# Incorporating Dependencies in Spectral Kernels for Gaussian Processes

Kai Chen[1,2,3,5(✉)], Twan van Laarhoven[3,4], Jinsong Chen[1,2,5], and Elena Marchiori[3]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China
js.chen@siat.ac.cn

[2] Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, People's Republic of China

[3] Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands
{kai,elenam}@cs.ru.nl

[4] Faculty of Management, Science and Technology, Open University of The Netherlands, Heerlen, The Netherlands
twan.vanlaarhoven@ou.nl

[5] Shenzhen Engineering Laboratory of Ocean Environmental Big Data Analysis and Application, Shenzhen 518055, People's Republic of China

**Abstract.** Gaussian processes (GPs) are an elegant Bayesian approach to model an unknown function. The choice of the kernel characterizes one's assumption on how the unknown function autocovaries. It is a core aspect of a GP design, since the posterior distribution can significantly vary for different kernels. The spectral mixture (SM) kernel is derived by modelling a spectral density - the Fourier transform of a kernel - with a linear mixture of Gaussian components. As such, the SM kernel cannot model dependencies between components. In this paper we use cross convolution to model dependencies between components and derive a new kernel called Generalized Convolution Spectral Mixture (GCSM). Experimental analysis of GCSM on synthetic and real-life datasets indicates the benefit of modeling dependencies between components for reducing uncertainty and for improving performance in extrapolation tasks.

**Keywords:** Gaussian processes · Spectral mixture · Convolution · Dependency · Uncertainty

## 1 Introduction

Gaussian processes (GPs) provide regression models where a posterior distribution over the unknown function is maintained as evidence is accumulated.

**Electronic supplementary material** The online version of this chapter (https://doi.org/10.1007/978-3-030-46147-8_34) contains supplementary material, which is available to authorized users.

U. Brefeld et al. (Eds.): ECML PKDD 2019, LNAI 11907, pp. 565–581, 2020.
https://doi.org/10.1007/978-3-030-46147-8_34

This allows GPs to learn complex functions when a large amount of evidence is available, and it makes them robust against overfitting in the presence of little evidence. GPs can model a large class of phenomena through the choice of the kernel, which characterizes one's assumption on how the unknown function autocovaries [17,18]. The choice of the kernel is a core aspect of a GP design, since the posterior distribution can significantly vary for different kernels. In particular, in [24] a flexible kernel called Spectral Mixture (SM) was defined, by modelings the kernel's spectrum with a mixture of Gaussians. An SM kernel can be represented by a sum of components, and can be derived from Bochner's theorem as the inverse Fourier Transform (FT) of its corresponding spectral density. SM kernels assume mutually independence of its components [24–26].

Here we propose a generalization of SM kernels that explicitly incorporates dependencies between components. We use cross convolution to model dependencies between components, and derive a new kernel called Generalized Convolution Spectral Mixture (GCSM) kernel. The number of hyper-parameters remains equal to that of SM, and there is no increase in computational complexity. A stochastic variational inference technique is used to perform scalable inference. In the proposed framework, GCSM without cross components (that is, by only considering auto-convolution of base components) reduces to the SM kernel.

We assess the performance of GCSM kernels through extensive experiments on real-life datasets. The results show that GCSM is able to capture dependence structure in time series and multi-dimensional data containing correlated patterns. Furthermore, we show the benefits of the proposed kernel for reducing uncertainty, overestimation and underestimation in extrapolation tasks. Our main contributions can be summarized as follows:

- a new spectral mixture kernel that captures dependencies between components;
- two metrics, posterior correlation (see Eq. 10) and learned dependency (see Eq. 19) to analyze intrinsic dependencies between components in the SM kernel and dependencies captured by our kernel, respectively;
- an extensive comparison between the proposed GCSM and other SM kernels in terms of spectral density, covariance, posterior predictive density and sampling, as well as in terms of performance gain.

The remainder of this paper is organized as follows. We start by giving a background on GPs, SM kernels, and we briefly describe related work. Next, we introduce the GCSM kernel, and discuss the differences between the GCSM and SM kernels. Then we describe the experimental setting and show results on synthetic and real-world datasets. We conclude with a summary and discussion on future work.

## 2 Background

A GP is any distribution over functions such that any finite set of function values has a joint Gaussian distribution. A GP model, before conditioning on the data, is completely specified by its mean function $m(\mathbf{x}) = \mathbb{E}(f(\mathbf{x}))$ and its covariance function (also called *kernel*) $k(\mathbf{x}, \mathbf{x}') = \text{cov}(f(\mathbf{x}), f(\mathbf{x}'))$ for input vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^P$. It is common practice to assume that the mean function is simply zero everywhere, since uncertainty about the mean function can be taken into account by adding an extra term to the kernel (cf. e.g. [18]).

The kernel induces a positive definite covariance matrix $K = k(X, X)$ of the training locations set $X$. For a regression task [18], by choosing a kernel and inferring its hyper-parameters $\Theta$, we can predict the unknown function value $\tilde{y}^*$ and its variance $\mathbb{V}[\tilde{y}^*]$ (the uncertainty) for a test point $\mathbf{x}^*$ as follows:

$$\tilde{y}^* = \mathbf{k}^{*\top}(K + \sigma_n^2 I)^{-1}\mathbf{y} \tag{1}$$

$$\mathbb{V}[\tilde{y}^*] = k^{**} - \mathbf{k}^{*\top}(K + \sigma_n^2 I)^{-1}\mathbf{k}^* \tag{2}$$

where $k^{**} = k(\mathbf{x}^*, \mathbf{x}^*)$, $\mathbf{k}^{*\top}$ is the vector of covariances between $\mathbf{x}^*$ and $X$, and $\mathbf{y}$ are the observed values at training locations in $X$. The hyper-parameters can be optimized by minimizing the Negative Log Marginal Likelihood (NLML) $-\log\, p(\mathbf{y}|\mathbf{x}, \Theta)$. Smoothness and generalization properties of GPs depend on the kernel function and its hyper-parameters $\Theta$ [18]. In particular, the SM kernel [26], here denoted by $k_{\text{SM}}$, is derived by modeling the empirical spectral density as a Gaussian mixture, using Bochner's Theorem [2,22], resulting in the following kernel:

