Gaussian processes with skewed Laplace spectral mixture kernels for long-term forecasting

Kai Chen[1,2] · Twan van Laarhoven[1] · Elena Marchiori[1]

Abstract

Long-term forecasting involves predicting a horizon that is far ahead of the last observation. It is a problem of high practical relevance, for instance for companies in order to decide upon expensive long-term investments. Despite the recent progress and success of Gaussian processes (GPs) based on spectral mixture kernels, long-term forecasting remains a challenging problem for these kernels because they decay exponentially at large horizons. This is mainly due to their use of a mixture of Gaussians to model spectral densities. Characteristics of the signal important for long-term forecasting can be unravelled by investigating the distribution of the Fourier coefficients of (the training part of) the signal, which is non-smooth, heavy-tailed, sparse, and skewed. The heavy tail and skewness characteristics of such distributions in the spectral domain allow to capture long-range covariance of the signal in the time domain. Motivated by these observations, we propose to model spectral densities using a skewed Laplace spectral mixture (SLSM) due to the skewness of its peaks, sparsity, non-smoothness, and heavy tail characteristics. By applying the inverse Fourier Transform to this spectral density we obtain a new GP kernel for long-term forecasting. In addition, we adapt the lottery ticket method, originally developed to prune weights of a neural network, to GPs in order to automatically select the number of kernel components. Results of extensive experiments, including a multivariate time series, show the beneficial effect of the proposed SLSM kernel for long-term extrapolation and robustness to the choice of the number of mixture components.

Keywords Kernels for Gaussian processes · Skewed Laplace distribution · Empirical spectral densities · Long-range forecasting
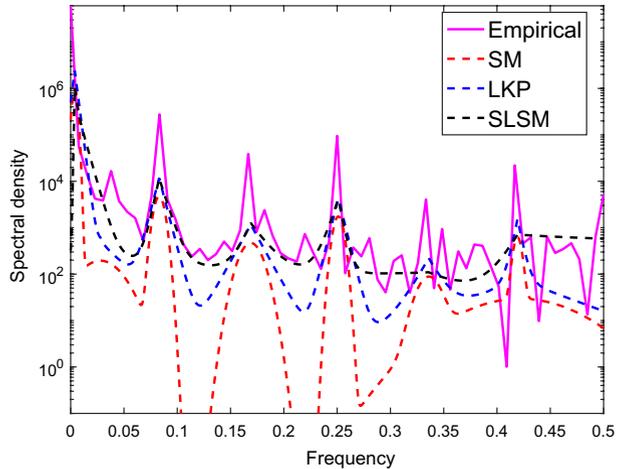
Fig. 1 Empirical spectral density (magenta solid line) of the rail passenger miles (training) time series dataset described in Sect. 6.1. SM kernel (red dashed line), LKP kernel (blue dashed), and SLSM kernel (black dashed) fit (obtained with the least squares method) of this empirical spectral density (Color figure online)


1 Introduction

Gaussian processes (GPs) are an elegant Bayesian approach to modeling an unknown function and representing its predictive uncertainty. They provide regression models where a posterior distribution over the unknown function is maintained as evidence is accumulated. This allows GPs to learn involved functions when a large amount of evidence is available, and makes them robust against overfitting in the presence of little evidence (Rasmussen and Nickisch 2010; Rasmussen and Williams 2006). A GP can model a large class of phenomena through the choice of its kernel, which characterizes one's assumption about the autocovariance of the unknown function. The choice of the kernel is a core step when designing a GP, since the posterior distribution can significantly vary for different kernels.

Important advances in the design of kernel functions for GPs have been achieved in recent years. In particular, kernels based on exponential distributions (Wilson and Adams 2013; Wilson 2014; Duvenaud et al. 2013b; Remes et al. 2017; Jang et al. 2017), notably the flexible Spectral Mixture (SM) kernels, have been successfully applied to many real-life extrapolation tasks. However, it remains challenging for SM kernels to perform long-term forecasting. Long-term forecasting involves predictive horizons that are far ahead of the observed training data.

Our experimental analysis indicates that characteristics of the signal important for long-term forecasting can be unravelled by analyzing the distribution of the Fourier coefficients of the signal, as shown in Fig. 1 for the (training part of the) Rail Passenger Miles time series dataset (fully described in Sect. 6.1). The distribution of the Fourier coefficients is non-smooth, heavy-tailed, sparse, and contains skewed peaks. Skewness of the peaks in the empirical density can be observed: the first, third, and fourth peak are right tailed. The heavy tails and skewness of peaks in the empirical spectral density correspond to long-range covariance in the time domain. We have fitted an SM kernel (red dashed line), Lévy kernel process (LKP) kernel (blue dashed), and SLSM kernel (black dashed) on this empirical spectral density, using the least squares method. SM and LKP cannot capture one-side heavy tail characteristics of the spectral density peaks, while SLSM can, yielding a better fit. Note that differences between the kernels are large, although in the figure they look negligible because an exponential scale is used.
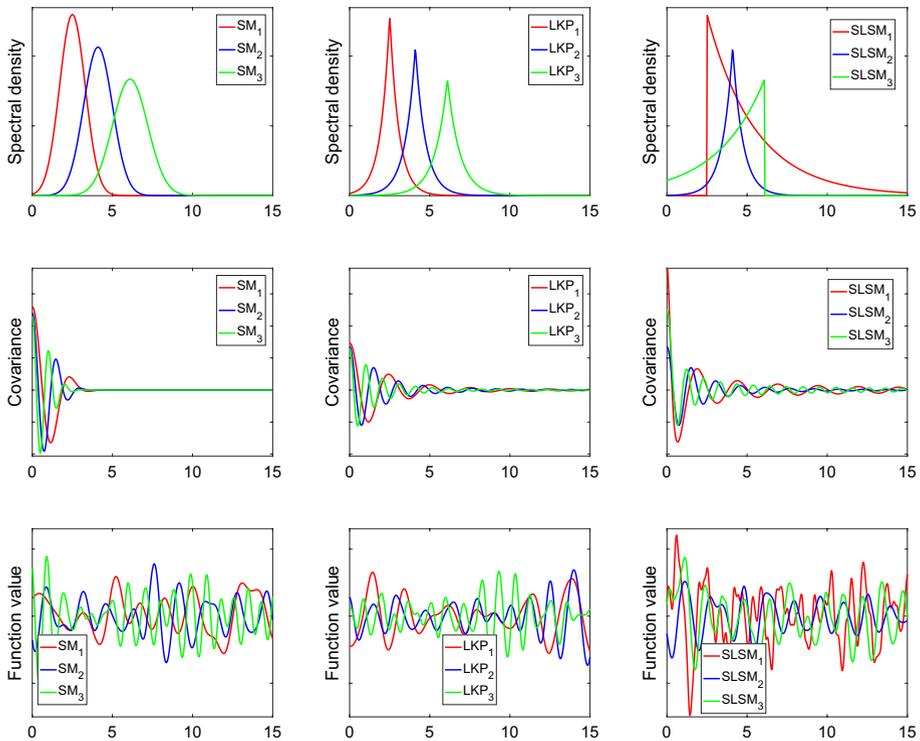
Fig. 2 SM (left), LKP (middle), and SLSM (right) kernels with three different choices of amplitude, mean and variance. Top row: different shapes of the spectral density. Middle row: covariance functions. Bottom row: samples drawn from Gaussian Processes with these kernels

