Shen, Zheyang; Heinonen, Markus; Kaski, Samuel

Harmonizable mixture kernels with variational Fourier features

Published in:
The 22nd International Conference on Artificial Intelligence and Statistics

Published: 01/05/2019

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

Please cite the original version:
Shen, Z., Heinonen, M., & Kaski, S. (2019). Harmonizable mixture kernels with variational Fourier features. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1812-1821). (Proceedings of Machine Learning Research; Vol. 89). JMLR. http://proceedings.mlr.press/v89/shen19c/shen19c.pdf

Harmonizable mixture kernels with variational Fourier features

Zheyang Shen **Markus Heinonen** **Samuel Kaski**
Aalto University
Helsinki Institute for Information Technology HIIT

Abstract

The expressive power of Gaussian processes depends heavily on the choice of kernel. In this work we propose the novel harmonizable mixture kernel (HMK), a family of expressive, interpretable, non-stationary kernels derived from mixture models on the generalized spectral representation. As a theoretically sound treatment of non-stationary kernels, HMK supports harmonizable covariances, a wide subset of kernels including all stationary and many non-stationary covariances. We also propose variational Fourier features, an inter-domain sparse GP inference framework that offers a representative set of 'inducing frequencies'. We show that harmonizable mixture kernels interpolate between local patterns, and that variational Fourier features offers a robust kernel learning framework for the new kernel family.

1 INTRODUCTION

Kernel methods are one of the cornerstones of machine learning and pattern recognition. Kernels, as a measure of similarity between two objects, depart from common linear hypotheses by allowing for complex nonlinear patterns (Vapnik, 2013). In a Bayesian framework, kernels are interpreted probabilistically as covariance functions of random processes, such as for the Gaussian processes (GP) in Bayesian nonparametrics. As rich distributions over functions, GPs serve as an intuitive nonparametric inference paradigm, with well-defined posterior distributions.

The kernel of a GP encodes the prior knowledge of the underlying function. The *squared exponential* (SE) kernel is a common choice which, however, can only model global monotonic covariance patterns, while generalisations have explored local monotonicities (Gibbs, 1998; Paciorek and Schervish, 2004). In contrast, expressive kernels can learn hidden representations of the data (Wilson and Adams, 2013).

The two main approaches to construct expressive kernels are composition of simple kernel functions (Archambeau and Bach, 2011; Durrande et al., 2016; Gönen and Alpaydın, 2011; Rasmussen and Williams, 2006; Sun et al., 2018), and modelling of the spectral representation of the kernel (Wilson and Adams, 2013; Samo and Roberts, 2015; Remes et al., 2017). In the compositional approach kernels are composed of simpler kernels, whose choice often remains ad-hoc.

The spectral representation approach proposed by Quiñonero Candela et al. (2010), and extended by Wilson and Adams (2013), constructs *stationary* kernels as the Fourier transform of a Gaussian mixture, with theoretical support from the Bochner's theorem. Stationary kernels are unsuitable for large-scale datasets that are typically rife with locally-varying patterns (Samo and Roberts, 2016). Remes et al. (2017) proposed a practical *non-stationary* spectral kernel generalisation based on Gaussian process frequency functions, but with explicitly unclear theoretical foundations. An earlier technical report studied a non-stationary spectral kernel family derived via the generalised Fourier transform (Samo and Roberts, 2015). Samo (2017) expanded the analysis into non-stationary continuous bounded kernels.

The cubic time complexity of GP models significantly hinders their scalability. Sparse Gaussian process models (Herbrich et al., 2003; Snelson and Ghahramani, 2006; Titsias, 2009; Hensman et al., 2015) scale GP models with variational inference on pseudo-input points as a concise representation of the input data. Inter-domain Gaussian processes generalize sparse GP models by linearly transforming the original GP and computing cross-covariances, thus putting the inducing points on the transformed domain (Lázaro-Gredilla and Figueiras-Vidal, 2009).

In this paper we propose a theoretically sound treat-

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

Kernel	Harmonizable	Non-stationary	Spectral inference	Reference
SE: squared exponential	✓	✗	✓	Rasmussen and Williams (2006)
SS: sparse spectral	✓	✗	✓	Quiñonero Candela et al. (2010)
SM: spectral mixture	✓	✗	✓	Wilson and Adams (2013)
GSK: generalised spectral kernel	✓	✓	✗	Samo (2017)
GSM: generalised spectral mixture	?	✓	✗	Remes et al. (2017)
HMK: harmonizable mixture kernel	✓	✓	✓	current work

Table 1: Overview of proposed spectral kernels. The SE, SS and SM kernels are stationary with scalable spectral inference paradigms (Lázaro-Gredilla and Figueiras-Vidal, 2009; Quiñonero Candela et al., 2010; Gal and Turner, 2015). The GSM kernel is theoretically poorly defined with unknown harmonizable properties. HMK is well-defined with variational Fourier features as spectral inference.

ment of non-stationary kernels, with main contributions:

- We present a detailed analysis of *harmonizability*, a concept mainly existent in statistics literature. Harmonizable kernels are non-stationary kernels interpretable with their *generalized* spectral representations, similar to stationary ones.
- We propose practical *harmonizable mixture kernels* (HMK), a class of kernels dense in the set of harmonizable covariances with a mixture generalized spectral distribution.
- We propose *variational Fourier features*, an inter-domain GP inference framework for GPs equipped with HMK. Functions drawn from such GP priors have a well-defined Fourier transform, a desirable property not found in stationary GPs.

2 HARMONIZABLE KERNELS

In this section we introduce *harmonizability*, a generalization of stationarity largely unknown to the field of machine learning. We first define harmonizable kernel, and then analyze two existing special cases of harmonizable kernels, stationary and locally stationary kernels. We present a theorem demonstrating the expressiveness of previous stationary spectral kernels. Finally, we introduce Wigner transform as a tool to interpret and analyze these kernels.

Throughout the discussion in the paper, we consider complex-valued kernels with vectorial input $k(\mathbf{x}, \mathbf{x}') : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{C}$, and we denote vectors from the input (data) domain with symbols $\mathbf{x}, \mathbf{x}', \boldsymbol{\tau}, \mathbf{t}$, while we denote frequencies with symbols $\boldsymbol{\xi}, \boldsymbol{\omega}$.