$$k_{\text{SM}}(\tau) = \sum_{i=1}^{Q} w_i k_{\text{SM}i}(\tau), \tag{3}$$

$$k_{\text{SM}i}(\tau) = \cos\left(2\pi\tau^\top \boldsymbol{\mu}_i\right) \prod_{p=1}^{P} \exp\left(-2\pi^2\tau^2\Sigma_{i,p}\right), \tag{4}$$

where $\tau = \mathbf{x} - \mathbf{x}'$, $Q$ denotes the number of components, $k_{\text{SM}i}$ is the $i$-th component, $P$ denotes the input dimension, and $w_i$, $\boldsymbol{\mu}_i = [\mu_{i,1}, ..., \mu_{i,P}]$, and $\Sigma_i = \text{diag}\left(\left[\sigma_{i,1}^2, ..., \sigma_{i,P}^2\right]\right)$ are the weight, mean, and variance of the $i$-th component in the frequency domain, respectively. The variance $\sigma_i^2$ can be thought of as an inverse length-scale, $\mu_i$ as a frequency, and $w_i$ as a contribution. For SM kernel, we have $\hat{k}_{\text{SM}i}(\mathbf{s}) = [\varphi_{\text{SM}i}(\mathbf{s}) + \varphi_{\text{SM}i}(-\mathbf{s})]/2$ where $\varphi_{\text{SM}i}(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_i, \Sigma_i)$ is a symmetrized scale-location Gaussian in the frequency domain.

The SM kernel does not consider dependencies between components, because it is a linear combination of $\{k_{\text{SM}i}\}_{i=1}^{Q}$ (see Eq. 3). Therefore its underlying assumption is that such components are mutually independent. One should not confuse the spectral mixture components that make up the spectral density of the SM kernel with the base components of the Fourier Transform (FT): (1) FT components are periodic trigonometric functions, such as sine and cosine functions, while SM kernel components are quasi-periodic Gaussian functions; (2) FT components are orthogonal (i.e. the product of an arbitrary pair of Fourier series components is zero) while the product of two arbitrary SM components is not necessarily equal to zero; (3) the SM component in the frequency domain is a

Gaussian function covering wide frequency range while an FT component is just a sharp peak at a single frequency, which is covered by multiple SM components.

## 3 Related Work

Various kernel functions have been proposed [18], such as Squared Exponential (SE), Periodic (PER), and general Matérn (MA). Recently, Spectral Mixture (SM) kernels have been proposed in [24]. Additive GPs have been proposed in [4], a GP model whose kernel implicitly sums over all possible products of one-dimensional base kernels. Extensions of these kernels include the spectral mixture product kernel (SMP) [25] $k_{\mathrm{SMP}}(\tau|\Theta) = \prod_{p=1}^{P} k_{\mathrm{SM}}(\tau_p|\Theta_p)$, which uses multi-dimensional SM kernels, and extends the application scope of SM kernels to image data and spatial time data. Other interesting families of kernels include non-stationary kernels [7,10,19,21], which are capable to learn input-dependent covariances between inputs. All these mentioned kernels do not consider dependencies between components. To the best of our knowledge, our proposed kernel is the first attempt to explicitly model dependencies between components.

The problem of expressing structure present in the data being modeled with kernels has been investigated also in the context of kernel composition. For instance, in [3] a framework was introduced for composing kernel structures. A space of kernel structures is defined compositionally in terms of sums and products of a small number of base kernel structures. Then an automatic search over this space of kernel structures is performed using marginal likelihood as search criterion. Although composing kernels allows one to produce kernels combining several high-level properties, they depend on the choice of base kernel families, composition operators, and search strategy. Instead, here we directly enhance SM kernels by incorporating dependency between components.

## 4 Dependencies Between SM Components

Since the SM kernel is additive, any $f \sim \mathcal{GP}(0, k_{\mathrm{SM}})$ can be expressed as

$$f = \sum_{i=1}^{Q} f_i, \tag{5}$$

where each $f_i \sim \mathcal{GP}(0, w_i k_{\mathrm{SM}i})$ is drawn from a GP with kernel $w_i k_{\mathrm{SM}i}$. With a slight abuse of notation we denote by $\boldsymbol{f}_i$ the function values at training locations $X$, and by $\boldsymbol{f}_i^*$ the function values at some set of query locations $X^*$.

From the additivity of the SM kernel it follows that the $f_i$'s are *a priori* independent. Then, by using the formula for Gaussian conditionals we can give the conditional distribution of a GP-distributed function $\boldsymbol{f}_i^*$ conditioned on its sum with another GP-distributed function $\boldsymbol{f}_j$ :

$$\boldsymbol{f}_i^* | \boldsymbol{f}_{i+j} \sim \mathcal{N}\Big(K_i^{*\top} K_{i+j}^{-1} \boldsymbol{f}_{i+j},\ K_i^{**} - K_i^{*\top} K_{i+j}^{-1} K_i^*\Big) \tag{6}$$

where $\boldsymbol{f}_{i+j} = \boldsymbol{f}_i + \boldsymbol{f}_j$ and $K_{i+j} = K_i + K_j$. The reader is referred to [3] (Sect. 2.4.5) for the derivation of these results. The Gaussian conditionals express the model's posterior uncertainty about the different components of the signal, integrating over the possible configurations of the other components.

In particular, we have:

$$\mathbb{V}(\boldsymbol{f}_i^*|\boldsymbol{f}_i) = K_i^{**} - K_i^{*\top} K_i^{-1} K_i^*, \tag{7}$$

$$\mathbb{V}(\boldsymbol{f}_i^*|\boldsymbol{f}_i, \boldsymbol{f}_j) = K_i^{**} - K_i^{*\top} K_{i+j}^{-1} K_i^*. \tag{8}$$

In general $\mathbb{V}(\boldsymbol{f}_i^*|\boldsymbol{f}_i) \neq \mathbb{V}(\boldsymbol{f}_i^*|\boldsymbol{f}_i, \boldsymbol{f}_j)$ when dependencies between components are present. We can also compute the posterior covariance between the height of any two functions, conditioned on their sum [3]:

$$\mathrm{Cov}\left(\boldsymbol{f}_i^*, \boldsymbol{f}_j^*|\boldsymbol{f}_i, \boldsymbol{f}_j\right) = -K_i^{*\top} K_{i+j}^{-1} K_j^*. \tag{9}$$

We define posterior correlation $\rho_{ij}^*$ as normalized posterior covariance:

$$\rho_{ij}^* = \frac{\mathrm{Cov}\left(\boldsymbol{f}_i^*, \boldsymbol{f}_j^*|\boldsymbol{f}_i, \boldsymbol{f}_j\right)}{\left(\mathbb{V}\left(\boldsymbol{f}_i^*|\boldsymbol{f}_i, \boldsymbol{f}_j\right)\mathbb{V}\left(\boldsymbol{f}_j^*|\boldsymbol{f}_i, \boldsymbol{f}_j\right)\right)^{1/2}}. \tag{10}$$

We can use $\rho_{ij}^* \neq 0$ as indicator of statistical dependence between components $i$ and $j$. In our experiments, we will use the normalized posterior covariance to illustrate the presence of dependencies between components in SM kernels for GPs.