Let us analyze the differences between these three spectral kernels in more detail. The inverse Fourier transform of an SM spectral density is an exponential function, so in the time domain, we have an exponential decay of covariance. When we use LKP to fit the empirical spectral density and apply the inverse Fourier transform, we get a Cauchy function, which has a much slower decay than that of the exponential function. Using SLSM we extend the tails of the Laplace distribution [by means of the γ term in Eq. (8)], which in the time domain yields a further reduction of the decay rate of the covariance. These observations are illustrated in Fig. 2.

In the top row, we can see the shapes of the spectral density for the three kernels, and the flexibility provided by the skewness parameter γ in the SLSM kernel. Covariance functions are shown in the middle row. They indicate that SM has a very short range covariance (range $\tau < 5$). LKP has longer range covariance (range $\tau < 15$) because of the heavier tails in the spectral density. SLSM has the longest covariance representation range ($\tau \gg 15$) due to skewness in the frequency domain.

In the bottom row, samples drawn from Gaussian Processes with these kernels are plotted. The SLSM kernel (red line) results in a larger variance, which is more flexible and less smooth than SM and LKP, because of heavy tails at high frequency position in the frequency domain. Both positive and negative skewness (determined by the value of γ) in SLSM can extend the covariance range and therefore increase the variance in the sampled

functions. A positive value of γ contributes more to the signal variance. The above observations justify the usefulness of an asymmetric Laplace kernel, capable to capture characteristics of the signal important for long-range forecasting.

Building on previous work on modeling the asymmetric tails of discrete Fourier transform coefficients distributions (Gazor and Zhang 2003; Eltoft et al. 2006; Minguillón and Pujol 2001), we investigate the use of the skewed Laplace mixture to model the spectral density, that is, the Fourier transform (FT) of a kernel. We then use the inverse Fourier transform to construct a kernel for GPs, which we call the Skewed Laplace Spectral Mixture (SLSM) kernel. Results of experiments indicate improved performance of GPs with the SLSM kernel and robustness to overfitting (when considering many mixture components).

Our contributions can be summarized as follows:

- we show how long-term forecasting with GPs is linked to long-range characteristics of covariance between random vectors, as determined by properties of the spectral density, namely sparsity, skewed peaks, long tails, and non-smoothness;
- we propose a new GP kernel for long-term forecasting, which includes a term to model skewness of each spectral peak;
- we compare SLSM with other spectral kernels, and show that SLSM is linked to the popular Rational Quadratic kernel;
- we introduce an algorithm for pruning kernel components and show experimentally its beneficial effect;
- we perform extensive long-term forecasting experiments with various time series data, including multivariate and large-scale ones.

The rest of the paper is organized as follows. In the next section we summarize related works. Sections 3.1 and 3.2 introduce GPs and the SM kernel. Then in Sect. 4 we introduce the SLSM kernel. Hyper-parameters initialization and automatic pruning of components are introduced in Sect. 5. Section 6 describes experiments on real-world datasets. Concluding remarks and future work are given in Sect. 7.

2 Related work

Because of the fundamental role of kernels in GPs, various works tackled the problem of kernel design, using two main approaches: the compositional approach, where kernels are constructed compositionally from a small set of simple base kernels, and the spectral representation approach, where kernels are constructed by modeling a spectral density, the Fourier transform of a kernel, with a Gaussian mixture, with theoretical support from the Bochner's theorem. In Duvenaud et al. (2013a) the kernel learning problem is formalised as structure discovery. A space of kernel structures is defined compositionally in terms of sums and products of a small number of base kernel structures. This provides an expressive modeling language that concisely captures many widely used techniques for constructing kernels. An algorithm that searches over this space using a greedy search is proposed. Recent works based on this approach include (Sun et al. 2018), which introduces a flexible family of kernels represented by a neural network, whose architecture is based on the composition rules for kernels, so that each unit of the network corresponds to a valid kernel; and Pearce et al. (2020), which derives Bayesian Neural Network architectures mirroring

compositional kernel combinations. The SLSM kernel with pruning strategy as proposed in this paper can be interpreted as a form of structure discovery, where unimportant components representing irrelevant structure are removed while important components representing the true underlying structure are kept during learning.

Spectral mixture kernels were introduced in Wilson and Adams (2013) and are described in detail in the next section. The success of these kernels has stimulated research on their extensions and improvements. For instance, the spectral mixture product kernel (Wilson et al. 2014) was introduced for modeling Cartesian structured datasets. Various works address the scalability of GPs with spectral kernels. In particular, in Oliva et al. (2016) a Bayesian non-parametric kernel-learning is proposed, which places a non-parametric prior on the spectral distribution of random frequencies allowing it to both learn kernels and scale to large datasets. The limitations of stationary kernels, notably their generalization to the non-stationary setting has been addressed in various works. For instance, the non-stationary SM kernel allows to model input-dependent GPs (Remes et al. 2017). In Shen et al. (2019) a family of non-stationary kernels, called harmonizable mixture kernels, is introduced, which supports stationary and many non-stationary covariances. A robust kernel learning framework for this kernel family is proposed. Spectral kernels with desirable properties, called generalized spectral kernels, have been proposed in Samo and Roberts (2015): they can approximate any (possibly non-stationary) bounded kernel, and allow inference of the degree of differentiability of the corresponding stochastic process when used as a covariance function. The SLSM kernel is stationary, a limitation discussed in the conclusion section. In Chen et al. (2019) a new kernel called generalized convolution spectral mixture is proposed, which uses cross convolution to model dependencies between components of a spectral kernel. The SLSM kernel does not explicitly model dependencies between components.

Our investigation differs from previous works mainly because of the focus on long-term forecasting. The above-mentioned spectral kernels consider a short tailed and smooth Gaussian as a base representation of the spectral density, which corresponds to an exponential function for the covariance in the time domain. Therefore these kernels are less suited for long-term forecasting.