2.1 Harmonizable kernel definition

A harmonizable kernel (Kakihara, 1985; Yaglom, 1987; Loève, 1994) is a kernel with a *generalized spectral distribution* defined by a generalized Fourier transform:

Definition 1. A complex-valued bounded continuous kernel $k : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{C}$ is *harmonizable* when it can be represented as

$$k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^D \times \mathbb{R}^D} e^{2i\pi(\boldsymbol{\omega}^\top \mathbf{x} - \boldsymbol{\xi}^\top \mathbf{x}')} \mu_{\Psi_k}(\mathrm{d}\boldsymbol{\omega}, \mathrm{d}\boldsymbol{\xi}), \quad (1)$$

where μ_{Ψ_k} is the Lebesgue-Stieltjes measure associated to some positive definite function $\Psi_k(\boldsymbol{\omega}, \boldsymbol{\xi})$ with bounded variations.

Harmonizability is a property shared by kernels and random processes with such kernels. The positive definite measure induced by function Ψ_k is defined as the generalized spectral distribution of the kernel, and when μ_{Ψ_k} is twice differentiable, the derivative $S_k(\boldsymbol{\omega}, \boldsymbol{\xi}) = \dfrac{\partial^2 \Psi_k}{\partial \boldsymbol{\omega} \partial \boldsymbol{\xi}}$ is defined as *generalized spectral density* (GSD).

Harmonizable kernel is a very general class in the sense that it contains a large portion of bounded, continuous kernels (See Table 1) with only a handful of (somewhat pathological) exceptions (Yaglom, 1987).

2.2 Comparison with Bochner's theorem

Stationary kernels are kernels whose value only depends on the distance $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$, and therefore is invariant to translation of the input. Bochner's theorem (Bochner, 1959; Stein, 2012) expresses similar relation between finite measures and kernels:

Theorem 1. *(Bochner) A complex-valued function $k : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{C}$ is the covariance function of a weakly stationary mean square continuous complex-valued random process on $\mathbb{R}^D$ if and only if it can be represented as*

$$k(\boldsymbol{\tau}) = \int_{\mathbb{R}^D} e^{2i\pi \boldsymbol{\omega}^\top \boldsymbol{\tau}} \psi_k(d\boldsymbol{\omega}). \quad (2)$$

where ψ_k is a positive finite measure.

Bochner's theorem draws duality between the space of finite measures to the space of stationary kernels. The *spectral distribution* ψ_k of a stationary kernel is the

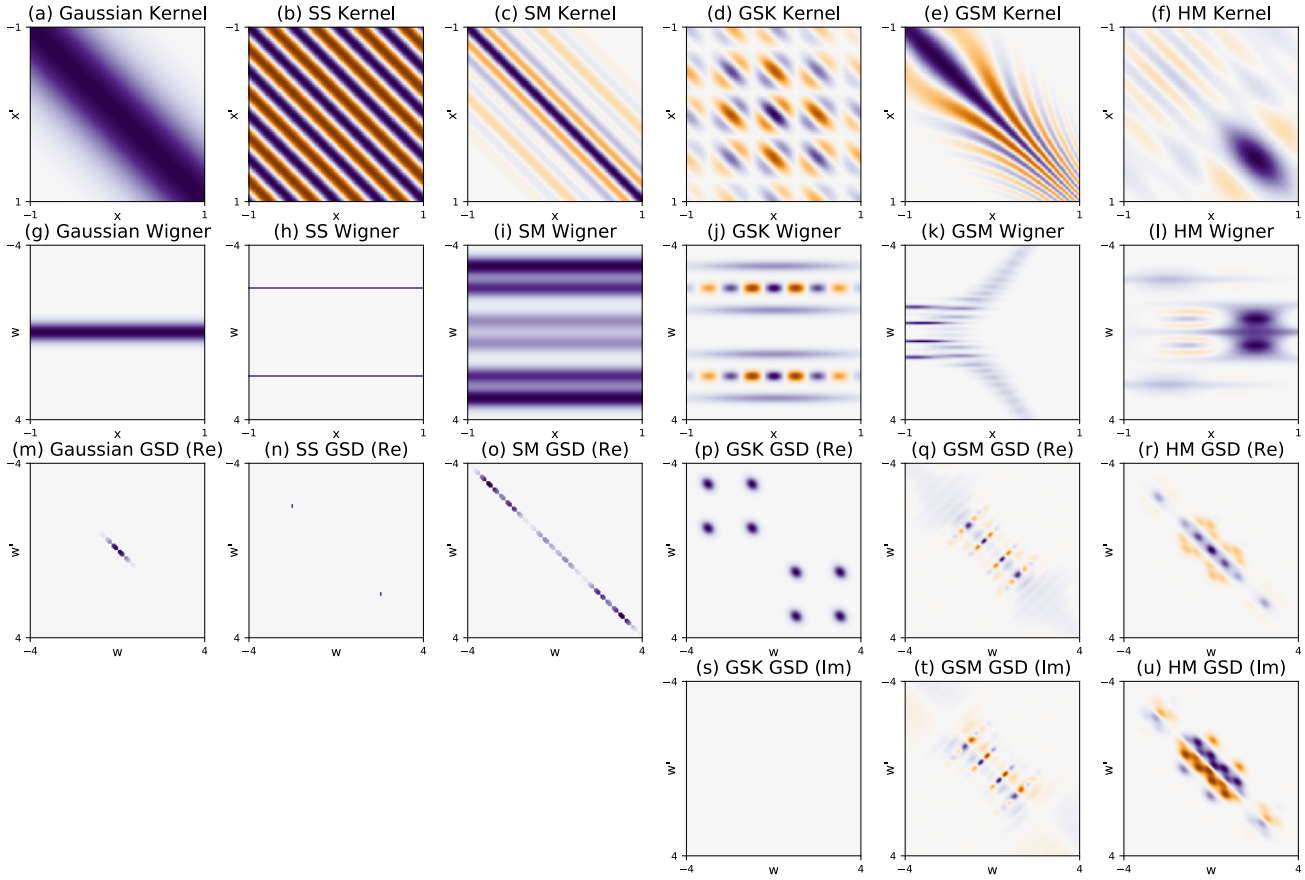

Figure 1: Comparison of Gaussian, SS, SM, GSK, GSM and HM kernels (columns) with respect to the kernel, Wigner distribution, and the generalized spectral density including real and imaginary part (rows).

finite measure induced by a Fourier transform. And when ψ_k is absolutely continuous with respect to the Lebesgue measure, its density is called *spectral density* (SD), $S_k(\boldsymbol{\omega}) = \dfrac{\mathrm{d}\psi_k(\boldsymbol{\omega})}{\mathrm{d}\boldsymbol{\omega}}$.