## 5 Generalized Convolution SM Kernels

We propose to generalize SM kernels by incorporating cross component terms. To this aim we use versions of the seminal Convolution theorem, which states that under suitable conditions the Fourier transform of a convolution of two signals is the pointwise product of their Fourier transforms. In particular, convolution in the time domain equals point-wise multiplication in the frequency domain. The construction of our kernel relies on the fact that any stationary kernel $k(\mathbf{x}, \mathbf{x}')$ can be represented as a convolution form on $\mathbb{R}^P$ (see e.g. [5,6,13])

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^P} g(\mathbf{u})\ g(\tau - \mathbf{u})\, d\mathbf{u} = (g * g)(\tau). \tag{11}$$

By applying a Fourier transformation to the above general convolution form of the kernel we obtain $\hat{k}(\mathbf{s}) = (\hat{g}(\mathbf{s}))^2$ in the frequency domain. For each weighted component $w_i k_{\mathrm{SM}i}(\tau)$ in the SM kernel, we can define the function $\hat{g}_{\mathrm{SM}i}(\mathbf{s})$ as

$$\hat{g}_{\mathrm{SM}i}(\mathbf{s}) = \left(w_i \hat{k}_{\mathrm{SM}i}(\mathbf{s})\right)^{1/2} = {w_i}^{\frac{1}{2}} \frac{\exp\left(-\frac{1}{4}(\mathbf{s} - \boldsymbol{\mu}_i)^\top \Sigma_i^{-1}(\mathbf{s} - \boldsymbol{\mu}_i)\right)}{\left((2\pi)^P |\Sigma_i|\right)^{1/4}}, \tag{12}$$

which is the basis function of the $i$-th weighted spectral density. We use cross-correlation, which is similar in nature to the convolution of two functions.

The cross-correlation of functions $f(\tau)$ and $g(\tau)$ is equivalent to the convolution of $\overline{f}(-\tau)$ and $g(\tau)$ [1]. we have that the cross-correlation between two components $f_i \sim \mathcal{GP}(0, w_i k_{\mathrm{SM}i})$ and $f_j \sim \mathcal{GP}(0, w_j k_{\mathrm{SM}j})$ is as

$$k_{\mathrm{GCSM}}^{i\times j}(\tau) = w_i k_{\mathrm{SM}i}(\tau) \star w_j k_{\mathrm{SM}j}(\tau) = \mathcal{F}_{s\to\tau}^{-1}\left[w_i \varphi_{\mathrm{SM}i}(\mathbf{s}) \cdot \overline{w_j \varphi_{\mathrm{SM}j}}(\mathbf{s})\right](\tau) \quad (13)$$

where $\mathcal{F}_{s\to\tau}^{-1}$, $\star$, and $\overline{(-)}$ denote the inverse FT, the cross-correlation operator, and the complex conjugate operator, respectively. Here $\varphi_{\mathrm{SM}i}(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_i, \Sigma_i)$ is a symmetrized scale-location Gaussian in the frequency domain ($\varphi_{\mathrm{SM}i}(\mathbf{s}) = \overline{\varphi_{\mathrm{SM}i}}(\mathbf{s})$). The product of Gaussians $\varphi_{\mathrm{SM}i}(\mathbf{s})$ and $\varphi_{\mathrm{SM}j}(\mathbf{s})$ is also a Gaussian. Therefore, the cross-correlation term in the frequency domain has also a Gaussian form and must be greater than zero, which implies the presence of dependencies between $f_i$ and $f_j$.

The cross-correlation term $k_{\mathrm{GCSM}}^{i\times j}(\tau)$ of our new kernel, obtained as cross-correlation of the $i$-th and $j$-th base components in SM, corresponds to the cross spectral density term

$$\hat{k}_{\mathrm{GCSM}}^{i\times j}(\mathbf{s}) = \hat{g}_{\mathrm{SM}i}(\mathbf{s}) \cdot \overline{\hat{g}_{\mathrm{SM}j}}(\mathbf{s}) \quad (14)$$

in the frequency domain. From (12) and (14) we obtain

$$\hat{k}_{\mathrm{GCSM}}^{i\times j}(\mathbf{s}) = w_{ij} a_{ij} \frac{\exp\left(-\frac{1}{2}(\mathbf{s}-\boldsymbol{\mu}_{ij})^\top {\Sigma_{ij}}^{-1}(\mathbf{s}-\boldsymbol{\mu}_{ij})\right)}{\sqrt{(2\pi)^P |\Sigma_{ij}|}}. \quad (15)$$

The parameters for the cross spectral density term $\hat{k}_{\mathrm{GCSM}}^{i\times j}(\mathbf{s})$ corresponding to the cross convolution component $k_{\mathrm{GCSM}}^{i\times j}(\tau)$ are:

- cross weight: $w_{ij} = \sqrt{w_i w_j}$
- cross amplitude: $a_{ij} = \left|\frac{\sqrt{4\Sigma_i \Sigma_j}}{\Sigma_i + \Sigma_j}\right|^{\frac{1}{2}} \exp\left(-\frac{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^\top (\Sigma_i + \Sigma_j)^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}{4}\right)$
- cross mean: $\boldsymbol{\mu}_{ij} = \frac{\Sigma_i \boldsymbol{\mu}_j + \Sigma_j \boldsymbol{\mu}_i}{\Sigma_i + \Sigma_j}$;
- cross covariance: $\Sigma_{ij} = \frac{2\Sigma_i \Sigma_j}{\Sigma_i + \Sigma_j}$

Parameters $\boldsymbol{\mu}_{ij}$ and $\Sigma_{ij}$ can be interpreted as frequency and inverse length-scale of the cross component $k_{\mathrm{GCSM}}^{i\times j}(\tau)$, respectively. Cross amplitude $a_{ij}$ is a normalization constant which does not depend on $\mathbf{s}$.