In this respect, the most related work is the Lévy kernel process (LKP) (Jang et al. 2017). LKP was used for modeling the sharply peaked spectral density with a location-scale mixture of Laplace components, which is more sparse than the SM kernel, hence more robust to the choice of the number of mixture components. LKP can discover heavy-tailed patterns which are useful for long-range extrapolation. However it is only available for one dimensional inputs and does not model skewness characteristics of the spectral density investigated in the proposed SLSM, which, as demonstrated by our extensive experiments, is beneficial for long-term forecasting. Specifically, SLSM extends the tails of the Laplace distribution (by means of the γ term in Equation (8)), which in the time domain yields a slower decrease of the decay rate of the covariance. The value of γ gives increasing weight to the right ($\lim \gamma \to 1$) or to the left ($\lim \gamma \to -1$) tail of the density.

In general, the Laplace distribution has been successfully used in the past within diverse signal processing applications. Here we mention few examples. In Bhowmick et al. (2006) a Laplace mixture model is proposed, as a long-tailed alternative to the Normal distribution when identifying differentially expressed genes in micro-array experiments, and provide an extension to asymmetric over- or under-expression by means of an asymmetric extension of the mixture model. In Eltoft et al. (2006) the multivariate Laplace probability model in the context of a Normal variance mixture model is analyzed and an application of the model to represent the statistics of the discrete Fourier transform coefficients of a speech signal is given.

In Gazor and Zhang (2003) samples of a speech signal are considered as samples of a random variable, and the speech signal's probability density function in the time domain is investigated. It is shown that the distribution of speech samples is well described by a Laplace distribution. In Hyun (2004) a method for capturing nonlinear dependencies in images of natural scenes is introduced, which builds a hierarchical model based on independent component analysis and a mixture of Laplace distributions. These works mainly focus on the representation and modeling of spectral density using the Laplace distribution. Instead, our focus is on kernel design for GPs.

3 Background

3.1 Gaussian processes

A Gaussian process defines a distribution over functions, specified by its mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}')$ for given input vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^P$ (Rasmussen and Williams 2006). Thus we can define a GP as

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

Without loss of generality we assume the mean of a GP to be zero. The covariance function is used to construct a positive definite covariance matrix on the set X of training points, called the kernel and denoted by $K = k(X, X)$.

By placing a GP prior over functions through the choice of a kernel and parameter initialization, from the training data X we can predict the unknown function value $\tilde{y}_*$ and its variance $\mathbb{V}[\tilde{y}_*]$ (that is, its uncertainty) for a test point $\mathbf{x}_*$ using the following key predictive equations for GP regression (Rasmussen and Williams 2006):

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

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

where $\mathbf{k}_* = k(\mathbf{x}_*, X)$ is the covariance vector between $\mathbf{x}_*$ and X, and $\mathbf{y}$ are the observed function values corresponding to X. Typically, kernels contain free hyper-parameters, denoted by Θ, which can be optimized by minimizing the Negative Log Marginal Likelihood (NLML) of the observed values:

$$\begin{aligned} \text{NLML} &= -\log p(\mathbf{y}|X, \Theta) \\ &\propto \overbrace{\frac{1}{2}\mathbf{y}^\top (K + \sigma_n^2 I)^{-1} \mathbf{y}}^{\text{model fit}} + \overbrace{\frac{1}{2} \log |K + \sigma_n^2 I|}^{\text{complexity penalty}} \end{aligned} \tag{3}$$

where σ_n^2 is the noise level and NLML is obtained through marginalization over the latent function (Rasmussen and Williams 2006). This formulation follows directly from the fact that $\mathbf{y} \sim \mathcal{N}(0, K + \sigma_n^2 I)$.

In this paper we consider only stationary kernels, which are invariant to translation of the inputs. These can be described as functions of $\tau = \mathbf{x} - \mathbf{x}'$,

$$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x} - \mathbf{x}'). \tag{4}$$

3.2 Spectral mixture kernels

The class of flexible stationary kernels for GPs called Spectral Mixture (SM) kernels, was introduced in Wilson and Adams (2013) and Wilson (2014). An SM kernel, here denoted by k_{SM}, is derived through modeling its spectral density (the Fourier transform of a kernel) with a mixture of Gaussians. Such modeling is possible because of Bochner's Theorem (Bochner 2016; Stein 1999), which states that the properties of a stationary kernel entirely depend on its spectral density:

Theorem 1 (Bochner) *A complex-valued function k on $\mathbb{R}^P$ is the covariance function of a weakly stationary mean square continuous complex-valued random process on $\mathbb{R}^P$ if and only if it can be represented as $k(\tau) = \int_{\mathbb{R}^P} e^{2\pi J s^\top \tau} \psi(d\mathbf{s})$, where ψ is a positive finite measure, and J denotes the imaginary unit.*

If ψ has a density $\hat{k}(\mathbf{s})$, then $\hat{k}$ is called the spectral density or power spectrum of k, and k and $\hat{k}$ are Fourier duals, that is:

$$k(\tau) = \int \hat{k}(\mathbf{s}) e^{2\pi J s^\top \tau} d\mathbf{s}$$

and

$$\hat{k}(\mathbf{s}) = \int k(\tau) e^{2\pi J s^\top \tau} d\tau.$$

The SM kernel k_{SM} is defined as the inverse Fourier transform (that is, the Fourier dual) of a mixture of Q Gaussians in the frequency domain:

$$k_{SM}(\tau) = \mathcal{F}_{s\to\tau}^{-1}\left[\sum_{i=1}^{Q} w_i \hat{k}_{SM,i} \right](\tau), \tag{5}$$

where $\mathcal{F}_{s\to\tau}^{-1}$ denotes the inverse Fourier transform operator from the frequency to the time domain, and w_i is the weight of component $\hat{k}_{SM,i}$. Here the spectral density $\hat{k}_{SM,i}(\mathbf{s})$ of i-th component, is

$$\hat{k}_{SM,i}(\mathbf{s}) = [\varphi_{SM,i}(\mathbf{s}) + \varphi_{SM,i}(-\mathbf{s})]/2,$$

where $\varphi_{SM,i}(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_i, \Sigma_i)$ is a scale-location Gaussian with parameters $\boldsymbol{\mu}_i$, Σ_i. The symmetrization makes $\hat{k}(\mathbf{s})$ even, that is, $\hat{k}(\mathbf{s}) = \hat{k}(-\mathbf{s})$ for all $\mathbf{s}$. So the Fourier transform of $\hat{k}$, that is our kernel, is real, since the Fourier transform of a real even function is real. Furthermore k is symmetric, since the Fourier transform of an even function is even (cf. e.g. Hoffman 1997).