Harmonizable kernels include stationary kernels as a special case. When the mass of the measure μ_Ψ is concentrated on the diagonal $\boldsymbol{\omega} = \boldsymbol{\xi}$, the generalized inverse Fourier transform devolves into an inverse Fourier transform with respect to $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$, and therefore recovers the exact form in Bochner's theorem.

A key distinction between the two spectral distributions is that the spectral distribution is a nonnegative finite measure, but the generalized spectral distribution is a complex-valued measure with subsets assigned to complex numbers. Even with a real-valued harmonizable kernel, Ψ_k can be complex-valued.

2.3 Stationary spectral kernels

The perspective of viewing the spectral distribution as a normalized probability measure makes it possible to construct expressive stationary kernels by modeling their spectral distributions. Notable examples include the sparse spectrum (SS) kernel (Quiñonero Candela et al., 2010), and spectral mixture (SM) kernel (Wilson and Adams, 2013),

$$k_{SS}(\boldsymbol{\tau}) = \sum_{q=1}^{Q} \alpha_q \cos(2\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}), \tag{3}$$

$$k_{SM}(\boldsymbol{\tau}) = \sum_{q=1}^{Q} \alpha_q e^{-2\pi^2 \boldsymbol{\tau}^\top \boldsymbol{\Sigma}_q \boldsymbol{\tau}} \cos(2\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}), \tag{4}$$

with number of components $Q \in \mathbb{N}_+$, the component weights (amplitudes) $\alpha_q \in \mathbb{R}_+$, the (mean) frequencies $\boldsymbol{\omega}_q \in \mathbb{R}_+^D$, and the frequency covariances $\boldsymbol{\Sigma}_q \succeq \mathbf{0}$. Here we prove a theorem demonstrating the expressiveness of the above two kernels.

Theorem 2. *Let h be a complex-valued positive definite, continuous and integrable function. Then the family of* generalized spectral kernels

$$k_{GS}(\boldsymbol{\tau}) = \sum_{q=1}^{Q} \alpha_q h(\boldsymbol{\tau} \circ \boldsymbol{\gamma}_q) e^{2i\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}}, \tag{5}$$

is dense in the family of stationary, complex-valued kernels with respect to pointwise convergence of functions. Here $\circ$ denotes the Hadamard product, $\alpha_q \in \mathbb{R}_+$, $\boldsymbol{\omega}_k \in \mathbb{R}^D$, $\boldsymbol{\gamma}_k \in \mathbb{R}_+^D$, $Q \in \mathbb{N}_+$.

Proof sketch. We know that discrete measures are dense in the Banach space of finite measures. Therefore, the complex extension of sparse spectrum kernel $k_{SS}(\boldsymbol{\tau}) = \sum_{k=1}^{K} \alpha_k e^{2i\pi \boldsymbol{\omega}_k^\top \boldsymbol{\tau}}$ is dense in stationary kernels.

For each q, the function $\frac{\alpha_q}{h(0)} h(\boldsymbol{\tau} \circ \boldsymbol{\gamma}_q) e^{2i\pi \boldsymbol{\omega}_k^\top \boldsymbol{\tau}}$ converges to $\alpha_q e^{2i\pi \boldsymbol{\omega}_q^\top \boldsymbol{\tau}}$ pointwise as $\boldsymbol{\gamma}_q \downarrow \mathbf{0}$. Therefore, the proposed kernel form is dense in the set of sparse spectrum kernels, and by extension, stationary kernels. See Section 1 in supplementary materials for a more detailed proof. □

We strengthen the claim of Samo and Roberts (2015) by adding a constraint $\alpha_k > 0$ that restricts the family of functions to only valid PSD kernels (Samo, 2017). The spectral distribution of k_{GS} is

$$\psi_{k_{GS}}(\boldsymbol{\xi}) = \sum_{q=1}^{Q} \frac{\alpha_q}{\prod_{d=1}^{D} \gamma_{kd}} \psi_h((\boldsymbol{\xi} - \boldsymbol{\omega}_k) \oslash \boldsymbol{\gamma}_k), \tag{6}$$

with $\oslash$ denoting elementwise division of vectors. A real-valued kernel can be obtained by averaging a complex kernel with its complex conjugate, which induces a symmetry on the spectral distribution, $\psi_k(\boldsymbol{\xi}) = \psi_k(-\boldsymbol{\xi})$. For instance, the SM kernel has the symmetric Gaussian mixture spectral distribution

$$\psi_{k_{SM}}(\boldsymbol{\xi}) = \frac{1}{2} \sum_{q=1}^{Q} \alpha_q (\mathcal{N}(\boldsymbol{\xi} | \boldsymbol{\omega}_q, \boldsymbol{\Sigma}_q) + \mathcal{N}(\boldsymbol{\xi} | -\boldsymbol{\omega}_q, \boldsymbol{\Sigma}_q)). \tag{7}$$

2.4 Locally stationary kernels

As a generalization of stationary kernels, the locally stationary kernels (Silverman, 1957) are a simple yet unexplored concept in machine learning. A locally stationary kernel is a stationary kernel multiplied by a sliding power factor:

$$k_{LS}(\mathbf{x}, \mathbf{x}') = k_1 \left(\frac{\mathbf{x} + \mathbf{x}'}{2} \right) k_2(\mathbf{x} - \mathbf{x}'). \tag{8}$$

where $k_1 : \mathbb{R}^D \mapsto \mathbb{R}_{\geq 0}$ is an arbitrary nonnegative function, and $k_2 : \mathbb{R}^D \mapsto \mathbb{C}$ is a stationary kernel. k_1 is a function of the *centroid* between $\mathbf{x}$ and $\mathbf{x}'$, describing the scale of covariance on a global structure, while k_2 as a stationary covariance describes the local structure (Genton, 2001). It is straightforward to see that locally stationary kernels reduce into stationary kernels when k_1 is constant.

