) share the same spectrogram, demonstrating an equivalence up to second-order effects.

3.3 Advantages of covariance function DGPs

Given the equivalence drawn in 3.2, we delineate the pros and cons between compositional DGPs and DGPs formulated with covariance functions (Dunlop et al., 2018), which comprise a layered structure of nonparametric kernels $k_{\boldsymbol{\theta}}$:

$$f_0(x) \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}}(x, x'|\boldsymbol{\theta} = \boldsymbol{\theta}_0)), \tag{29}$$

$$f_l(x)|f_{l-1}(x) \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}}(x, x'|\boldsymbol{\theta} = F \circ f_{l-1})). \tag{30}$$

Here $k_{\boldsymbol{\theta}}$ denotes such nonparametric kernels as NSQ (11) (Paciorek and Schervish, 2004), $\boldsymbol{\theta}$ their functional hyperparameters, and F a warping function (14) mapping GP samples to valid parameters. Despite the notable equivalence, the two types of GP models behave differently in practice (Dunlop et al., 2018). This discrepancy warrants a closer investigation, which evidences the superiority of covariance function DGPs.

We posit that the deep compositional probabilistic models (24) could benefit from monotonicity in hidden layers $f_0, f_1 \cdots, f_{L-1}$, a constraint not required for covariance function DGPs (30). Zero-mean compostional DGPs exhibit a pathology where the prior space of $f_L \circ \cdots \circ f_0$ degenerates into piecewise constant functions as $L \to \infty$ (Duvenaud et al., 2014; Dunlop et al., 2018). The state-of-the-art DGPs (Salimbeni and Deisenroth, 2017; Havasi et al., 2018) remedy this pathology with calibrated mean functions, which are likely to generate monotonic sample paths, despite the seemingly detrimental effect on model expressivity. Our derivation of approximate equivalence holds when the function $f_{l-1}(x)$ has nonzero derivatives almost everywhere, and is consequently monotonic. The "equivalence conditioned on monotonicity" marks the absence of rank pathology in covariance function DGPs (30) and provides alternate justification on calibrated mean functions.

Covariance function DGP avoids multi-modality by directly modeling lengthscale values. It is worth noting that the conditional kernel (26) stays invariant under *translation* and *reflection* of f_{l-1}, thus defining an equivalence class containing the above transformations. The invariance yields highly multi-modal posterior distributions (Havasi et al., 2018), a major obstacle to effective inference. The linear mean function (Salimbeni and Deisenroth, 2017) solves reflection but not translation of f_{l-1}. Meanwhile, functions belonging to the same equivalence class correspond to a singular lengthscale (28), and consequently a single covariance function DGP, which has notably more concentrated posterior densities. We demonstrate the two posteriors of a toy example in Figure 3.

While the covariance function DGP appears to claim a minor turf (Paciorek and Schervish, 2004; Heinonen et al., 2016; Remes et al., 2017) in deep probabilistic modeling, we advocate that it replace the customary compositional DGPs as an equivalent but more effective alternative.

3.4 CSK as deep Gaussian processes

Applying CSK (13) in the covariance function DGP recursion (30) yields a DGP that nests the NSQ construction (Dunlop et al., 2018) as a special case, defining a proven *ergodic* Markov chain (Dunlop et al., 2018). While the ergodicity of the DGP-CSK model upper bounds the model complexity with its mixing time, the application of CSK, nevertheless, significantly enriches the prior space (Figure 4).