By expanding equation (5) we get

$$k_{SM}(\tau) = \sum_{i=1}^{Q} w_i k_{SM,i}(\tau) \tag{6a}$$

$$k_{\mathrm{SM},i}(\tau) = \cos\left(2\pi\tau^\top\boldsymbol{\mu}_i\right)\prod_{p=1}^{P}\exp\left(-2\pi^2\tau^2\,\Sigma_i^{(p)}\right) \tag{6b}$$

where $k_{\mathrm{SM},i}$ is the i-th component, P is the dimension of input, w_i, $\boldsymbol{\mu}_i = \left[\mu_i^{(1)}, \ldots, \mu_i^{(P)}\right]$, and $\Sigma_i = \mathrm{diag}\left(\left[(\sigma_i^2)^{(1)}, \ldots, (\sigma_i^2)^{(P)}\right]\right)$ are weight, mean, and variance of the i-th component in the frequency domain, respectively.

The inverse mean $1/\boldsymbol{\mu}_i$ of component i is the period, and the inverse standard deviation $1/\sqrt{\Sigma_i}$ the length scale, determining how quickly a component varies with the inputs. So the variance $(\sigma_i^2)^{(P)}$ can be thought of as an inverse length-scale, and $\mu_i^{(P)}$ as a frequency. The weight w_i specifies the relative contribution of the i-th mixture component.

In addition, by modeling the spectral density as a mixture of Q univariate Laplacians, the kernel form of LKP (Jang et al. 2017) can written as follows:

$$k_{\mathrm{LKP}}(\tau) = \sum_{i=1}^{Q} w_i k_{\mathrm{LKP},i}(\tau) \tag{7a}$$

$$k_{\mathrm{LKP},i}(\tau) = \cos\left(2\pi\tau^\top\mu_i\right)\left(1 + \frac{1}{2}\sigma_i^2\tau^2\right)^{-1} \tag{7b}$$

4 Skewed Laplace spectral mixture kernel

Any stationary covariance kernel can be approximated to arbitrary precision by an SM kernel, given enough components of mixture in the spectral representation, because mixture of Gaussians are dense in the set of all distribution functions (Kostantinos 2000). So it would seem useless to introduce a new stationary kernel. However, using a large number of components can lead to overfitting when there is not enough training data. Therefore, it makes sense to introduce and study new stationary kernels, that are robust to overfitting in the presence of a large number of components. An example of such a kernel is LKP. However, as mentioned in the introduction, LKP does not model skewness characteristics of the spectral density beneficial for long-range forecasting.

Here we propose to overcome the limitations of the SM and LKP kernels by using a mixture of Skewed Laplace (SL) distributions, which can better capture skewness of peaks of the spectral density and its heavy tail characteristic. We consider the three-parameter (μ, γ, σ) family of skewed Laplace distributions introduced in Kotz et al. (2001), with density

$$\varphi(s;\mu,\gamma,\sigma) = \frac{\sqrt{2}}{\sigma}\frac{\kappa}{1+\kappa^2}\begin{cases}\exp\left(-\frac{\sqrt{2}}{\sigma\kappa}(\mu-s)\right) & \text{if } s < \mu \\ \exp\left(-\frac{\sqrt{2}\kappa}{\sigma}(s-\mu)\right) & \text{if } s \geq \mu\end{cases} \tag{8}$$

where $\kappa = \sqrt{2}\sigma/(\gamma + \sqrt{2\sigma^2+\gamma^2})$, γ is the skewness parameter, μ, and σ^2 are the mean and variance of the distribution, respectively. When $\gamma = 0$ this distribution reduces to the standard non-skewed Laplace distribution. The inverse Fourier transform of the i-th skewed Laplace density is (Fragiadakis and Meintanis 2011)

$$\mathcal{F}^{-1}_{s\to\tau}[\varphi_i(s;\mu_i,\gamma_i,\sigma_i)](\tau) = \frac{C_i(\tau) + J\gamma_i\tau}{C_i(\tau)^2 + \gamma_i^2\tau^2}e^{J\mu_i\tau}, \tag{9}$$

where $C_i(\tau) = 1 + \frac{1}{2}\sigma_i^2\tau^2$.

We proceed to construct a kernel from the skewed Laplace density in the same way as described in the previous section: (1) symmetrize $\varphi_i(s;\mu_i,\gamma_i,\sigma_i)$; (2) take the inverse Fourier transform of the resulting even real function; and (3) make a mixture of several such components.

First, by symmetrization we obtain the spectral density of the i-th component of the SL kernel:

$$\hat{k}_{\text{SLSM},i}(s) = \frac{1}{2}\left(\varphi_i(s;\mu_i,\gamma_i,\sigma_i) + \varphi_i(-s;\mu_i,\gamma_i,\sigma_i)\right) \tag{10}$$

Note that the above transformation does not destroy the skewness of the mixture components (the peaks) of the spectral density: although by construction the spectral density is symmetric around zero, each of its peaks is skewed around its mean (with skew parameter γ_i).

Next, by applying (9) to $\varphi_i(s;\mu_i,\gamma_i,\sigma_i)$ and $\varphi_i(-s;\mu_i,\gamma_i,\sigma_i)$, and adding the resulting functions we obtain the time domain representation of the i-th component $k_{\text{SLSM},i}$ of the SLSM kernel:

$$\begin{aligned}
k_{\text{SLSM},i}(\tau) &= \frac{1}{2}\left(\frac{C_i(\tau) + J\gamma_i\tau}{C_i^2(\tau) + \gamma_i^2\tau^2}\exp\left(J\mu_i\tau\right) + \frac{C_i(\tau) - J\gamma_i\tau}{C_i^2(\tau) + \gamma_i^2\tau^2}\exp\left(-J\mu_i\tau\right)\right) \\
&= \frac{C_i(\tau)\cos(\mu_i\tau) - \gamma_i\tau\sin(\mu_i\tau)}{C_i^2(\tau) + \gamma_i^2\tau^2}.
\end{aligned} \tag{11}$$

Each component $k_{\text{SLSM},i}(\tau)$ is positive semi-definite because its corresponding spectral density defined in (10) is non-negative everywhere (Bochner 2016; Stein 1999). Also, each component is real because its spectral density is even (Rasmussen and Nickisch 2010).

The denominator of $k_{\text{SLSM},i}(\tau)$ is a skew function of the inverse distance and the numerator term is a periodical trigonometry function. $C_i(\tau)$ is an inverse Cauchy function (the inverse Fourier transform of the zero positioned non-skewed Laplace) (Jang et al. 2017). The Cauchy function has slower decay over the distance τ than the exponential function used in SM based kernels.