Integrable locally stationary kernels are of particular interest because they are harmonizable with a GSD. Consider a locally stationary Gaussian kernel (LSG) defined as a SE kernel multiplied by a Gaussian density on the centroid $\widetilde{\mathbf{x}} = (\mathbf{x} + \mathbf{x}')/2$. Its GSD can be obtained using the generalized Wiener-Khintchin relations (Silverman, 1957).

$$k_{\text{LSG}}(\mathbf{x}, \mathbf{x}') = e^{-2\pi^2 \widetilde{\mathbf{x}}^\top \boldsymbol{\Sigma}_1 \widetilde{\mathbf{x}}} e^{-2\pi^2 \boldsymbol{\tau}^\top \boldsymbol{\Sigma}_2 \boldsymbol{\tau}}, \tag{9}$$

$$S_{k_{\text{LSG}}}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \mathcal{N}\left(\frac{\boldsymbol{\omega} + \boldsymbol{\xi}}{2} \middle| 0, \boldsymbol{\Sigma}_2 \right) \mathcal{N}(\boldsymbol{\omega} - \boldsymbol{\xi} | 0, \boldsymbol{\Sigma}_1). \tag{10}$$

2.5 Interpreting spectral kernels

While the spectral distribution of a stationary kernel can be easily interpreted as a 'spectrum', the analogy does not apply to harmonizable kernels. In this section, we introduce the Wigner transform (Flandrin, 1998) which adds interpretability to kernels with spectral representations.

Definition 2. The *Wigner distribution function* (WDF) of a kernel $k(\cdot, \cdot) : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{C}$ is defined as $W_k : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$:

$$W_k(\mathbf{x}, \boldsymbol{\omega}) = \int_{\mathbb{R}^D} k\left(\mathbf{x} + \frac{\boldsymbol{\tau}}{2}, \mathbf{x} - \frac{\boldsymbol{\tau}}{2} \right) e^{-2i\pi \boldsymbol{\omega}^\top \boldsymbol{\tau}} \, \mathrm{d}\boldsymbol{\tau}. \tag{11}$$

The Wigner transform first changes the kernel form k into a function of the centroid of the input: $(\mathbf{x} + \mathbf{x}')/2$ and the lag $\mathbf{x} - \mathbf{x}'$, and then takes the Fourier transform of the lag. The Wigner distribution functions are fully equivalent to non-stationary kernels. Given the domain of WDF, we can view WDF as a 'spectrogram' demonstrating the relation between input and frequency. Converting an arbitrary kernel into its Wigner distribution sheds light into the frequency structure of the kernel (See Figure 1).

The WDFs of locally stationary kernels adhere to the intuitive notion of local stationarity where frequencies remain constant at a local scale. Take locally stationary Gaussian kernel k_{LSG} as an example:

$$W_{k_{\text{LSG}}}(\mathbf{x}, \boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\omega} | \mathbf{0}, \boldsymbol{\Sigma}_2) e^{-2\pi^2 \mathbf{x}^\top \boldsymbol{\Sigma}_1 \mathbf{x}}. \tag{12}$$

3 HARMONIZABLE MIXTURE KERNEL

In this section we propose a novel *harmonizable mixture kernel*, a family of kernels dense in harmonizable covariance functions. We present the kernel in an intentionally concise form, and refer the reader to the Section 2 in the Supplements for a full derivation.

3.1 Kernel form and spectral representations

The *harmonizable mixture kernel* (HMK) is defined with an additive structure:

$$k_{\text{HM}}(\mathbf{x}, \mathbf{x}') = \sum_{p=1}^{P} k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p), \tag{13}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k_{\text{LSG}}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p)\phi_p(\mathbf{x})^\top \mathbf{B}_p \phi_p(\mathbf{x}'), \tag{14}$$

where $P \in \mathbb{N}_+$ is the number of centers, $(\phi_p(\mathbf{x}))_{q=1}^{Q_p} = e^{2i\pi \boldsymbol{\mu}_{pq}^\top \mathbf{x}}$ are sinusoidal feature maps, $\mathbf{B}_p \succeq \mathbf{0}_{Q_p}$ are spectral amplitudes, $\boldsymbol{\gamma}_p \in \mathbb{R}_+^D$ are input scalings, $\mathbf{x}_p \in \mathbb{R}^D$ are input shifts, and $\boldsymbol{\mu}_{pq} \in \mathbb{R}^D$ are frequencies. It is easy to verify k_{HM} as a valid kernel, for each k_p is a product of an LSG kernel and an inner product with finite basis expansion of sinusoidal functions.

HMKs have closed form spectral representations such as *generalized spectral density* (See Section 2 in the Supplement for detailed derivation):

$$S_{k_{\text{HM}}}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \sum_{p=1}^{P} S_{k_p}(\boldsymbol{\omega}, \boldsymbol{\xi}) e^{-2i\pi \mathbf{x}_p^\top (\boldsymbol{\omega} - \boldsymbol{\xi})}, \tag{15}$$

$$S_{k_p}(\boldsymbol{\omega}, \boldsymbol{\xi}) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}^2} \sum_{1 \le i,j \le Q_p} b_{pij} S_{pij}(\boldsymbol{\omega}, \boldsymbol{\xi}), \tag{16}$$

$$S_{pij}(\boldsymbol{\omega}, \boldsymbol{\xi}) = S_{k_{\text{LSG}}}((\boldsymbol{\omega} - \boldsymbol{\mu}_{pi}) \oslash \boldsymbol{\gamma}_p, (\boldsymbol{\xi} - \boldsymbol{\mu}_{pj}) \oslash \boldsymbol{\gamma}_p). \tag{17}$$

The *Wigner distribution function* can be obtained in a similar fashion

$$W_{k_{\text{HM}}}(\mathbf{x}, \boldsymbol{\omega}) = \sum_{p=1}^{P} W_{k_p}(\mathbf{x} - \mathbf{x}_p, \boldsymbol{\omega}), \tag{18}$$

$$W_{k_p}(\mathbf{x}, \omega) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}} \sum_{1 \le i,j \le Q_p} W_{pij}(\mathbf{x}, \boldsymbol{\omega}), \tag{19}$$

$$W_{pij}(\mathbf{x}, \boldsymbol{\omega}) = W_{k_{\text{LSG}}}\left(\mathbf{x} \circ \boldsymbol{\gamma}_p, \left(\boldsymbol{\omega} - (\boldsymbol{\mu}_{pi} + \boldsymbol{\mu}_{pj})/2\right) \oslash \boldsymbol{\gamma}_p\right) \times \cos(2\pi(\boldsymbol{\mu}_{pi} - \boldsymbol{\mu}_{pj})^\top \mathbf{x}). \tag{20}$$

The kernel form, GSD and WDF both take a normal density form. It is straightforward to see $S_{k_{\text{HM}}}$ is PSD, and that $k_{\text{Hm}}(-\mathbf{x}, -\mathbf{x}')$ is the GSD of $S_{k_{\text{HM}}}$. A real-valued kernel k_r is obtained by averaging with its complex conjugate: $W_{k_r}(\mathbf{x}, \boldsymbol{\omega}) = W_{k_r}(\mathbf{x}, -\boldsymbol{\omega})$, $S_{k_r}(\boldsymbol{\omega}, \boldsymbol{\xi}) = S_{k_r}(-\boldsymbol{\omega}, -\boldsymbol{\xi})$.