Observe that when $\hat{g}_{\mathrm{SM}i}(\mathbf{s})$ is equal to $\hat{g}_{\mathrm{SM}j}(\mathbf{s})$, $w_{ij}$ $a_{ij}$, $\boldsymbol{\mu}_{ij}$, and $\Sigma_{ij}$ reduce to $w_i$, 1, $\boldsymbol{\mu}_i$, and $\Sigma_i$, respectively. In this case, the cross spectral density $\hat{k}_{\mathrm{GCSM}}^{i\times j}(\mathbf{s})$ is equal to $\hat{k}_{\mathrm{SM}i}(\mathbf{s})$. We can observe that the closer the frequencies $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ are and as closer the scales $\Sigma_i$ and $\Sigma_j$ between components $i$ and $j$ in the SM kernel are, the higher the cross convolution components contribution in GCSM will be.

Using the inverse FT, by the distributivity of the convolution operator and by the symmetry of the spectral density, we can obtain the GCSM kernel with

$Q$ (auto-convolution) components as:

$$k_{\mathrm{GCSM}}(\tau) = \sum_{i=1}^{Q}\sum_{j=1}^{Q} c_{ij} \exp\left(-2\pi^2\tau^\top \Sigma_{ij}\tau\right) \cos\left(2\pi\tau^\top \boldsymbol{\mu}_{ij}\right) \tag{16}$$

where $c_{ij} = w_{ij}a_{ij}$ is the cross contribution incorporating cross weight and cross amplitude to quantify the dependency between components in the GCSM kernel. The proof that GCSM is positive semi-definite is given in the Appendix. The auto-convolution cross-terms in GCSM correspond to the components in SM since $k^{i\times i}_{\mathrm{GCSM}}(\tau) = k_{\mathrm{SM}i}(\tau)$. It is a mixture of periodic cosine kernels and their dependencies, weighted by exponential weights.

## 6 Comparisons Between GCSM and SM

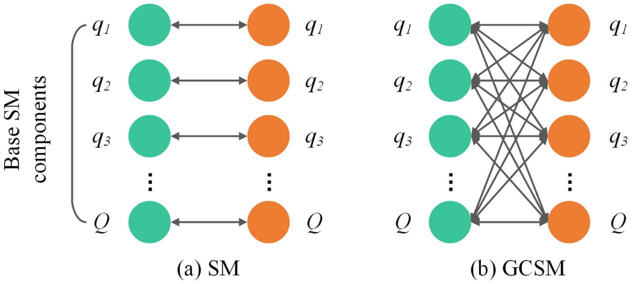

**Fig. 1.** SM and GCSM with $Q$ components. (a) SM models only auto-convolution between base components. (b) GCSM models both auto- and cross-convolution between base components.

Figure 1 illustrates the difference between SM and GCSM, where each connection represents a convolution component of the kernel. SM is an auto-convolution spectral mixture kernel that ignores the cross-correlation between base components. The figure also shows that SM is a special case of GCSM since the latter involves both cross convolution and auto-convolution of base components. In GCSM, dependencies are explicitly modeled and quantified. In the experiment illustrated in Fig. 2, SM and GCSM have the same initial parameters the same noise term. The observations are sampled from a $\mathcal{GP}(0, K_{\mathrm{SM}} + K_{\mathrm{GCSM}})$. From Fig. 2 we can observe clear differences (in terms of amplitude, peak, and trend from SM) for the kernel functions (SM: top, in dashed red; GCSM: bottom, in dashed blue). For the corresponding spectral densities, the dependence (in magenta) modeled by GCSM is also a Gaussian in the frequency domain, which yields a spectral mixture with different magnitude. The posterior distribution and sampling are obtained from GCSM and SM conditioned on six observations (black crosses). One can observe that the predictive distribution of GCSM has a tighter confidence interval (in blue shadow) than SM (in red shadow).

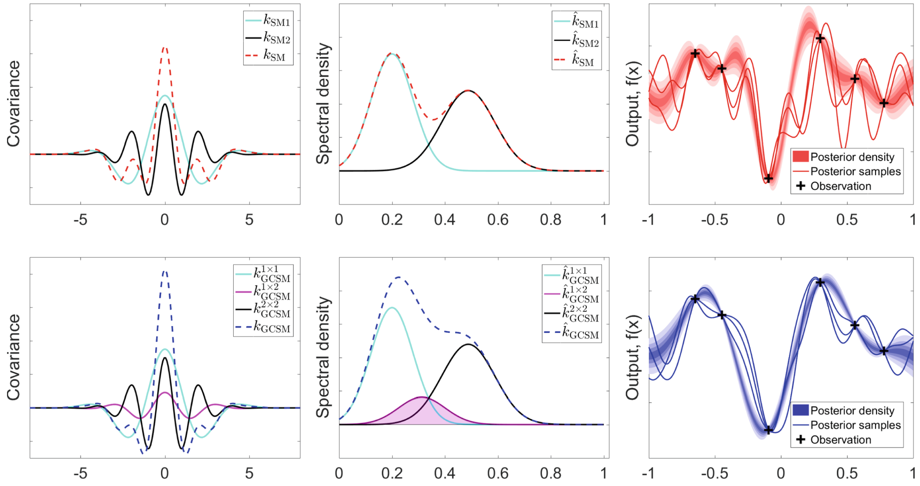

**Fig. 2.** Covariance, spectral density, and posterior functions drawn from GPs with SM and GCSM kernels conditioning on six samples. In the first row two SM components ($w_1 k_{\mathrm{SM1}}(\tau)$ and $w_2 k_{\mathrm{SM2}}(\tau)$) correspond to two solid lines (in cyan and black). In the second row two GCSM components with dependent structures ($k_{\mathrm{GCSM}}^{1\times2}(\tau)$) (in magenta). SM and GCSM plots have the same axes. (Color figure online)

# 7 Scalable Inference

Exact inference for GPs is prohibitively slow for more than a few thousand datapoints, as it involves inverting the covariance matrix $(K+\sigma_n^2 I)^{-1}$ and computing the determinant of the covariance $|K+\sigma_n^2 I|$. This issues are addressed by covariance matrix approximation [16,20,23] and inference approximation [8,9].