Finally, the skewed Laplace spectral mixture (SLSM) kernel k_{SLSM} corresponds to a mixture of Q skewed Laplace components, and is defined as

$$\begin{aligned}
k_{\text{SLSM}}(\tau) &= \mathcal{F}^{-1}_{s\to\tau}\left[\sum_{i=1}^{Q} w_i\hat{k}_{\text{SLSM},i}\right](\tau) \\
&= \sum_{i=1}^{Q} w_i\frac{C_i(\tau)\cos\left(\mu_i\tau\right) - \gamma_i\tau\sin\left(\mu_i\tau\right)}{C_i^2(\tau) + \gamma_i^2\tau^2}.
\end{aligned} \tag{12}$$

The Skewed Laplace distribution has a natural extension to higher dimensions, described, e.g., in Kotz et al. (2001), Kozubowski and Podgórski (2000) and Visk (2009). Therefore our kernel can be directly extended to the multivariate setting as follows:

$$k_{\text{SLSM}}(\boldsymbol{\tau}) = \sum_{i=1}^{Q} w_i \frac{C_i(\boldsymbol{\tau}) \cos\left(\boldsymbol{\tau}^\top \boldsymbol{\mu}_i\right) - (\boldsymbol{\tau}^\top \boldsymbol{\gamma}_i) \sin\left(\boldsymbol{\tau}^\top \boldsymbol{\mu}_i\right)}{C_i^2(\boldsymbol{\tau}) + (\boldsymbol{\tau}^\top \boldsymbol{\gamma}_i)^2}, \tag{13}$$

where $C_i(\boldsymbol{\tau}) = 1 + \frac{1}{2}\boldsymbol{\tau}^\top \Sigma_i \boldsymbol{\tau}$. Here Σ_i denotes the covariance matrix of component i, and $\boldsymbol{\mu}_i$ is the vector of means of this component.

4.1 Comparison with other kernels

Figure 2 provides a visual comparison of the SM, LKP and SLSM distributions used for fitting the empirical spectral density. The differences among SLSM, LKP and SM are illustrated in terms of covariance, spectral density, and sampling functions. The subplots in the second row of Fig. 2 show random functions drawn from a GP with SM, LKP, and SLSM kernel, respectively. The sampled function values were obtained using 500 equally-spaced discrete points.

The inverse FT of the SM spectral density is an exponential function, so in the time domain we have an exponential decay of covariance. When we use LKP to fit the empirical spectral density and apply the inverse FT we get a Cauchy function, which decays in a much slower way than the exponential one. Using SLSM we extend the tails of the Laplace distribution [by means of the γ term in Eq. (8)], which in the time domain yields a further reduction of the decay rate of the covariance. This phenomenon is illustrated in Fig. 2, which shows longer range covariance of SLSM (middle of right plots). The main characteristics of SM, LKP and SLSM can be summarized as follows:

- SM: multivariate Gaussian, dense, non-skewed peaks, short-tailed, smooth, symmetry of each peak, exponential decay of covariance;
- LKP: univariate Laplacian, sparse, non-skewed peaks, heavy-tailed, non-smooth, symmetry of each peak, fast decay of covariance;
- SLSM: multivariate skewed Laplacians, skewed peaks, one side more heavily-tailed, non-smooth, slower decay of covariance.

Figure 1 shows an example where SLSM provides a better fit of the empirical spectral density than SM and LKP. SLSM extends the LKP kernel (apart from the way inference is performed), which can be obtained by removing the skewness term, i.e., by setting the γ_i's to 0. More precisely, LKP as defined in Jang et al. (2017), uses Reversed-Jump MCMC (RJ-MCMC) to perform inference. While, SLSM uses LBFGS implemented in GPML toolbox to perform inference.

Also, SLSM is related to the rational quadratic (RQ) kernel,

$$k_{\text{RQ}}(\tau) = \theta_f \left(1 + \frac{\tau^2}{2\alpha\ell^2}\right)^{-\alpha} \tag{14}$$

The RQ kernel can be interpreted as a scale mixture (an infinite sum) of squared scale mixture SE kernels with different characteristic length-scales (Rasmussen and Williams 2006). The RQ kernel is considered the most general representation defining a valid isotropic covariance function in $\mathbb{R}^P$ (Wilson 2014; Stein 1999). From Eqs. (11) and (14) it follows that a non-skewed SLSM component $k_{\text{SLSM},i}(\tau)$ is the product of a RQ kernel with $\alpha = 1$, $\theta_f = w_i$, $\ell^{-2} = \sigma_i^2$ and a cosine kernel. Thus the RQ kernel with $\alpha = 1$ can be viewed as modeling the spectral density, as one Laplace distribution at zero mean position in the

frequency domain. That is, the RQ kernel can be viewed as part of a component of a SLSM kernel.

SLSM component can be approximated at arbitrary precision with a mixture of Gaussians using sufficiently many components. Thus a SM kernel can approximate the SLSM kernel. However, as observed e.g. by Samo and Roberts (2015), although mixture of Gaussian distributions can be used to approximate any stationary distribution, a large number of SM components might be required to account for the lower degrees of smoothness in the data. This would affect inference, which would become more expensive and be more vulnerable to the presence of local optima. In Fig. 2 SM, LKP, and SLSM with 3 components are illustrated in terms of spectral densities, covariance, and sampling paths.

5 SLSM's hyper-parameter initialization, automatic pruning of components, and scalable inference

Kernels hyper-parameters are inferred by optimizing the negative log marginal likelihood (NLML), which for kernels such as SM and SLSM, amounts to solve a non-convex optimization problem. The initialization of these hyper-parameters has a direct impact on the ability of the optimization process to find a good local optimum of the NLML. Below, we describe the initialization procedure used in our experiments. Another parameter of spectral kernels to be set is the number of kernel components. A way to circumvent this problem is to automatically prune irrelevant components. In LKP, automatic pruning of components is part of the method, because of the sparsity-inducing property of the Lévy priors [see Jang et al. (2017), supplementary material]. For SLSM and SM kernels, we propose a method for automatic pruning kernel components. In order to perform experiments with large datasets, scalable inference is needed, because in GPs exact inference is computationally expensive. Therefore, for the large-scale experiment in Sect. 6.5 we use the scalable inference procedure described in this section.