3.2 Expressiveness of HMK

Similar to the construction of *generalized spectral kernels*, we can construct a generalized version k_h where k_{LSG} is replaced by k_{LS}, a locally stationary kernel with a GSD.

Theorem 3. *Given a continuous, integrable kernel k_{LS} with a valid* generalized spectral density*, the harmonizable mixture kernel*

$$k_h(\mathbf{x}, \mathbf{x}') = \sum_{p=1}^{P} k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p), \tag{21}$$

$$k_p(\mathbf{x}, \mathbf{x}') = k_{LS}(\mathbf{x} \circ \boldsymbol{\gamma}_p, \mathbf{x}' \circ \boldsymbol{\gamma}_p)\phi_p(\mathbf{x})^\top \mathbf{B}_p \phi_p(\mathbf{x}'), \tag{22}$$

is dense in the family of harmonizable covariances with respect to pointwise convergence of functions. Here $P \in \mathbb{N}_+$, $(\phi_p(\mathbf{x}))_q = e^{2i\pi \boldsymbol{\mu}_{pq}^\top \mathbf{x}}$, $q = 1, \ldots, Q_p$, $\boldsymbol{\gamma}_p \in \mathbb{R}_+^D$, $\mathbf{x}_p \in \mathbb{R}^D$, $\boldsymbol{\mu}_{pq} \in \mathbb{R}^D$, $\mathbf{B}_p$ as positive definite Hermitian matrices.

Proof. See Section 3 in the supplementary materials. □

4 VARIATIONAL FOURIER FEATURES

In this section we propose variational inference for the harmonizable kernels applied in Gaussian process models.

We assume a dataset of n input $X = \{\mathbf{x}_i\}_{i=1}^n$ and output $\mathbf{y} = \{y_i\} \in \mathbb{R}^n$ observations from some function $f(\mathbf{x})$ with a Gaussian observation model:

$$y = f(\mathbf{x}) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_y^2). \tag{23}$$

4.1 Gaussian processes

Gaussian processes (GP) are a family of Bayesian models that characterise distributions of functions (Rasmussen and Williams, 2006). We assume a zero-mean Gaussian process prior on a latent function $f(\mathbf{x}) \in \mathbb{R}$ over vector inputs $\mathbf{x} \in \mathbb{R}^D$

$$f(\mathbf{x}) \sim \mathcal{GP}(0, K(\mathbf{x}, \mathbf{x}')), \tag{24}$$

which defines a priori distribution over function values $f(\mathbf{x})$ with mean $\mathbb{E}[f(\mathbf{x})] = 0$ and covariance

$$\mathbf{cov}[f(\mathbf{x}), f(\mathbf{x}')] = K(\mathbf{x}, \mathbf{x}'). \tag{25}$$

A GP prior specifies that for any collection of n inputs X, the corresponding function values $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^\top \in \mathbb{R}^n$ are coupled by following a multivariate normal distribution $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{ff})$, where $\mathbf{K}_{ff} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n \times n}$ is the kernel matrix over input pairs. The key property of GP's is that output predictions $f(\mathbf{x})$ and $f(\mathbf{x}')$ correlate according to how similar are their inputs $\mathbf{x}$ and $\mathbf{x}'$ as defined by the kernel $K(\mathbf{x}, \mathbf{x}') \in \mathbb{R}$.

4.2 Variational inference with inducing features

In this section, we introduce variational inference of sparse GPs in an inter-domain setting. Consider a GP prior $f(\mathbf{x}) \sim \mathcal{GP}(0, k)$, and a valid linear transform $\mathcal{L}$ projecting f to another GP $\mathcal{L}_f(\mathbf{z}) \sim \mathcal{GP}(0, k')$.

We begin by *augmenting* the Gaussian process with $m < n$ inducing variables $u_j = \mathcal{L}_f(\mathbf{z}_j)$ using a Gaussian model. $\mathbf{z}_j$ are *inducing features* placed on the domain of $\mathcal{L}_f(\mathbf{z})$, with prior $p(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K}_{uu})$ and a conditional model (Hensman et al., 2015)

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{A}\mathbf{u}, \mathbf{K}_{ff} - \mathbf{A}\mathbf{K}_{uu}\mathbf{A}^\dagger), \tag{26}$$

where $\mathbf{A} = \mathbf{K}_{fu}\mathbf{K}_{uu}^{-1}$, and $\mathbf{A}^\dagger$ denotes the Hermitian transpose of $\mathbf{A}$ allowing for complex GPs. The kernel $\mathbf{K}_{uu}$ is between the $m \times m$ inducing variables and the kernel $\mathbf{K}_{fu}$ is the cross covariance of $\mathcal{L}$, $(\mathbf{K}_{fu})_{is} = \mathbf{cov}(f(\mathbf{x}_i), \mathcal{L}_f(\mathbf{z}_s))$. Next, we define a variational approximation $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m}, \mathbf{S})$ with the Gaussian interpolation model (26),

$$q(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\mathbf{A}\mathbf{m}, \mathbf{K}_{ff} - \mathbf{A}(\mathbf{S} - \mathbf{K}_{uu})\mathbf{A}^\dagger), \tag{27}$$

with free variational mean $\mathbf{m} \in \mathbb{R}^m$ and variational covariance $\mathbf{S} \in \mathbb{R}^{m \times m}$ to be optimised. Finally, variational inference (Blei et al., 2016) describes an evidence lower bound (ELBO) of augmented Gaussian processes as

$$\log p(\mathbf{y}) \geq \sum_{i=1}^{n} \mathbb{E}_{q(f_i)} \log p(y_i|f_i) - \mathrm{KL}[q(\mathbf{u})||p(\mathbf{u})]. \tag{28}$$

4.3 Fourier transform of a harmonizable GP

In this section, we compute cross-covariances between a GP and the Fourier transform of the GP. Consider a GP prior $f \sim \mathcal{GP}(0, k)$ where the kernel k is harmonizable with a GSD S_k and where $\hat{f}$ is the Fourier transform of f,

$$\hat{f}(\boldsymbol{\omega}) \triangleq \int_{\mathbb{R}^D} f(\mathbf{x}) e^{-2i\pi\boldsymbol{\omega}^\top \mathbf{x}} \, \mathrm{d}\mathbf{x}. \tag{29}$$