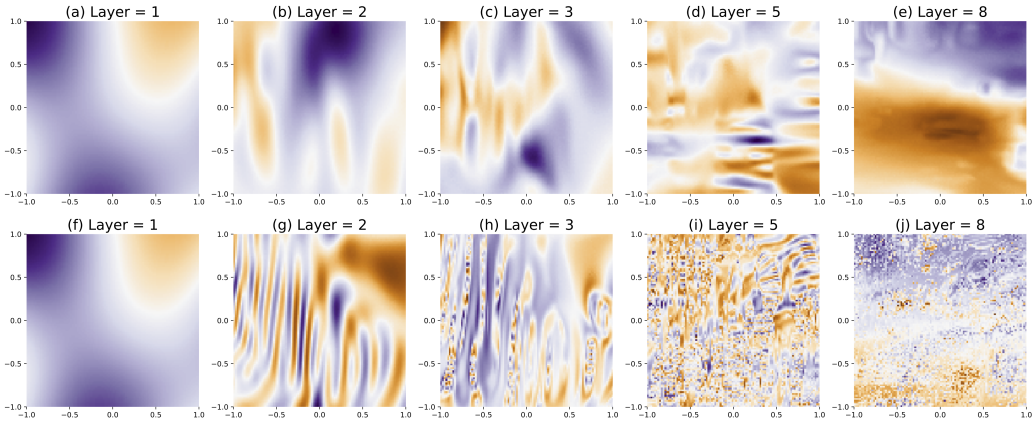

Figure 4: 2D prior draws from covariance function DGPs with kernels NSQ (11) (first row) and CSK (6) (second row).

4 Scalable inference for covariance function DGPs

In this section, we present a framework of scalable inference for covariance function DGPs (30) (Dunlop et al., 2018), which we formulate with f being a zero-mean GP with nonparametric kernel $k_{\boldsymbol{\theta}}$ ($f \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}})$), and the kernel parameters $\boldsymbol{\theta} = \{\theta_d\}_{d=1}^D$ have warped GP priors (14)-(15). Inserting a set of pseudo-observations denoted as $\mathbf{u}_{\theta_d} = F_d(h_d(\mathbf{z}_{\theta_d})), \mathbf{u}_f = f(\mathbf{z}_f)$ on each GP, we can factorize the joint likelihood $p(\mathbf{y}, \mathbf{f}, \boldsymbol{\theta}, \mathbf{u})$ as

$$p(\mathbf{y}, \mathbf{f}, \boldsymbol{\theta}, \mathbf{u}) = \underbrace{p(\mathbf{y}|\mathbf{f})}_{\text{likelihood}} \underbrace{p(\mathbf{f}|\mathbf{u}_f, \boldsymbol{\theta}) p(\mathbf{u}_f|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{u}_{\boldsymbol{\theta}}) p(\mathbf{u}_{\boldsymbol{\theta}})}_{\text{prior}}. \tag{31}$$

Similar to Hensman et al. (2015), we assume a *free-form* variational distribution $q(\mathbf{u}_f, \mathbf{u}_{\boldsymbol{\theta}})$, and formulate the variational posterior process as

$$q(f) = p(f|\mathbf{u}_f, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{u}_{\boldsymbol{\theta}}) q(\mathbf{u}_f, \mathbf{u}_{\boldsymbol{\theta}}). \tag{32}$$

Minimizing the term $\mathrm{KL}[q(f, \mathbf{u}_f, \mathbf{u}_{\boldsymbol{\theta}}) || p(f, \mathbf{u}_f, \mathbf{u}_{\boldsymbol{\theta}}|\mathbf{X}, \mathbf{y})]$ yields an optimal solution q^* with (un-normalized) log-likelihood with a normalizing constant C,

$$\begin{aligned} \log q^*(\mathbf{u}_f, \mathbf{u}_{\boldsymbol{\theta}}) =& \mathbb{E}_{p(\mathbf{f}|\mathbf{u}_f, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{u}_{\boldsymbol{\theta}})} \log p(\mathbf{y}|\mathbf{f}) p(\mathbf{u}_f|\boldsymbol{\theta}) \\ &+ \log p(\mathbf{u}_{\boldsymbol{\theta}}) - \log C. \end{aligned} \tag{33}$$

While the expectation term renders q^* intractable, we can, however, unbiasedly estimate the expectation with Monte Carlo samples of $\boldsymbol{\theta}$, and a subsample of the data (Salimbeni and Deisenroth, 2017; Havasi et al., 2018). Therefore, we can jointly infer the approximate posterior and optimize the hyperparameter values with a combination of stochastic gradient Hamiltonian Monte Carlo (Chen et al., 2014; Springenberg et al., 2016) and moving window Monte Carlo Expectation Maximization algorithm (Havasi et al., 2018).