5.1 Hyper-parameter initialization

In general, SM-based kernels are particularly sensitive to the initialization of their hyper-parameters, and the SLSM kernel shares this initialization problem. Here we apply an initialization strategy that has been shown to be effective in previous works (Wilson and Adams 2013; Herlands et al. 2016). These works suggest that non-smooth peaks of the empirical spectral densities are near the true frequencies and use a Gaussian Mixture Model to fit the empirical spectral density, which is then used to initialize the hyper-parameters of an SM kernel. We make use of this result and initialize the hyper-parameters of the SLSM kernel using a Laplace Mixture Model,

$$p_{\text{LMM}}(\Theta|\mathbf{s}) = \sum_{i=1}^{Q} \tilde{w}_i \phi_i(\tilde{\mu}_i, \tilde{\Sigma}_i) \tag{15}$$

to fit the empirical spectral density in order to get Q cluster centers, where Q is the number of components. We use the Expectation Maximization algorithm (Moon 1997) to estimate the parameters of this mixture model. The resulting estimates $\tilde{w}_i$, $\tilde{\mu}_i$, and $\tilde{\Sigma}_i$ are used as initial hyper-parameters values of the SLSM kernel. The skewness parameter γ_i is initialized

randomly between − 1 and 1. We use the above initialization strategies in all our experiments on real-world datasets.

5.2 Automatic pruning of kernel components

In order to automatically select the number of components, we adopt a method originally introduced for pruning the weights of a neural network, based on the lottery ticket hypothesis (LTH): dense, randomly-initialized, feed-forward networks contain subnetworks (winning tickets) that—when trained in isolation—reach test accuracy comparable to the original network in a similar number of iterations (Frankle and Carbin 2019). The above mentioned method trains a neural network and prunes a percentage of small weights; the process is repeated in a recursive fashion for each subsequent network. At each iteration, each unpruned weight value is reset to its initial value before training.

In our context, we want to prune Laplace kernel components as part of training the model. This means that the components that we prune are the Laplace kernel components that make up Eq. (12). Note that kernel components should not be confused with the sines and cosines terms of the Fourier decomposition of the empirical spectral density, which are often also referred to as components. Each component of the kernel covers many frequencies in the spectral density.

Specifically, in our case components with magnitude of the weight w_i smaller than 1 are pruned. The choice of this threshold is heuristically motivated: since the energy of spectral density is very big, if a weight is smaller than 1, then the contribution to its component is very small and can be ignored. Figure 6 illustrate this phenomenon, and shows the weights of components of kernels of a GP trained on the sunspot data (described in Sect. 6.3) using $Q = 100$ components. Weights of many components (almost 65%) of SLSM are smaller than 1, while for SM and LKP there are less than 30% components with weights smaller than 1.

In our experiments we use the following LTH algorithm to prune components of SM and SLSM kernels for GP:

1. Initialize a GP with Q components.
2. Train the GP for 100 iterations using the LBFGS algorithm (see description below).
3. Prune component with weight smaller than 1.
4. Set weights of unpruned components to their initial value and re-train the GP.

We train the GP model by means of the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) algorithm with 100 iterations. The computational cost of this pruning procedure depends on the number of iterations performed. In practice, two pruning rounds are sufficient.

Our LTH pruning algorithm is applicable to SM and SLSM. We also consider the specific variant of SLSM obtained by setting $\gamma = 0$ (we call this variant SLSM ($\gamma = 0$)). This variant is strongly related with the LKP kernel (see Sect. 6). As already mentioned, in the LKP method, the Lévy process introduces a sparsity-inducing prior over mixture components, allowing automatic pruning of components.

5.3 Scalable inference

For GPs, exact inference is prohibitively slow for more than a few thousand of points because of its $\mathcal{O}(n^3)$ computational complexity and $\mathcal{O}(n^2)$ memory complexity (Rasmussen and Williams 2006; Rasmussen and Nickisch 2010; Quiñonero-Candela and Rasmussen 2005; Chalupka et al. 2013) in the Cholesky decomposition of covariance matrix. The most computationally expensive steps are inverting the covariance matrix and computing the determinant of the covariance (i.e. $(K + \sigma_n^2 I)$), see Eqs. (2b) and (3). These problems have been addressed by covariance matrix approximation or inference approximation.

In our experiment with big time series of length $n \geq 10{,}000$, we adopt the robust Bayesian Committee Machine (rBCM) of distributed GPs (Deisenroth and Ng 2015) framework to approximate the covariance matrix structures for scalable inference. rBCM is a scalable product-of-experts model for large-scale GP regression, which recursively distributes computations to M independent local computational units and, recombine them to form an overall result. Due to the independence between local GP models, the marginal likelihood [see Eq. (3)] of a full GP using rBCM can be factorized into a product form of M marginal likelihoods of local GP models, thus performing scalable inference of the full GP model.

Using rBCM, we can train a GP model with SLSM kernel as follows:

$$p(\mathbf{y}|X, \Theta) \approx \prod_{i=1}^{M} p^{(i)}(\mathbf{y}^{(i)}|X^{(i)}, \Theta) \tag{16}$$

where M is the number of local GPs and $p^{(i)}(\mathbf{y}^{(i)}|X^{(i)}, \Theta)$ is the marginal likelihood of the i-th local GP using the i-th subset $\{X^{(i)}, \mathbf{y}^{(i)}\}$ of dataset $\{X, \mathbf{y}\}$. The predictive probability is computed as the product of the predictive probabilities of the independent local GPs, as follows:

$$p(f_*|\mathbf{x}_*, \mathbf{y}, X) \approx \prod_{i=1}^{M} p^{\beta_i}(f_*|\mathbf{x}_*, \mathbf{y}^{(i)}, X^{(i)}). \tag{17}$$

where β_i is the weight of i-th local GP.

6 Experiments

We assess the performance of SLSM and of the following popular stationary kernels available in the GPML toolbox (Rasmussen and Williams 2006; Rasmussen and Nickisch 2010): linear with bias (LIN), squared exponential (SE), piecewise polynomial (Poly), periodic (PER), rational quadratic (RQ), Matérn 5/2 (MA), Gabor, neural network (NN), additive kernel based on SE using unary and pairwise interactions (ADD), as well as SM and LKP kernels. Also, we assess the performance of the LKP kernel as an instance of SLSM with $\gamma = 0$, we denote this variant by SLSM ($\gamma = 0$). LKP and SLSM ($\gamma = 0$) have the same kernel structure, but inference is performed using different optimization methods: LKP uses reversible jump Markov chain Monte Carlo (RJ-MCMC), while SLSM ($\gamma = 0$) uses the standard LBFGS method based on gradient-based optimization.

Also, we consider popular traditional forecasting methods, namely autoregressive integrated moving average (ARIMA), automatic ARIMA (Auto-ARIMA) and the error, trend, seasonality (ETS) model with Holt-Winters seasonal method. These methods are different from GP models and have been used for forecasting time series with trend and seasonal

components. We use the implementation of the ARIMA, Auto-ARIMA, and ETS in the ECOTOOL toolbox, with default parameters values (Pedregal 2019).