The validity of this setting is easily verified because f is square integrable on expectation,

$$\mathbb{E}\left\{\int_{\mathbb{R}^D} |f(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}\right\} = \int_{\mathbb{R}^D} k(\mathbf{x}, \mathbf{x}) \, \mathrm{d}\mathbf{x} < \infty. \tag{30}$$

We can therefore derive the cross-covariances

$$\mathbf{cov}(\hat{f}(\boldsymbol{\omega}), f(\mathbf{x})) = \int_{\mathbb{R}^D} k(\mathbf{t}, \mathbf{x}) e^{-2i\pi\boldsymbol{\omega}^\top \mathbf{t}} \, \mathrm{d}\mathbf{t} \tag{31}$$

$$\mathbf{cov}(\hat{f}(\boldsymbol{\omega}), \hat{f}(\boldsymbol{\xi})) = S_k(\boldsymbol{\omega}, \boldsymbol{\xi}). \tag{32}$$

The above derivation is valid for any harmonizable kernel with a GSD. The Fourier transform of $\mathcal{GP}(0, k)$ is a complex-valued GP with kernel S_k, which correlates to the original GP.

For harmonizable, integrable kernel k, we can construct an inter-domain sparse GP model defined in 4.2 by plugging in $\mathcal{L}_f = \hat{f}$.

4.4 Variational Fourier features of the harmonizable mixture kernel

HMK belongs to the kernel family discussed in 4.3, but we can utilize the additive structure of an HMK $k_{HM} = \sum_{p=1}^{P} k_p(\mathbf{x} - \mathbf{x}_p, \mathbf{x}' - \mathbf{x}_p)$. A GP with kernel k_{HM} can be decomposed into P independent GPs:

$$f(\mathbf{x}) = \sum_{p=1}^{P} f_p(\mathbf{x} - \mathbf{x}_p), \tag{33}$$

$$f_p(\mathbf{x}) \sim \mathcal{GP}(0, k_p(\mathbf{x}, \mathbf{x}')). \tag{34}$$

Given this formulation, we can derive *variational Fourier features* with inducing frequencies conditioned on one f_p. For the p^{th} component, we have m_p inducing frequencies $(\boldsymbol{\omega}_{p1}, \ldots, \boldsymbol{\omega}_{pm_p})$ and m_p inducing values $(u_{p1}, \cdots, u_{pm_p})$. We can compute inter-domain covariances in a similar fashion:

$$\begin{aligned} \mathbf{K}_{fu}(\boldsymbol{\omega}_{qj}, \mathbf{x}) &\triangleq \mathbf{cov}(f(x), u_{qj}) \\ &= \sum_{p=1}^{P} \mathbf{cov}(f_p(\mathbf{x} - \mathbf{x}_p), u_{qj}) \\ &= \mathbf{cov}(f_q(\mathbf{x} - \mathbf{x}_q), \hat{f}_q(\boldsymbol{\omega}_{qj})). \end{aligned} \tag{35}$$

Similarly, we compute entries of the matrix K_{uu}

$$\mathbf{K}_{uu}(\boldsymbol{\omega}_{pi}, \boldsymbol{\omega}_{qj}) \triangleq \mathbf{cov}(u_{pi}, u_{qj}) = \begin{cases} S_p(\boldsymbol{\omega}_{pi}, \boldsymbol{\omega}_{qj}), p = q, \\ 0, p \neq q. \end{cases} \tag{36}$$

The matrix $\mathbf{K}_{uu}$ allows for a block diagonal structure, which allows for faster matrix inversion. The variational Fourier features are then completed by plugging in entries in $\mathbf{K}_{fu}$ (35) and $\mathbf{K}_{uu}$ (36) into the evidence lower bound (28).

4.5 Connection to previous work

In this section we demonstrate that an inter-domain stationary GP with windowed Fourier transform (Lázaro-Gredilla and Figueiras-Vidal, 2009) is equivalent to a rescaled VFF with a tweaked kernel. GPs with stationary kernels do not have valid Fourier transform, therefore, previous attempts of using Fourier transforms of GPs have been accompanied by a window

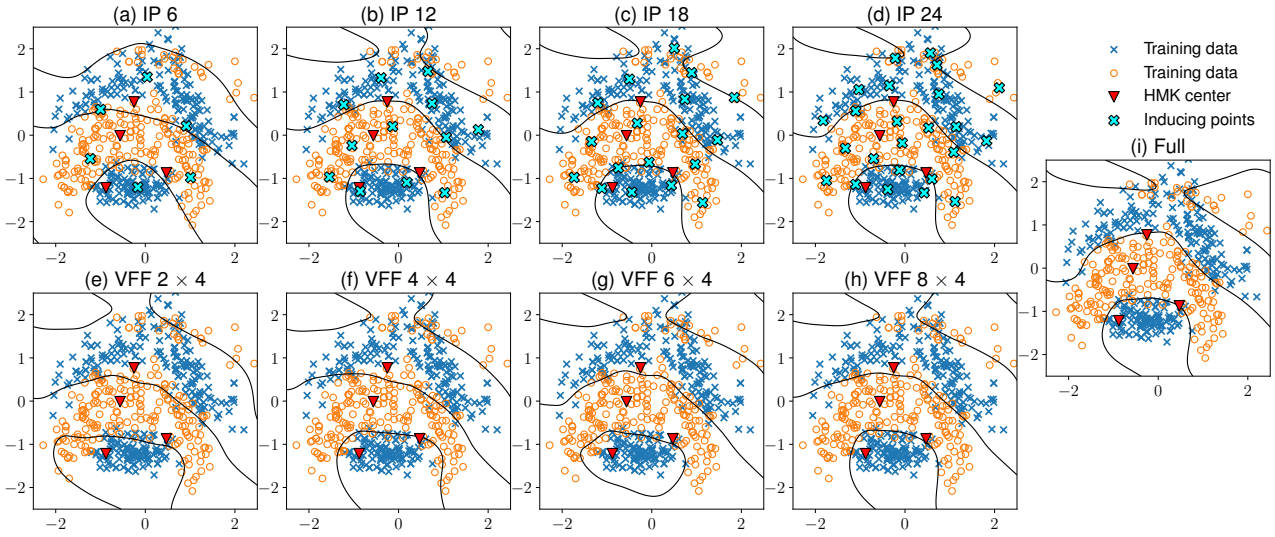

Figure 2: Sparse GP classification with the banana dataset. The model is learned by an HMK with $P = 4$ components, and thus 2 inducing frequencies for each component constitute a total of 2×4 inducing frequencies.