5 Experiments

5.1 Recovering chirp signal

We first test our methods on a simulated dataset. A chirp signal is a non-stationary signal taking the generic form $x(t) = \cos(\phi(t))$. In this setting, we take 400 noisy

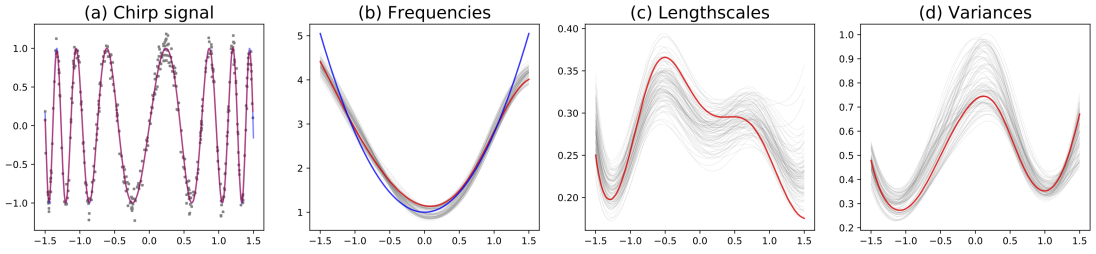

Figure 5: Regression on synthetic chirp signal. **(a)**: The training data (grey points) with the ground truth line (blue) and predicted mean (red); ground truth and predict mean overlap. **(b)**: HMC samples of the learned frequency function, with ground truth instantaneous frequency (blue line) and MAP estimate (red line). **(c)**, **(d)**: HMC samples of lengthscale and variance functions, with MAP estimates marked with red lines.