Here we employ stochastic variational inference (SVI) which provides a generalized framework for combining inducing points $\mathbf{u}$ and variational inference yielding impressive efficiency and precision. Specifically, SVI approximates the true GP posterior with a GP conditioned on a small set of inducing points $\mathbf{u}$, which as a set of global variables summarise the training data and are used to perform variational inference. The variational distribution $P(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_\mathbf{u}, \Sigma_\mathbf{u})$ gives a variational lower bound $\mathcal{L}_3(\mathbf{u}; \boldsymbol{\mu}_\mathbf{u}, \Sigma_\mathbf{u})$, also called Evidence Lower Bound (ELBO) of the quantity $p(\mathbf{y}|X)$. From [9], the variational distribution $\mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_\mathbf{u}, \Sigma_\mathbf{u})$ contains all the information in the posterior approximation, which represents the distribution on function values at the inducing points $\mathbf{u}$. From $\frac{\partial \mathcal{L}_3}{\partial \boldsymbol{\mu}_\mathbf{u}} = 0$ and $\frac{\partial \mathcal{L}_3}{\partial \Sigma_\mathbf{u}} = 0$, we can obtain an optimal solution of the variational distribution. The posterior distribution of testing data can be written as

$$p(f^*|X,\mathbf{y}) = \mathcal{N}(\mathbf{k}_\mathbf{u}^* K_{\mathbf{uu}}^{-1} \boldsymbol{\mu}_\mathbf{u}, k^{**} + \mathbf{k}_\mathbf{u}^{*\top} (K_{\mathbf{uu}}^{-1} \Sigma_\mathbf{u} K_{\mathbf{uu}}^{-1} - K_{\mathbf{uu}}^{-1}) \mathbf{k}_\mathbf{u}^*) \qquad (17)$$

where $\mathbf{k}_\mathbf{u}^*$ is the GCSM covariance vector between $\mathbf{u}$ and test point $\mathbf{x}^*$. The complexity of SVI is $\mathcal{O}(m^3)$ where $m$ is the number of inducing points.

### 7.1 Hyper-parameter Initialization

In our experiments, we use the empirical spectral densities to initialize the hyper-parameters, as recommend in [10,24]. Different from these works, we apply a Blackman window function to the training data to improve the quality of empirical spectral densities, e.g. the signal to noise ratio (SNR), and to more easily discover certain characteristics of the signal, e.g. magnitude and frequency. We consider the windowed empirical spectral densities $p(\Theta|\mathbf{s})$ as derived from the data, and then apply a Bayesian Gaussian mixture model (GMM) in order to get the $Q$ cluster centers of the Gaussian spectral densities [10].

$$p(\Theta|\mathbf{s}) = \sum_{i=1}^{Q} \tilde{w_i}\mathcal{N}(\tilde{\boldsymbol{\mu}}_i, \tilde{\Sigma}_i) \tag{18}$$

We use the Expectation Maximization algorithm [15] to estimate the parameters $\tilde{w_i}$, $\tilde{\boldsymbol{\mu}}_i$, and $\tilde{\Sigma}_i$. The results are used as initial values of $w_i$, $\boldsymbol{\mu}_i$, and $\Sigma_i$, respectively.

## 8 Experiments

We comparatively assess the performance of GCSM on real-world datasets. Three of these datasets have been used in the literature of GP methods. The other is a relative new dataset which we use to illustrate the capability of GPs with the considered kernels to model irregular long term increasing trends. We use Mean Squared Error ($\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{y}_i\right)^2$) as the performance metric for all tasks. We used the 95% confidence interval (instead of, e.g., error bar) to quantify uncertainty (see Eq. (2)). In addition to these performance metrics, we also consider the posterior correlation $\rho^*_{ij}$ (see Eq. (10)) to illustrate the underlying dependency between SM components. Moreover, to illustrate the dependency between components captured by the cross-components in our GCSM kernel, we use the normalized cross-correlation term:

$$\gamma_{ij}(\tau) = \frac{k^{i\times j}_{\text{GCSM}}(\tau)}{\sqrt{k_{\text{SM}i}(\tau)k_{\text{SM}j}(\tau)}} \tag{19}$$

We call $\gamma_{ij}$ *learned dependency* between component $i$ and $j$. Note that $\gamma_{ij} = 1$ when $i = j$. In our experiments we will analyze dependency between components in SM kernel for GPs as expressed by the posterior covariance, and dependency modeled by GCSM kernels for GPs as expressed by $\gamma_{ij}$'s. We compare GCSM with ordinary SM for prediction tasks on four real-life datasets: monthly average atmospheric $\text{CO}_2$ concentrations [12,18], monthly ozone concentrations, air revenue passenger miles, and the larger multidimensional alabone dataset.

As baselines for comparison we consider the popular kernels implemented in the GPML toolbox [18]: linear with bias (LIN), SE, polynomial (Poly), PER, rational quadratic (RQ), MA, Gabor, fractional Brownian motion covariance

(FBM), underdamped linear Langevin process covariance (ULL), neural network (NN) and SM kernels. For the considered multidimensional dataset, we use automatic relevance determination (ARD) for other kernels to remove irrelevant input. FBM and ULL kernels are only available for time series type of data, thus they are not applied to this dataset. We use the GPML toolbox [17] and GPflow [14] for ordinary and scalable inference, respectively. For GCSM, we calculate the gradient of the parameters using an analytical derivative technique. In all experiments we use the hyper-parameter initialization previously described for SM and GCSM kenels.