function:

$$\mathcal{L}_f(\boldsymbol{\omega}) = \int_{\mathbb{R}^D} f(\mathbf{x}) w(\mathbf{x}) e^{-2i\pi \boldsymbol{\omega}^\top \mathbf{x}} \, \mathrm{d}\mathbf{x}. \tag{37}$$

The windowing function $w(\mathbf{x})$ can be a soft Gaussian window $w(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Lázaro-Gredilla and Figueiras-Vidal, 2009) or a hard interval window $w(x) = \mathbb{I}_{[a \le x \le b]} e^{2i\pi a}$ (Hensman et al., 2017). The windowing approach shares the caveat of a blurred version of the frequency space, caused by an inaccurate Fourier transform(Lázaro-Gredilla and Figueiras-Vidal, 2009).

Consider $f \sim \mathcal{GP}(0, k)$ where k is a stationary kernel, and $w(\mathbf{x}) = \mathcal{N}(x|\mu, \boldsymbol{\Sigma})$, we see that $g(\mathbf{x}) = w(\mathbf{x}) f(\mathbf{x}) \sim \mathcal{GP}(0, w(\mathbf{x}) w(\mathbf{x}') k(\mathbf{x} - \mathbf{x}'))$. It is easy to verify that the kernel of $g(\mathbf{x})$ is locally stationary. There exist the following relations of cross-covariances:

$$\mathbf{cov}(f(\mathbf{x}), \mathcal{L}_f(\boldsymbol{\omega})) = \frac{\mathbf{cov}(g(\mathbf{x}), \hat{g}(\boldsymbol{\omega}))}{w(\mathbf{x})}, \tag{38}$$

$$\mathbf{cov}(\mathcal{L}_f(\boldsymbol{\omega}), \mathcal{L}_f(\boldsymbol{\xi})) = \mathbf{cov}(\hat{g}(\boldsymbol{\omega}), \hat{g}(\boldsymbol{\xi})). \tag{39}$$

Therefore, windowed inter-domain GPs are equivalent to rescaled GPs with a tweaked kernel.

5 EXPERIMENTS

In this section, we experiment with the harmonizable mixture kernels for kernel recovery, GP classification and regression. We use a simplied version of the harmonizable kernel where the two matrices of the locally stationary k_{LSG} are diagonals: $\boldsymbol{\Sigma}_1 = \mathrm{diag}(\sigma_d^2)$, $\boldsymbol{\Sigma}_2 = \lambda^2 I$. See Section 6 in the supplement for more detailed information.

5.1 Kernel recovery

We demonstrate the expressiveness of HMK by using it to recover certain non-stationary kernels. We choose the non-stationary *generalized spectral mixture kernel* (GSM) (Remes et al., 2017) and the covariance function of a time-inverted fractional Brownian motion (IFBM):

$$k_{\mathrm{GSM}}(x, x') = w(x) w(x') k_{\mathrm{Gibbs}}(x, x') \cos(2\pi(\mu(x)x - \mu(x')x')),$$
$$k_{\mathrm{Gibbs}}(x, x') = \sqrt{\frac{2l(x)l(x')}{l(x)^2 + l(x')^2}} \exp\left(-\frac{(x - x')^2}{l(x)^2 + l(x')^2}\right),$$
$$k_{\mathrm{IFBM}}(t, s) = \frac{1}{2}\left(\frac{1}{t^{2h}} + \frac{1}{s^{2h}} - \left|\frac{1}{t} - \frac{1}{s}\right|^{2h}\right),$$

where $s, t \in (0.1, 1.1]$ and $x, x' \in [-1, 1]$. The hyperparameters of k_{HM} are randomly initialized, and optimized with stochastic gradient descent.

Both kernels can be recovered almost perfectly with mean squared errors of 0.0033 and 0.0008. The result indicates that we can use the GSD and the Wigner distribution of the approximating HM kernel to interpret the GSM kernel (see Section 5 in supplementary materials).

5.2 GP classification with banana dataset

In this section, we show the effectiveness of variational Fourier fetures in GP classification with HMK. We use an HMK with $P = 4$ components to classify the banana dataset, and compare SVGP with inducing points (IP) (Hensman et al., 2015) and SVGP with variational Fourier features (VFF). The model parameters are learned by alternating optimization rounds of natural gradients for the variational parameters, and Adam

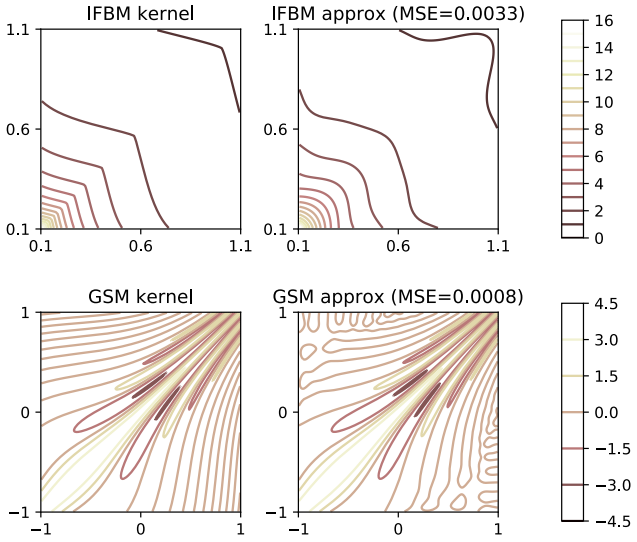

Figure 3: Kernel recovery experiment with true kernels (left) against SM kernel approximations (right).

optimizer for the other parameters (Salimbeni et al., 2018).

Figure 2 shows the decision boundaries of the two methods over the number of inducing points. For both variants, we experiment with model complexities from 6 to 24 inducing points in IP, and from 2 to 8 inducing frequencies for each component of HMK in the VFF. The centers of HMK (red triangles) spread to support the data distribution. The IP method is slightly more complex compared to VFF at the same parameter counts in terms of nonzero entries in the variational parameters.

The VFF method recovers roughly the correct decision boundary even with a small number of inducing frequencies, while converging faster to the decision boundaries as the number of inducing frequencies increases.