For SM, LKP, and SLSM, we consider $Q = 10$ components, which are then automatically pruned. The choice of $Q = 10$ is heuristic, as e.g. in Wilson and Adams (2013). The underlying assumption is that each component corresponds to a distribution in the frequency domain, which covers a much wider range of frequencies than a sine or cosine located at a single frequency position. Therefore a mixture of 10 components located at different positions can cover the structure of the underlying spectral density reasonably well, that is, the underlying peaks (determined by the weight), frequency range (determined by the variance), skewness, peak positions (determined by the mean) of spectral density in the learned model other than empirical spectral density).

For SM we initialize the hyper-parameters using a Gaussian Mixture Model $p_{\text{GMM}}(\Theta|\mathbf{s}) = \sum_{i=1}^{Q} \tilde{w}_i \mathcal{N}(\tilde{\mu}_i, \tilde{\Sigma}_i)$. For LKP we use a non-skewed Laplace Mixture Model. All other kernels in our comparison use the standard initialization in GPML toolbox.

The LKP method involves a Lévy process prior and RJ-MCMC to prune extraneous components and to infer the posterior distribution, respectively, as described in Jang et al. (2017). For all other kernels, we use exact inference together with LBFGS for optimizing the kernel hyper-parameters.

For SM and SLSM, we also report results obtained by pruning extraneous components using the LTH algorithm described in Sect. 5.2. Also, plots of the spectral density, specific skewness values, instance of skewed component, and experiments with varying values of Q in SLSM, LKP, SM are considered to provide a deeper insight into their performance.

In all experiments we use the following evaluation metrics: mean squared error (MSE) $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$, mean absolute error (MAE) $\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \tilde{y}_i|$, where $\tilde{y}_i$ denotes the predicted value, and NLML (computed over the training data). Results are average over 10 independent runs. Standard deviations are also reported.

We consider many forecasting tasks with real-world datasets: rail miles, monthly electricity, mean sunspot, computer hardware, the monthly M3-competition dataset with more than 1000 time series, and the large pole telecom dataset. Source code of the proposed SLSM kernel is available at https://github.com/ck2019ML/GPs-for-long-range-forecatsing.

6.1 Long-term forecasting of rail passenger miles: comparing spectral densities of SM, LKP and SLSM

In this experiment, we illustrate on a real-life long-term forecasting problem the differences between SM, LKP and SLSM kernels in terms of their spectral densities. We consider the rail miles dataset.[1] This dataset consists of 220 values covering the period from Jan. 2000 to Apr. 2018. We use $Q = 10$ components for the SM, LKP and SMLS kernels The top plots in Fig. 3 show extrapolation results. For test points far from the last (training) observation, the SLSM kernel achieves best extrapolation performance. The bottom part of Fig. 3 shows that the spectral density of the SLSM kernel has components with a higher variance than that of the other kernels as well as components with higher mean.

In Fig. 3 we can observe that better fit on empirical spectral density of SM and LKP lead to limited generalization in terms of extrapolation. This indicates that the fitting on empirical spectral density is a poor proxy for the generalization error of testing, since the

---

[1] https://fred.stlouisfed.org/series/RAILPM.
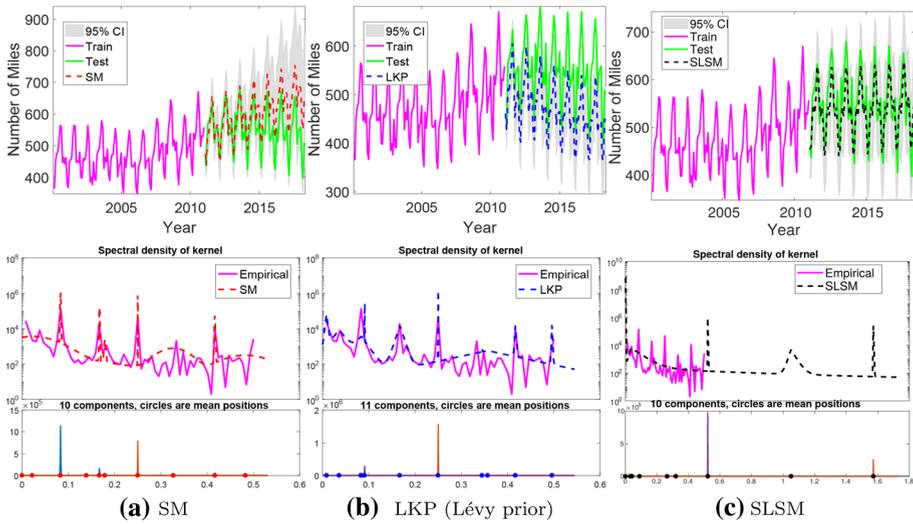
Fig. 3 Extrapolation results on the Rail Passenger Miles dataset. Bottom plots: spectral densities of optimized kernels, together with the empirical spectral density of the training data. Note that the better fitting of SM and LKP on the empirical spectral densities (training data) may lead to overfitting and negatively affect extrapolation performance
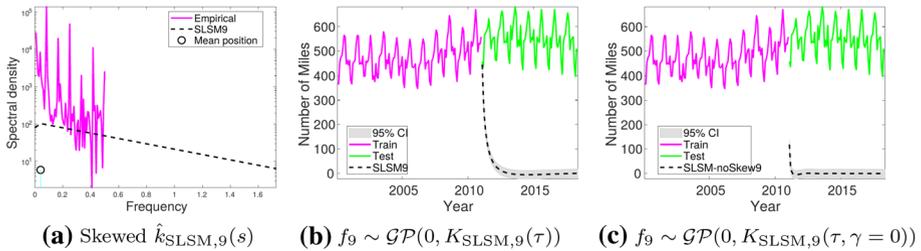
Fig. 4 An instance of skewed peak ($\mu_9 = 0.04, \gamma_9 = 0.45$) learnt by SLSM, its posterior extrapolation, and non-skewed ($\gamma = 0$) posterior extrapolation. With skewness, $f_9 \sim \mathcal{GP}(0, K_{\text{SLSM},9}(\tau))$ can keep trend for a longer time than $f_9 \sim \mathcal{GP}(0, K_{\text{SLSM},9}(\tau, \gamma = 0))$

model may be overfitted, leading to low training error but poor generalization (testing) performance. In contrast, the optimized SLSM kernel looks different from the empirical spectral density. We believe that this is the result of the extra flexibility that skewness provides, which allows the optimizer to find a better local optimum for the kernel hyper-parameters. Which in turn leads to better generalization, since it can capture patterns not present in the empirical spectral density.