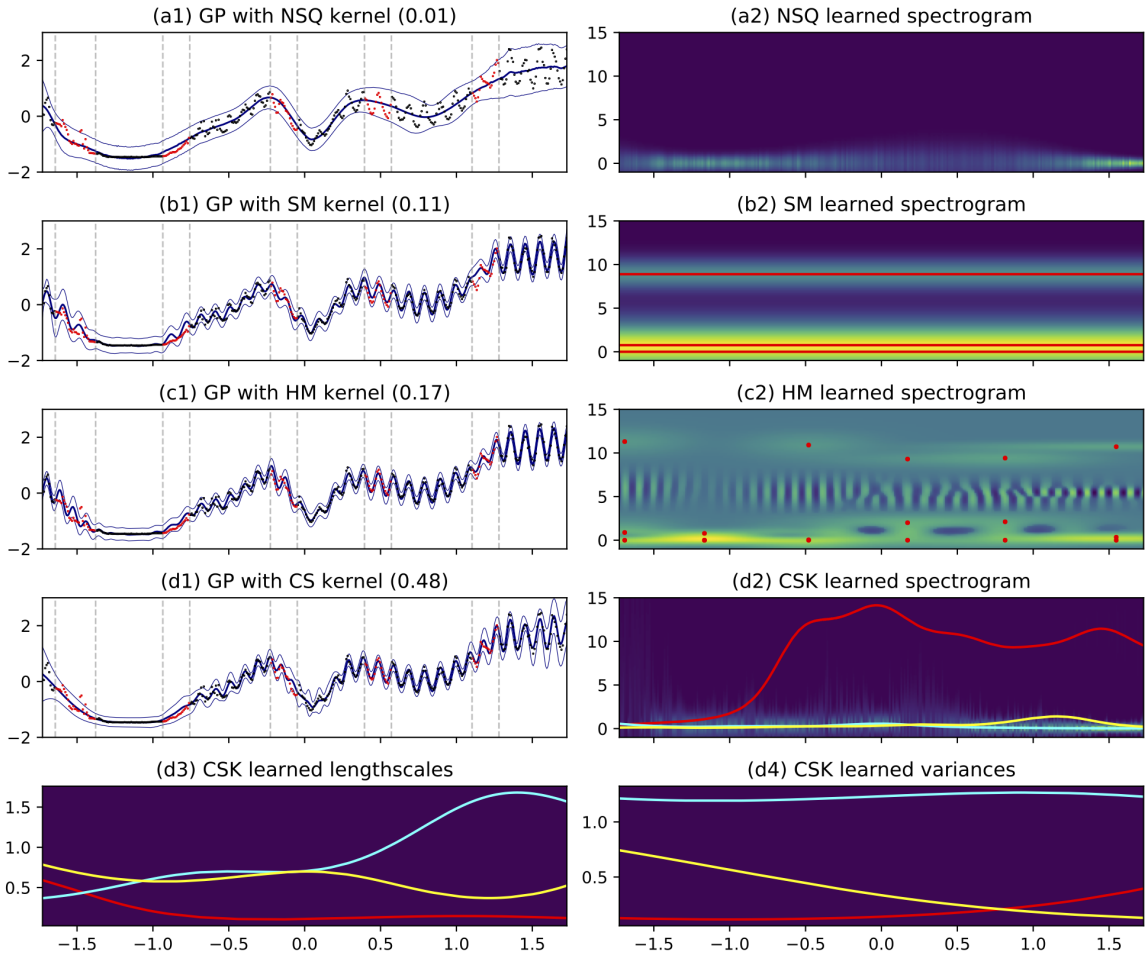

Figure 6: GP regression with solar irradiance. Training and test points are respectively marked by black and red dots in **(a1)**-**(d1)**, where the test log-likelihoods are shown in the parentheses. The frequency parameter obtained via kernel learning are marked as lines or points in the spectrogram (parametric values for SM and HM kernels in **(b2)** and **(c2)**, and the posterior mean value for each of the 3 components of CSK in **(d2)**). We color-consistently plot lengthscales (Σ_i in (3)) and variances ($\sigma(x_i)$ in (13)) in figures **(d3)** and **(d4)**, respectively.

observations of a chirp signal $x(t) = \cos(2\pi(t + 0.6t^3))$, and train a one-component CSK with 30 inducing points. The synthetic instantaneous frequency of this signal is $\phi'(t)/2\pi = 1 + 1.8t^2$, which is recovered with the frequency term in CSK (see figure 5).

5.2 Solar irradiance

We consider regression on the solar irradiance dataset (Lean, 2004), which exhibits some non-stationarities. We compare GP models with NSQ, SM, HMK (Shen et al., 2019) and CSK, where the inference for SM and HMK were done with sparse GP regression with inducing points (Titsias, 2009), and the functional hyperparameters of NSQ and CSK are inferred with SGHMC as illustrated in the previous section. While it is not immediately clear from the spectrogram visualizations (see figure 6), the spectral kernels (SM, HMK and CSK) learn similar frequency patterns: SM learns global frequency peaks; HMK learns and interpolates local patterns; CSK learns a varying global pattern, while also accounting for the non-Gaussianity of the data.

5.3 Air temperature anomaly dataset

We conduct spatiotemporal analysis on the air temperature anomaly dataset (Jones, 1994), which contains monthly air temperature deviations from the monthly mean temperature measured on locations on a global grid. We subsampled the data from 1988-1993 with 32910 readings and partitioned an 80%-20% split on training and testing data. Figure 7 demonstrates the predictive temperature anomaly between a GP with SE kernel and CSK with 5 components, which significantly