5.3 GP regression with solar irradiance

In this section, we demonstrate the effectiveness of HMK in interpolation for the non-stationary solar irradiance dataset. We run sparse GP regression with squared exponential, spectral mixture and harmonizable mixture kernels, and show the predicted mean, and 95% confidence intervals for each model (See Figure 2).

We use sparse GP regression proposed in (Titsias, 2009) with 50 inducing points marked at the x axis. The SE kernel can not estimate the periodic pattern and overestimates the signal smoothness. The SM kernel fits the training data well, but misidentifies frequencies on the first and fourth interval of the test set.

For sparse GP with HMK, we use the same framework

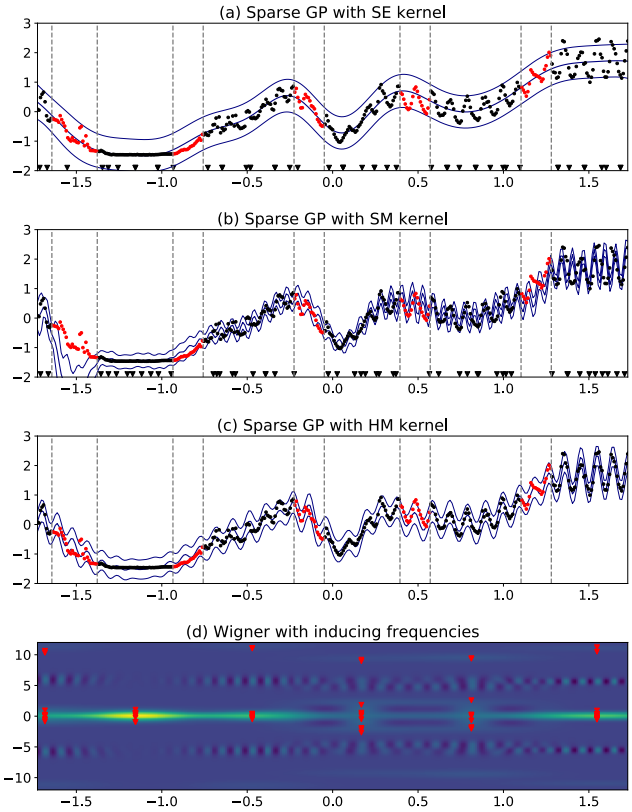

Figure 4: Sparse GP regression with solar irradiance dataset.

where the variational lower bound is adjusted for VFF. The model extrapolates better for the added flexibility of nonstationarity, and the inducing frequencies aggregate near the learned frequencies. Both first and last test intervals are well fitted. The Wigner distribution with inducing frequencies of the optimised HM kernel is shown in Figure 2d.

6 CONCLUSION

In this paper, we extend the generalization of Gaussian processes by proposing harmonizable mixture kernel, a non-stationary kernel spanning the wide class of harmonizable covariances. Such kernels can be used as an expressive tool for GP models. We also proposed variational Fourier features, an inter-domain inference framework used as drop-in replacements for sparse GPs. This work bridges previous research on spectral representation of kernels and sparse Gaussian processes.

Despite its expressiveness, one may brand the parametric form of HMK as not fully Bayesian, since it contradicts the nonparametric nature of GPs. A fully Bayesian approach would be to place a nonparametric prior over harmonizable mixture kernels, representing the uncertainty of the kernel form (Shah et al., 2014).

Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT. This work has been supported by the Academy of Finland grants no. 299915, 319264, 313195, 294238.

References

Cedric Archambeau and Francis Bach. Multiple Gaussian process models. *arXiv preprint arXiv:1110.5238*, 2011.

D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2016.

Salomon Bochner. *Lectures on Fourier Integrals: With an Author's Supplement on Monotonic Functions, Stieltjes Integrals and Harmonic Analysis; Translated from the Original German by Morris Tenenbaum and Harry Pollard*. Princeton University Press, 1959.

Nicolas Durrande, James Hensman, Magnus Rattray, and Neil D. Lawrence. Detecting periodicities with Gaussian processes. *PeerJ Computer Science*, 2:e50, 2016.

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

Yarin Gal and Richard Turner. Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In *International Conference on Machine Learning*, pages 655–664, 2015.

Marc G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2(Dec):299–312, 2001.

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

Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268, 2011.

James Hensman, Alexander G. de G. Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. *AISTATS*, 2015.

James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *The Journal of Machine Learning Research*, 18(1): 5537–5588, 2017.

Ralf Herbrich, Neil D. Lawrence, and Matthias Seeger. Fast sparse Gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems*, pages 625–632, 2003.

Y. Kakihara. A note on harmonizable and v-bounded processes. *Journal of Multivariate Analysis*, 16:140–156, 1985.

Miguel Lázaro-Gredilla and Anibal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *Advances in Neural Information Processing Systems*, pages 1087–1095, 2009.

Michel Loève. Probability theory II (graduate texts in mathematics), 1994.

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

Joaquin Quiñonero Candela, Carl Edward Rasmussen, Aníbal R. Figueiras-Vidal, et al. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11(Jun):1865–1881, 2010.

C.E. Rasmussen and K.I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

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

Hugh Salimbeni, Stefanos Eleftheriadis, and James Hensman. Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. *arXiv preprint arXiv:1803.09151*, 2018.

Yves-Laurent Kom Samo. *Advances in kernel methods: towards general-purpose and scalable models*. PhD thesis, University of Oxford, 2017.

Yves-Laurent Kom Samo and Stephen Roberts. Generalized spectral kernels. *arXiv preprint arXiv:1506.02236*, 2015.

Yves-Laurent Kom Samo and Stephen J. Roberts. String and membrane Gaussian processes. *The Journal of Machine Learning Research*, 17(1):4485–4571, 2016.

Amar Shah, Andrew Wilson, and Zoubin Ghahramani. Student-t processes as alternatives to Gaussian processes. In *Artificial Intelligence and Statistics*, pages 877–885, 2014.

R. Silverman. Locally stationary random processes. *IRE Transactions on Information Theory*, 3(3):182–187, 1957.

Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2006.

Michael L. Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Shengyang Sun, Guodong Zhang, Chaoqi Wang, Wenyuan Zeng, Jiaman Li, and Roger Grosse. Differentiable compositional kernel learning for Gaussian processes. *arXiv preprint arXiv:1806.04326*, 2018.

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

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Andrew Wilson and Ryan Adams. Gaussian rocess kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

A. M. Yaglom. *Correlation theory of stationary and related random functions: Volume I: Basic results*. Springer Series in Statistics. Springer, 1987.