Shen, Zheyang; Heinonen, Markus; Kaski, Samuel

Harmonizable mixture kernels with variational Fourier features

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

Published: 01/05/2019

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

Please cite the original version:
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 2001). 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 (Achambeau 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-
 Harmonizable mixture kernels with variational Fourier features

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Harmonizable</th>
<th>Non-stationary</th>
<th>Spectral inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE: squared exponential</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>SS: sparse spectral</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>SM: spectral mixture</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td>GSK: generalised spectral kernel</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>GSM: generalised spectral mixture</td>
<td>?</td>
<td>✓</td>
<td>×</td>
</tr>
<tr>
<td>HMK: harmonizable mixture kernel</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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; Quinonero 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 \( k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{C} \), and we denote vectors from the input (data) domain with symbols \( \mathbf{x}, \mathbf{x}' \), \( \tau, \mathbf{t} \), while we denote frequencies with symbols \( \mathbf{\xi}, \omega \).

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 \( k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{C} \) is **harmonizable** when it can be represented as

\[
k(\mathbf{x}, \mathbf{x}') = \int_{\mathbb{R}^D \times \mathbb{R}^D} e^{2\pi i (\omega \cdot \mathbf{x} + \mathbf{\xi} \cdot \mathbf{x}')} \mu_{\Psi_k}(d\omega, d\mathbf{\xi}),
\]

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

Harmonizability is a property shared by kernels and random processes with such kernels. The positive definite measure induced by function \( \Psi_k \) is defined as the generalized spectral distribution of the kernel, and when \( \mu_{\Psi_k} \) is twice differentiable, the derivative \( S_k(\omega, \mathbf{\xi}) = \frac{\partial^2 \Psi_k}{\partial \omega \partial \mathbf{\xi}} \) 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 \( \tau = \mathbf{x} - \mathbf{x}' \), and therefore is invariant to translation of the input. Bochner’s theorem (Bochner, 1932; Stein, 2012) expresses similar relation between finite measures and kernels:

**Theorem 1.** (Bochner) A complex-valued function \( k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{C} \) 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

\[
k(\tau) = \int_{\mathbb{R}^D} e^{2\pi i \omega^\top \tau} \psi_k(d\omega),
\]

where \( \psi_k \) 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 \( \psi_k \) of a stationary kernel is the
finite measure induced by a Fourier transform. And when \( \psi_k \) is absolutely continuous with respect to the Lebesgue measure, its density is called spectral density (SD), \( S_k(\omega) = \frac{d\psi_k(\omega)}{d\omega} \).

Harmonizable kernels include stationary kernels as a special case. When the mass of the measure \( \mu_\Psi \) is concentrated on the diagonal \( \omega = \xi \), the generalized inverse Fourier transform devolves into an inverse Fourier transform with respect to \( \tau = x - x' \), 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, \( \Psi_k \) 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),

\[
\begin{align*}
    k_{SS}(\tau) &= \sum_{q=1}^{Q} \alpha_q \cos(2\pi \omega_q^\top \tau), \\
    k_{SM}(\tau) &= \sum_{q=1}^{Q} \alpha_q e^{-2\pi^2 \tau^\top \Sigma_q \tau} \cos(2\pi \omega_q^\top \tau),
\end{align*}
\]

with number of components \( Q \in \mathbb{N}_+ \), the component weights (amplitudes) \( \alpha_q \in \mathbb{R}_+ \), the (mean) frequencies \( \omega_q \in \mathbb{R}_+^D \), and the frequency covariances \( \Sigma_q \succeq 0 \). 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

\[
k_{GS}(\tau) = \sum_{q=1}^{Q} \alpha_q h(\tau \circ \gamma_q)e^{2i\pi \omega_q^\top \tau},
\]
is dense in the family of stationary, complex-valued kernels with respect to pointwise convergence of functions. Here $\circ$ denotes the Hadamard product, $\alpha_q \in \mathbb{R}_+$, $\omega_q \in \mathbb{R}^+_D$, $\gamma_q \in \mathbb{R}^+_D$, $Q \in \mathbb{N}_+$.

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

For each $q$, the function $\frac{\alpha_q}{h(0)} h(\alpha_q) e^{2i\pi \omega_q^\top \tau}$ converges to $\alpha_q e^{2i\pi \omega_q^\top \tau}$ pointwise as $\gamma_q \downarrow 0$. 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. \qed

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

$$
\psi_{k_{GS}}(\xi) = \sum_{q=1}^Q \prod_{d=1}^D \alpha_k \psi_k((\xi - \omega_q) \circ \gamma_k),
$$

with $\circ$ 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, $\psi_k(\xi) = \psi_k(-\xi)$. For instance, the SM kernel has the symmetric Gaussian mixture spectral distribution

$$
\psi_{k_{SM}}(\xi) = \frac{1}{2} \sum_{q=1}^Q \alpha_q (\mathcal{N}(\xi | \omega_q, \Sigma_q) + \mathcal{N}(\xi | -\omega_q, \Sigma_q)).
$$