### 8.1 Compact Long Term Extrapolation

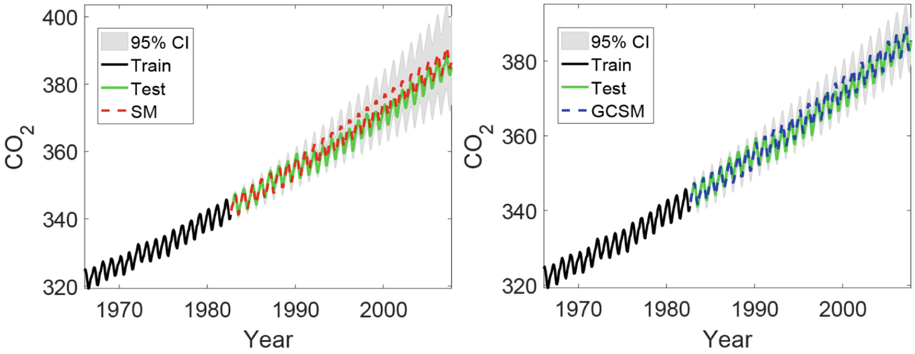

**Fig. 3.** Performance of SM (left) and GCSM (right) on the $CO_2$ concentration dataset.

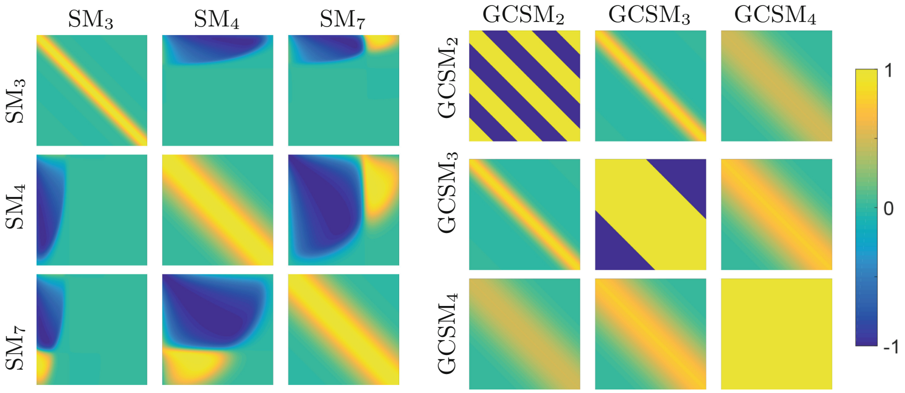

**Fig. 4.** Left: posterior correlations $\rho^*_{ij}$ in SM; Right: learned dependencies $\gamma_{ij}$ in GCSM.

The monthly average atmospheric $CO_2$ concentration dataset (cf. e.g. [18]) is a popular experiment which shows the advantage and flexibility of GPs due to multiple patterns with different scales in the data, such as long-term, seasonal and short-term trends. The dataset was collected at the Mauna Loa Observatory, Hawaii, between 1958 and 2003. We use 40% of the location points as training data and the rest 60% as testing data. For both GCSM and SM we consider $Q = 10$ components. The Gaussian mixture of the empirical spectral densities is considered to initialize the hyper-parameters.

Figure 3(a) shows that GCSM (in dashed blue) is better than ordinary SM (in red) in terms of predictive mean and variance. Moreover, GCSM yields a smaller confidence interval than SM. Unlike SM, GCSM does not overestimate the long-term trend. As for the analysis of the posterior correlation and learned dependency, evidence of posterior positive and negative correlations $\rho^*_{ij}$ can be observed for SM components (3, 4, 7) (left subplot in Fig. 4). These posterior correlations have been used for prediction (see Supplementary material). The right plot in Fig. 4 shows clear evidence of learned dependency $\gamma_{ij}$ for GCSM components (2, 3, 4). GCSM and SM are optimized independently, so component identifiers in the figures do not necessarily correspond to each other. Observe that plots for GCSM kernel with $i = j$ (right subplot) show stripes because of the normalization term in Eq. (19).

### 8.2 Modeling Irregular Long Term Decreasing Trends

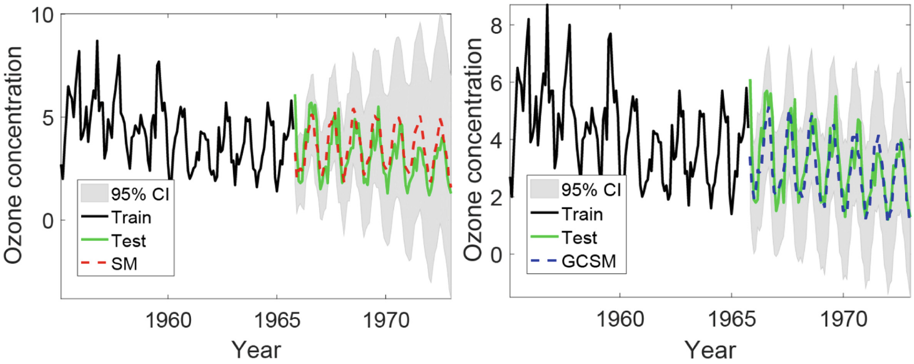

**Fig. 5.** Performance of SM (left) and GCSM (right) on the ozone concentration dataset.

We consider the monthly ozone concentration dataset (216 values) collected at Downtown L. A. from time range Jan 1955–Dec 1972. This dataset has different characteristics than the $CO_2$ concentration one, namely a gradual long term downtrend and irregular peak values in the training data which are much higher than those in the testing data. These characteristics make extrapolation a challenging task. Here we use the first 60% of observations for training, and the rest (40%) for testing (shown in black and green in Fig. 5, respectively). Again we consider $Q = 10$ components for both kernels.

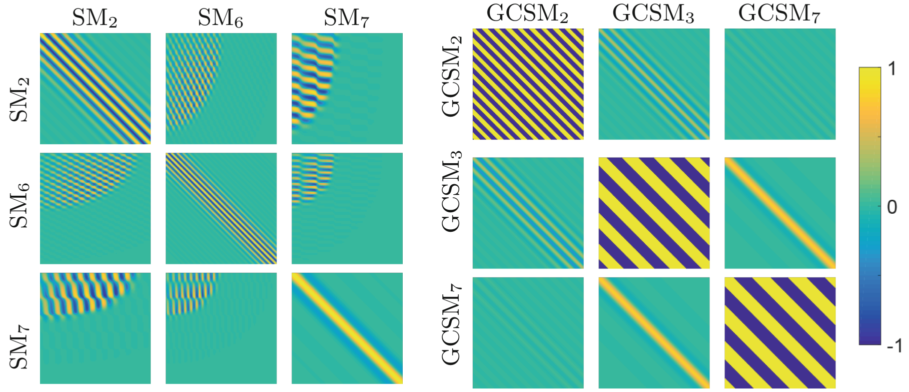

**Fig. 6.** Left: posterior correlations $\rho^*_{ij}$ in SM; Right: learned dependencies $\gamma_{ij}$ in GCSM.

Figure 5 shows that the ozone concentration signal has a long term decreasing tendency while the training part has a relatively stable evolution. Here SM fails to discover such long term decreasing tendency and overestimates the future trend with low confidence. Instead, GCSM is able to confidently capture the long term decreasing tendency. These results substantiate the beneficial effect of using cross-components for correcting overestimation and for reducing predictive uncertainty.

Results in Table 1 show that on this dataset GCSM consistently achieves a lower MSE compared with SM and other baselines.