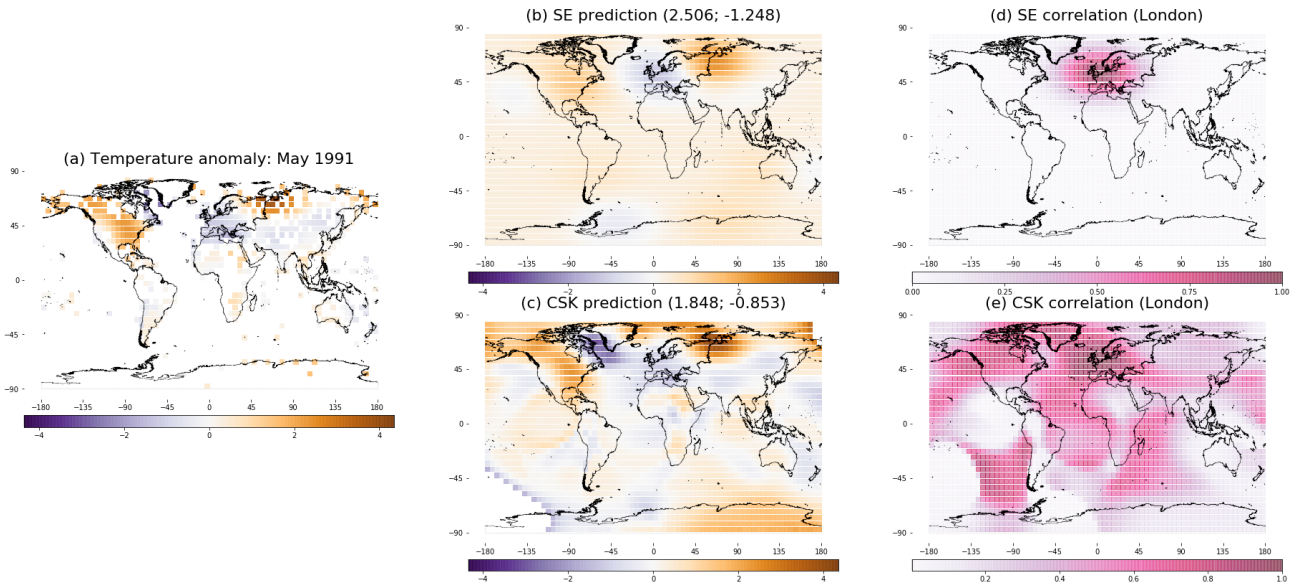

Figure 7: Air temperature anomaly dataset. **(a)** demonstrates the temperature anomaly readings from May 1991. **(b)**, **(c)** display the posterior predictive mean for May 1991 on a grid of global locations, with the numbers in parentheses denoting mean squared error (MSE) and mean log-likelihood, respectively. **(d)**, **(e)** depict the correlation between London and other geographical locations: it is worth noting that the SE kernel **(d)** only captures positive correlation on a small elliptical region.

Model	LINEAR	SVGP 100	SVGP 500	SM	DGP 4	NSQ	CSK
Test MLL	-1.277 (0.008)	-0.844 (0.015)	-0.828 (0.017)	-0.829 (0.017)	-0.627 (0.008)	-0.634 (0.009)	**-0.575 (0.012)**
Test MSE	0.753 (0.012)	0.326 (0.011)	0.315 (0.011)	0.313 (0.011)	0.307 (0.011)	0.310 (0.012)	**0.294 (0.008)**

Table 2: Results for NY taxi dataset, where we report the test mean log-likelihoods and mean square errors, with standard deviations denoted inside brackets, over 6 runs of the dataset.

improves the predictive performance. Nonstationarity is required to capture the correlation patterns demonstrated in the data.

5.4 New York Yellow Taxi dataset

We ran GP regression on a subset of the New York Yellow Taxi dataset[1], whose objective is to predict the taxi trip duration given the pickup and dropoff locations and the starting date and time. Given CSK's ability to handle periodicities, we treat the date and time as one feature. We ran 7 different models in total: Bayesian linear regression (LINEAR), standard stochastic variational GPs (Hensman et al., 2013) with 100 and 500 inducing points (SVGP 100 & 500), sparse GP with SM kernel (SM), 4-layer compositional deep GP with "monotonic" inner layers (DGP 4), GP with NSQ and CSK. Apart from SVGP 500, other models were run with 100 inducing points.

One can tell from comparison in Table 2 that the taxi dataset is nonlinear, non-Gaussian and exhibits nontrivial frequency patterns. CSK marginally outperforms other models by accounting for all three properties.

[1] `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`

6 Conclusion

In this work, we propose the *convolutional spectral kernel*, which generalizes the work of Paciorek and Schervish (2004) with spatially varying frequencies. We analyze commonly used kernels and GP models having input warping with spectrograms, which sheds light on the interpretation of deep models, and draws an equivalence between two types of DGPs. We propose a novel scalable inference framework for DGPs constructed via covariance functions, which empirically outperforms current compositional DGP methods.