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:

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

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 $x$ and $x'$, 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 $\bar{x} = (x + x')/2$. Its GSD can be obtained using the generalized Wiener-Khintchin relations [Silverman 1957].

$$
k_{LSG}(x, x') = e^{-2\pi^2 \bar{x}^\top \Sigma \bar{x}} e^{-2\pi^2 \tau^\top \Sigma_2 \tau},
$$

$$
S_{k_{LSG}}(\omega, \xi) = \mathcal{N} \left( \frac{\omega + \xi}{2} \right) \left| \mathcal{N} \left( \omega - \xi | 0, \Sigma_1 \right) \right|
$$

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 $W_k: \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$:

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

The Wigner transform first changes the kernel form $k$ into a function of the centroid of the input: $(x + x')/2$ and the lag $x - x'$, 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 relationship 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:

$$
W_{k_{LSG}}(x, \omega) = \mathcal{N}(\omega | 0, \Sigma_2) e^{-2\pi^2 \bar{x}^\top \Sigma_1 \bar{x}}.
$$

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:

\[ k_{\text{HM}}(x, x') = \sum_{p=1}^{P} k_p(x - x_p, x' - x_p), \]  
\[ k_p(x, x') = k_{\text{LSG}}(x \circ \gamma_p, x' \circ \gamma_p) \phi_p(x)^T B_p \phi_p(x'), \]  

where \( P \in \mathbb{N}^+ \) is the number of centers, \( (\phi_p(x))_{q=1}^{Q_p} = e^{2\pi i \mu_p^q x} \) are sinusoidal feature maps, \( B_p \geq 0_{Q_p \times Q_p} \) are spectral amplitudes, \( \gamma_p \in \mathbb{R}^D \) are input scalings, \( x_p \in \mathbb{R}^D \) are input shifts, and \( \mu_{pq} \in \mathbb{R}^D \) are frequencies. It is easy to verify \( k_{\text{HM}} \) 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):

\[ S_{k_{\text{HM}}} (\omega, \xi) = \sum_{p=1}^{P} S_{k_p}(\omega, \xi) e^{-2 \pi i x_p^T \gamma_p (\omega - \xi)}, \]  
\[ S_{k_p}(\omega, \xi) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}^2} \sum_{1 \leq i, j \leq Q_p} b_{pij} S_{pij}(\omega, \xi), \]  
\[ S_{pij}(\omega, \xi) = S_{k_{\text{LSG}}}( (\omega - \mu_p) \otimes \gamma_p, (\xi - \mu_p) \otimes \gamma_p). \]

The Wigner distribution function can be obtained in a similar fashion

\[ W_{k_{\text{HM}}}(x, \omega) = \sum_{p=1}^{P} W_{k_p}(x - x_p, \omega), \]  
\[ W_{k_p}(x, \omega) = \frac{1}{\prod_{d=1}^{D} \gamma_{pd}^2} \sum_{1 \leq i, j \leq Q_p} W_{pij}(x, \omega), \]  
\[ W_{pij}(x, \omega) = W_{k_{\text{LSG}}}(x \circ \gamma_p, (\omega - (\mu_p + \mu_p)/2) \otimes \gamma_p) \times \cos(2\pi (\mu_p - \mu_p)^T x). \]

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 \( k_{\text{HM}}(-x, -x') \) is the GSD of \( S_{k_{\text{HM}}} \). A real-valued kernel \( k_r \) is obtained by averaging with its complex conjugate: \( W_{k_r}(x, \omega) = W_{k_r}(x, -\omega), S_{k_r}(\omega, \xi) = S_{k_r}(-\omega, -\xi). \)

3.2 Expressiveness of HMK

Similar to the construction of generalized spectral kernels, we can construct a generalized version \( k_h \) where \( k_{\text{LSG}} \) is replaced by \( k_{\text{LS}} \), 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

\begin{align*}
  k_h(x, x') = & \sum_{p=1}^{P} k_p(x - x_p, x' - x_p), \\
  k_p(x, x') = & k_{LS}(x \circ \gamma_p, x' \circ \gamma_p) \phi_p(x)^T B_p \phi_p(x'),
\end{align*}

is dense in the family of harmonizable covariances with respect to pointwise convergence of functions. Here \( P \in \mathbb{N}^+, (\phi_p(x))_{q=1}^{Q_p} = e^{2\pi i \mu_p^q x}, q = 1, \ldots, Q_p, \gamma_p \in \mathbb{R}_+, x_p \in \mathbb{R}^D, \mu_{pq} \in \mathbb{R}^D, B_p \) as positive definite Hermitian matrices.