Figure 6 shows posterior correlation (left plot) and learned dependency (right plot), The texture of the posterior correlation $\rho^*_{ij}$ among SM components (2, 6, 7) demonstrates a more complicated posterior correlation between these components than that of the previous experiment. The learned dependency $\gamma_{ij}$ is clearly visible between components (2, 3, 7).

### 8.3 Modeling Irregular Long Term Increasing Trends

In this experiment we consider another challenging extrapolation task, using the air revenue passenger miles[1] with time range Jan 2000–Apr 2018, monthly collected by the U.S. Bureau of Transportation Statistics. Given 60% recordings at the beginning of the time series, we wish to extrapolate the remaining observations (40%). In this setting we can observe an apparent long term oscillation tendency in the training observations which is not present in the testing data. As shown in Fig. 7, even if at the beginning (in 2001) there seems to be a decreasing trend due to 9/11 attack and since 2010 was known as a disappointing year for safety, there is a positive trend as a result of a boosting of the airline market and extensive globalization.

[1] https://fred.stlouisfed.org/series/AIRRPMTSI.

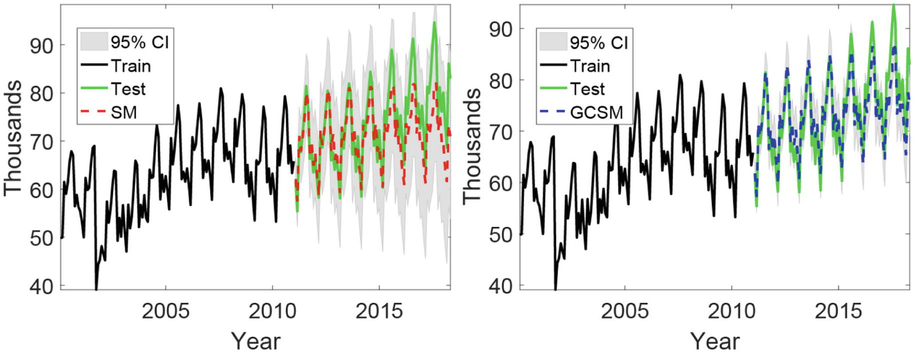

**Fig. 7.** Performance of SM (left) and GCSM (right) on air revenue passenger miles.

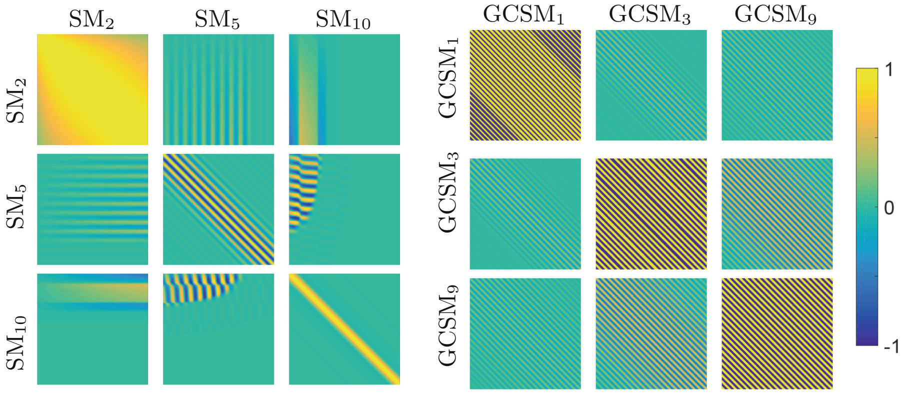

**Fig. 8.** Left: posterior correlations $\rho_{ij}^*$ in SM; Right: learned dependencies $\gamma_{ij}$ in GCSM.

In order to show the need for GCSM in a real-life scenarios, we consider the air revenue passenger miles dataset that contains a fake long term oscillation tendency happened in the training data but not in the testing data. The air revenue passenger miles[2] with time range Jan 2000–Apr 2018 was monthly collected by U.S. Bureau of Transportation Statistics.

Results in Table 1 show that on this dataset GCSM consistently achieves a lower MSE compared with SM and other baselines. In particular, kernels such as SE, Periodic and Matérn 5/2 have a poor performance on this extrapolation task.

In Fig. 8, the left plot shows the posterior correlation $\rho_{ij}^*$ among SM components (2, 5, 10), and the right subplot the learned dependency $\gamma_{ij}$ between components (1, 3, 9).

[2] https://fred.stlouisfed.org/series/AIRRPMTSI.

### 8.4 Prediction with Large Scale Multidimensional Data

After comparing GCSM and SM on extrapolation tasks on time series with diverse characteristics, we investigate comparatively its performance on a prediction task using a large multidimensional dataset, the abalone dataset. The dataset consists of 4177 instances with 8 attributes: Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight. The goal is to predict the age of an abalone from physical measurements. Abalone's age is measured by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. Thus the task is to predict the number of rings from the above mentioned attributes. We use the first 3377 instances as training data and the remaining 800 as testing data. For both GCSM and SM we used Q = 5 components. We use the windowed empirical density to initialize the hyper-parameters, as described in Sect. 7.1. Here components are multivariate Gaussian distributions in the frequency domain.

Results in Table 1 show that also on this type of task GCSM achieves lower MSE than SM.

**Table 1.** Performances between GCSM and other kernels. Left: MSE, right: NLML.

| Kernel | $CO_2$ | Ozon | Air | Abalone | Kernel | $CO_2$ | Ozon | Air | Abalone |
|---|---|---|---|---|---|---|---|---|---|
| LIN | 39.09 | 1.86 | 57.64 | 10.93 | LIN | 451.38 | 235.68 | 462.01 | 24261.35 |
| SE | 128502.50 | 10.40 | 4967.18 | 8.14 | SE | 399.90 | 208.53 | 456.68 | 21246.86 |
| Poly | 132369.70 | 11.36 | 5535.81 | 6.30 | Poly | 1444.80 | 375.86 | 735.39 | 17964.17 |
| PER | 53.37 | 3.87 | 276.07 | 7.98 | PER | 459.53 | 236.71 | 456.38 | 18775.23 |
| RQ | 985.39 | 1.86 | 168.33 | 5.38 | RQ | 222.17 | 196.96 | 430.86 | 15988.48 |
| MA | 110735.30 | 9.83 | 4711.33 | 7.52 | MA | 278.33 | 208.17 | 451.03 | 20288.56 |
| Gabor | 131931.30 | 2.09 | 5535.84 | 4.80 | Gabor | 1444.62 | 240.55 | 735.41 | 15400.84 |
| FBM | 193.18 | 2.56 | 172.01 | –.– | FBM | 910.61 | 202.42 | 457.792 | –.– |
| ULL | 117500.40 | 9.34 | 405.07 | –.– | ULL | 819.09 | 206.85 | 441.31 | –.– |
| NN | 326.81 | 1.69 | 116.66 | 5.60 | NN | 460.73 | 225.46 | 449.31 | 17695.80 |
| SM | 9.36 | 0.97 | 36.28 | 3.59 | SM | **62.09** | 160.75 | 328.56 | 8607.99 |
| GCSM | **1.19** | **0.59** | **10.02** | **3.29** | GCSM | 64.34 | **160.48** | **300.69** | **8566.35** |