Observing models through the lens of the spectrogram opens up points of interest for future work: While conceptually well-defined for multivariate inputs, the dimensionality of the spectrograms prevents straightforward visualization and thus needs further investigation; In addition, the theoretical results derived in this paper indicate that covariance function DGPs, as an appealing alternative to current DGPs, warrant further study.

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT. This work has been supported by the Academy of Finland (Flagship programme: Finnish Center for Artificial Intelligence, FCAI, and grants no. 299915, 319264, 313195, 294238, 292334).

References

M. Alvarez, D. Luengo, and N. D. Lawrence. Latent force models. In *Artificial Intelligence and Statistics*, pages 9–16, 2009.

T. Chen, E. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

M. M. Dunlop, M. A. Girolami, A. M. Stuart, and A. L. Teckentrup. How deep are deep Gaussian processes? *The Journal of Machine Learning Research*, 19(1): 2100–2145, 2018.

D. Duvenaud, J. R. Lloyd, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, pages 1166–1174, 2013.

D. Duvenaud, O. Rippel, R. Adams, and Z. Ghahramani. Avoiding pathologies in very deep networks. In *Artificial Intelligence and Statistics*, pages 202–210, 2014.

P. Flandrin. *Time-frequency/time-scale analysis*, volume 10. Academic press, 1998.

M. N. Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer, 1998.

M. Havasi, J. M. Hernández-Lobato, and J. J. Murillo-Fuentes. Inference in deep Gaussian processes using stochastic gradient Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 7517–7527, 2018.

M. Heinonen, H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki. Non-stationary Gaussian process regression with Hamiltonian Monte Carlo. In *Artificial Intelligence and Statistics*, pages 732–740, 2016.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*, 2013.

J. Hensman, A. G. Matthews, M. Filippone, and Z. Ghahramani. MCMC for variationally sparse Gaussian processes. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1648–1656. Curran Associates, Inc., 2015.

J. Hensman, N. Durrande, and A. Solin. Variational fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18:151–1, 2017.

D. Higdon, J. Swall, and J. Kern. Non-stationary spatial modeling. *Bayesian Statistics*, 6, 1998.

P. D. Jones. Hemispheric surface air temperature variations: a reanalysis and an update to 1993. *Journal of Climate*, 7(11):1794–1802, 1994.

J. Lean. Solar irradiance reconstruction. *IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series*, 35, 2004.

J Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. 2018. URL `https://openreview.net/pdf?id=B1EA-M-0Z`.

R. M. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. Technical report, University of Toronto, 1993.

C. J. Paciorek and M. J. Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 273–280, 2004.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, USA, January 2006.

S. Remes, M. Heinonen, and S. Kaski. Non-stationary spectral kernels. In *Advances in Neural Information Processing Systems*, pages 4642–4651, 2017.

H. Salimbeni and M. Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Y.-L. K. Samo. *Advances in kernel methods: towards general-purpose and scalable models*. PhD thesis, University of Oxford, 2017.

Z. Shen, M. Heinonen, and S. Kaski. Harmonizable mixture kernels with variational Fourier features. In *Artificial Intelligence and Statistics*, pages 3273–3282, 2019.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264, 2006.

E. Snelson, Z. Ghahramani, and C. E. Rasmussen. Warped Gaussian processes. In *Advances in neural information processing systems*, pages 337–344, 2004.

J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems*, pages 4134–4142, 2016.

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

F. Tobar. Bayesian nonparametric spectral estimation. In *Advances in Neural Information Processing Systems*, pages 10127–10137, 2018.

F. Tobar, T. D. Bui, and R. E. Turner. Learning stationary time series using Gaussian processes with nonparametric kernels. In *Advances in Neural Information Processing Systems*, pages 3501–3509, 2015.

C. K. I. Williams. Computing with infinite networks. In *Advances in Neural Information Processing Systems*, pages 295–301, 1997.

A. G. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378, 2016.

Andrew G Wilson, Christoph Dann, Chris Lucas, and Eric P Xing. The human kernel. In *Advances in neural information processing systems*, pages 2854–2862, 2015.