**Proof.** See Section 3 in the supplementary materials. \( \square \)

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 \( X = \{x_i\}_{i=1}^{n} \) and output \( y = \{y_i\} \in \mathbb{R}^n \) observations from some function \( f(x) \) with a Gaussian observation model:

\[ y = f(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_y^2). \]

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 \( f(x) \in \mathbb{R}^n \) over vector inputs \( x \in \mathbb{R}^D \)

\[ f(x) \sim \mathcal{GP}(0, K(x, x')), \]

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

\[ \text{cov}[f(x), f(x')] = K(x, x'). \]

A GP prior specifies that for any collection of \( n \) inputs \( X \), the corresponding function values \( f = (f(x_1), \ldots, f(x_n))^T \in \mathbb{R}^n \) are coupled by following a multivariate normal distribution \( f \sim \mathcal{N}(0, K_{ff}) \), where \( K_{ff} = (K(x_i, x_j))_{i,j=1}^{n} \in \mathbb{R}^{n \times n} \) is the kernel matrix over input pairs. The key property of GP’s is that output predictions \( f(x) \) and \( f(x') \) correlate according to how similar are their inputs \( x \) and \( x' \) as defined by the kernel \( K(x, x') \in \mathbb{R} \).
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 $f(x) \sim \mathcal{GP}(0, k)$, and a valid linear transform $L$ projecting $f$ to another GP $\mathcal{L}_f(z) \sim \mathcal{GP}(0, k')$.

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

$$p(f|u) = \mathcal{N}(Au, K_{ff} - AK_{uu}A^\dagger),$$

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

$$q(f) = \mathcal{N}(f|Am, K_{ff} - A(S - K_{uu})A^\dagger),$$

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

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

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 $f$,

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

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

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

We can therefore derive the cross-covariances

$$\text{cov}(\hat{f}(\omega), f(x)) = \int_{\mathbb{R}^D} k(t, x)e^{-2\pi i \omega^\top t} \, dt$$

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

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 $k_{HM} = \sum_{p=1}^P k_p(x - x_p, x' - x_p)$. A GP with kernel $k_{HM}$ can be decomposed into $P$ independent GPs:

$$f(x) = \sum_{p=1}^P f_p(x - x_p),$$

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

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 ($\omega_{p1}, \ldots, \omega_{pm_p}$) and $m_p$ inducing values ($u_{p1}, \ldots, u_{pm_p}$). We can compute inter-domain covariances in a similar fashion:

$$K_{fu}(\omega_{qj}, x) \triangleq \text{cov}(f(x), u_{qj})$$

$$= \sum_{p=1}^P \text{cov}(f_p(x - x_p), u_{qj})$$

$$= \text{cov}(f_q(x - x_q), \hat{f}(\omega_{qj})).$$

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

$$K_{uu}(\omega_{pi}, \omega_{qj}) \triangleq \text{cov}(u_{pi}, u_{qj}) = \begin{cases} S_p(\omega_{pi}, \omega_{qj}), & p = q, \\ 0, & p \neq q. \end{cases}$$

The matrix $K_{uu}$ allows for a block diagonal structure, which allows for faster matrix inversion. The variational Fourier features are then completed by plugging in entries in $K_{fu}$ [35] and $K_{uu}$ [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
function:

\[ \mathcal{L}_f(\omega) = \int_{\mathbb{R}^D} f(x)w(x)e^{-2i\pi \omega^T x} \, dx. \]  (37)

The windowing function \( w(x) \) can be a soft Gaussian window \( w(x) = N(x|\mu, \Sigma) \) (Lázaro-Gredilla and Figueiras-Vidal 2009) or a hard interval window \( w(x) = 1_{[a \leq x \leq b]}e^{2\pi i \omega x} \) (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 \( w(x) = N(x|\mu, \Sigma) \), we see that \( g(x) = w(x)f(x) \sim \mathcal{GP}(0, w(x)w(x')k(x-x')) \). It is easy to verify that the kernel of \( g(x) \) is locally stationary. There exist the following relations of cross-covariances:

\[ \text{cov}(f(x), L_f(\omega)) = \frac{\text{cov}(g(x), \hat{g}(\omega))}{w(x)}, \]  (38)

\[ \text{cov}(L_f(\omega), L_f(\xi)) = \text{cov}(\hat{g}(\omega), \hat{g}(\xi)). \]  (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 simplified version of the harmonizable kernel where the two matrices of the locally stationary \( k_{LSG} \) are diagonals: \( \Sigma_1 = \text{diag}(\sigma_1^2) \), \( \Sigma_2 = \lambda^2 I \). 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) (Reyes et al. 2017) and the covariance function of a time-inverted fractional Brownian motion (IFBM):

\[ k_{GSM}(x,x') = w(x)w(x')k_{GSM}(x,x') \cos(2\pi(\mu(x)x - \mu(x')x')), \]

\[ k_{IFBM}(x,x') = \frac{2(2\pi)^l l(x')^2}{l(x)^2 + l(x')^2} \exp\left(-\frac{(x-x')^2}{l(x)^2 + l(x')^2}\right), \]

\[ k_{IFBM}(t,s) = \frac{1}{2}\left(\frac{1}{2\pi^2} + \frac{1}{2\pi} \left|\frac{1}{2} - \frac{1}{2}\right]\right), \]

where \( s,t \in (0,1,1) \) 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 features 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
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 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