SM and GCSM kernels achieve comparable performance in terms of NLML (see right part of Table 1). This seems surprising, given the smaller uncertainty and MSE results obtained by GCSM. However, note that NLML is the sum of two terms (and a constant term that is ignored): a model fit and a complexity penalty term. The first term is the data fit term which is maximized when the data fits the model very well. The second term is a penalty on the complexity of the model, i.e. the smoother the better. When Optimizing NLML finds a balance between the two and this changes with the data observed.

Overall, results indicate the beneficial effect of modeling directly dependencies between components, as done in our kernel.

## 9 Conclusion

We proposed the generalized convolution spectral mixture (GCSM) kernel, a generalization of SM kernels with an expressive closed form to modeling dependencies between components using cross convolution in the frequency domain.

Experiments on real-life datasets indicate that the proposed kernel, when used in GPs, can identify and model the complex structure of the data and be used to perform long-term trends forecasting. Although here we do not focus on non-stationary kernels, GCSM can be transformed into a non-stationary GCSM, through parameterizing weights $w_i(x)$, means $\mu_i(x)$, and $\sigma_i(x)$ as kernel matrices by means of a Gaussian function. Future work includes the investigation of more generalized non-stationary GCSM.

An issue that remains to be investigated is efficient inference. This is a core issue in GP methods which needs to be addressed also for GPs with GCSM kernels. Lev́y process priors as proposed in [11] present a promising approach for tackling this problem, by regularizing spectral mixture for automatic selection of the number of components and pruning of unnecessary components.

**Acknowledgements.** Part of this work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA19030301 and Shenzhen Discipline Construction Project for Urban Computing and Data Intelligence.

## References

1. Gold, B.: Theory and Application of Digital Signal Processing. Prentice-Hall, Upper Saddle River (1975)
2. Bochner, S.: Lectures on Fourier Integrals (AM-42), vol. 42. Princeton University Press, Princeton (2016)
3. Duvenaud, D.: Automatic model construction with Gaussian processes, Doctoral thesis. Ph.D. thesis (2014). https://doi.org/10.17863/CAM.14087
4. Duvenaud, D., Nickisch, H., Rasmussen, C.E.: Additive Gaussian processes. In: Neural Information Processing Systems, pp. 226–234 (2012)
5. Gaspari, G., Cohn, S.E.: Construction of correlation functions in two and three dimensions. Q. J. Roy. Meteorol. Soc. **125**(554), 723–757 (1999)
6. Genton, M.G., Kleiber, W., et al.: Cross-covariance functions for multivariate geostatistics. Stat. Sci. **30**(2), 147–163 (2015)
7. Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., Lähdesmäki, H.: Non-stationary Gaussian process regression with hamiltonian monte carlo. In: Artificial Intelligence and Statistics, pp. 732–740 (2016)
8. Hensman, J., Durrande, N., Solin, A.: Variational Fourier features for Gaussian processes. J. Mach. Learn. Res. **18**, 151:1–151:52 (2017)
9. Hensman, J., Fusi, N., Lawrence, N.D.: Gaussian processes for big data. In: Nicholson, A., Smyth, P. (eds.) Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, 11–15 August 2013. AUAI Press (2013)
10. Herlands, W., et al.: Scalable Gaussian processes for characterizing multidimensional change surfaces. In: Artificial Intelligence and Statistics, pp. 1013–1021 (2016)

11. Jang, P.A., Loeb, A., Davidow, M., Wilson, A.G.: Scalable Levy process priors for spectral kernel learning. In: Advances in Neural Information Processing Systems, pp. 3943–3952 (2017)
12. Keeling, C.D.: Atmospheric $CO_2$ records from sites in the SIO air sampling network. In: Trends' 93: A Compendium of Data on Global Change, pp. 16–26 (1994)
13. Majumdar, A., Gelfand, A.E.: Multivariate spatial modeling for geostatistical data using convolved covariance functions. Math. Geol. **39**(2), 225–245 (2007)
14. Matthews, A.G.D.G., et al.: GPflow: a Gaussian process library using tensorflow. J. Mach. Learn. Res. **18**(40), 1–6 (2017)
15. Moon, T.K.: The expectation-maximization algorithm. IEEE Signal Process. Mag. **13**(6), 47–60 (1997)
16. Quiñonero-Candela, J., Rasmussen, C.E.: A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. **6**, 1939–1959 (2005)
17. Rasmussen, C.E., Nickisch, H.: Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. **11**, 3011–3015 (2010)
18. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning. In: Adaptive Computation and Machine Learning. MIT Press (2006)
19. Remes, S., Heinonen, M., Kaski, S.: Non-stationary spectral kernels. In: Advances in Neural Information Processing Systems, pp. 4645–4654 (2017)
20. Snelson, E., Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs. In: Advances in Neural Information Processing Systems, pp. 1257–1264 (2006)
21. Snoek, J., Swersky, K., Zemel, R., Adams, R.: Input warping for Bayesian optimization of non-stationary functions. In: International Conference on Machine Learning, pp. 1674–1682 (2014)
22. Stein, M.: Interpolation of Spatial Data: Some Theory for Kriging (1999)
23. Williams, C.K., Seeger, M.: Using the Nyström method to speed up kernel machines. In: Advances in Neural Information Processing Systems, pp. 682–688 (2001)
24. Wilson, A., Adams, R.: Gaussian process kernels for pattern discovery and extrapolation. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, pp. 1067–1075 (2013)
25. Wilson, A.G., Gilboa, E., Nehorai, A., Cunningham, J.P.: Fast kernel learning for multidimensional pattern extrapolation. In: Advances in Neural Information Processing Systems, pp. 3626–3634 (2014)
26. Wilson, A.G.: Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes. University of Cambridge (2014)