Looking in more detail, we can see that the optimized SLSM kernel has two components with large mean, $\mu_5 = 1.57$ and $\mu_5 = 1.05$, which may respectively correspond to smaller periods of about half month and one month, in line with half and one monthly variations that appear in the dataset. In SM and LKP, their spectral densities miss many of variations at higher frequencies, meaning they cannot learn structures located at higher frequencies. In Fig. 4 we show spectral density, extrapolation, and non-skewed extrapolation of the 9th skewed SLSM component in SLSM. Subplots (b) and (c) in Fig. 4 show that f_9 reduces

to zero fast if we just remove its skewness and keep the rest hyper-parameters unchanged, demonstrating that skewness extends covariance range in GP model, which can correlate test points far away from training points and thus is beneficial for long range extrapolation.

The specific values of weights w, means μ, variances σ^2, and skewness γ in SLSM determine the contribution, periodicity, length-scale, and long-range dependency of each SLSM component, respectively.

Overall, results indicate that GP with SLSM kernel achieves best performance in terms of MSE (see Table 1) and MAE (see Table 2), while its NLML is bigger than that of SM (see Table 3). This may seem surprising. However, note that the computation of NLML is based on training data [see Eq. (3)]. Therefore a lower NLML value usually indicates a better fitting of the training data, but does not guarantee better generalization. Also, 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.

6.2 Long-term forecasting of electricity consumption: varying the number of kernel components

In this experiment we investigate comparatively the performance of SM, LKP and SLSM when varying the number of components in a small range (2–10), using real-life data for long-term monthly electricity forecasting. Electricity forecasting is a problem of practical relevance for electric power system planning, tariff regulation and energy trading (Pessanha and Leon 2015). The dataset includes monthly residential electricity usage in Iowa city from Jan. 1971 to Oct. 1979. There are a total 106 time points, of which we use the first 60% for training and the rest for testing.

For SLSM, LKP, and SM we use $Q = 10$ components. Fig. 5, subplots (a), (b), and (c), show that SLSM has a better extrapolation performance than SM and LKP, in particular for small peaks corresponding to the winter months. Subplot (d) shows the importance of skewness in SLSM, with a component having a relatively very high skewness value.

On this data, LKP selects automatically 11 components. Fig. 5 [subplots (e) and (f)] shows the performance of SLSM, SLSM ($\gamma = 0$), and SM with varying number of components. Using the same value of Q for all kernels seems unfair, because SLSM has $4Q$ parameters while SM and LKP have $3Q$ parameters. Comparing performance at same value of the number of parameters would mean to increase the value of Q for SM and LKP. However, using a bigger Q does not necessarily yield better performance. For instance, results indicate that SLSM with $Q = 6$ is still better than SM and LKP with $Q = 10$. Results in terms of MSE, MAE, and NLML are given in Tables 1, 2 and 3, respectively.

6.3 Long-term prediction of mean sunspot numbers: large number of kernel components

Here we further assess long-term forecasting capability of the considered kernels, this time with a large number of components. We consider the mean total sunspot number dataset. For this dataset the monthly number of sunspots was collected from Jan. 1842 - Dec. 1933 at the Royal Observatory of Belgium, Brussels, resulting in 1104 time points. We consider
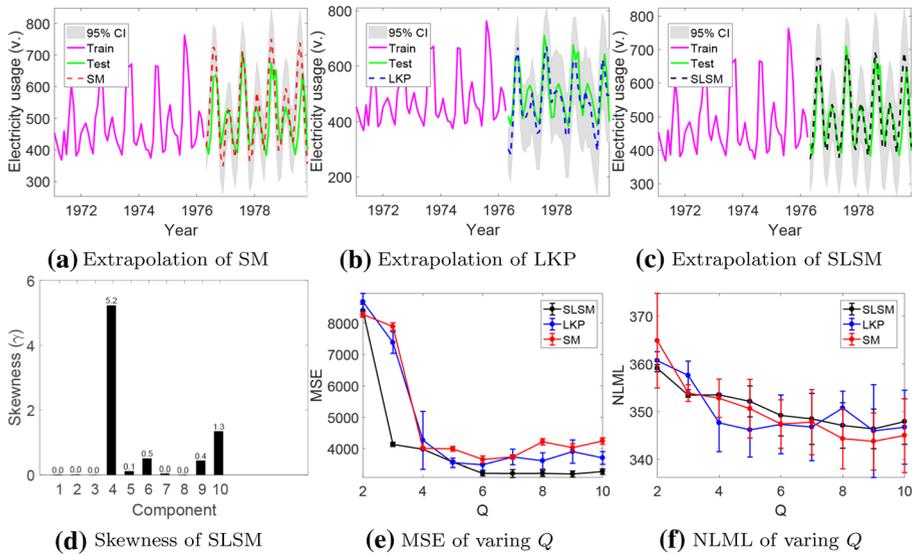
Fig. 5 Extrapolation using the SM kernel (**a**), LKP kernel using Lévy prior (**b**), and SLSM kernel (**c**) on the monthly electricity usage dataset. The solid magenta line shows the training data, while the dashed line is the mean prediction by a GP. Plots (**d**–**f**) show the skewness of the optimized SLSM kernels, together with MSE and NLML of varying Q. All plots for the other experiments use the same notation

the 13-month smoothed data. [2] We consider $Q = 100$ for the SM, LKP and SLSM kernels. In order to increase the difficulty of extrapolation and perform longer-term forecasting, we use less training samples (45%) than test samples (55%).

Results are shown in Fig. 6, in particular the hyper-parameters histogram of SM, LKP, and SLSM kernels, which indicates that in addition to long-range covariance, skewness is also beneficial for robust learning in the presence of a large number of components. However a large value of Q affects the optimization time due to the larger of hyper-parameter space. Table 1 and Table 2 shows improved MSE and MAE results of SLSM over the other kernels, while in this experiment NLML of SLSM is the smallest (see Table 3).

6.4 Extrapolating CPU relative performance: multidimensional computer hardware data

We now investigate the performance of the considered spectral kernels on an extrapolation task with multi-variate data. We consider a multidimensional computer hardware dataset. The dataset recorded relative CPU performance collected on Oct. 1987. There are 209 multivariate samples and the following 8 features: vendor name, model name, machine cycle time, minimum main memory, maximum main memory, cache memory, minimum channels, maximum channels. The task is to predict CPU relative performance. In this experiment we use $Q = 10$ components for SLSM, SLSM ($\gamma = 0$), and SM. We use the first 60% of the data for training and the rest for testing. Before training the GPs, we normalize the

---

[2] http://www.sidc.be/silso/datafiles.
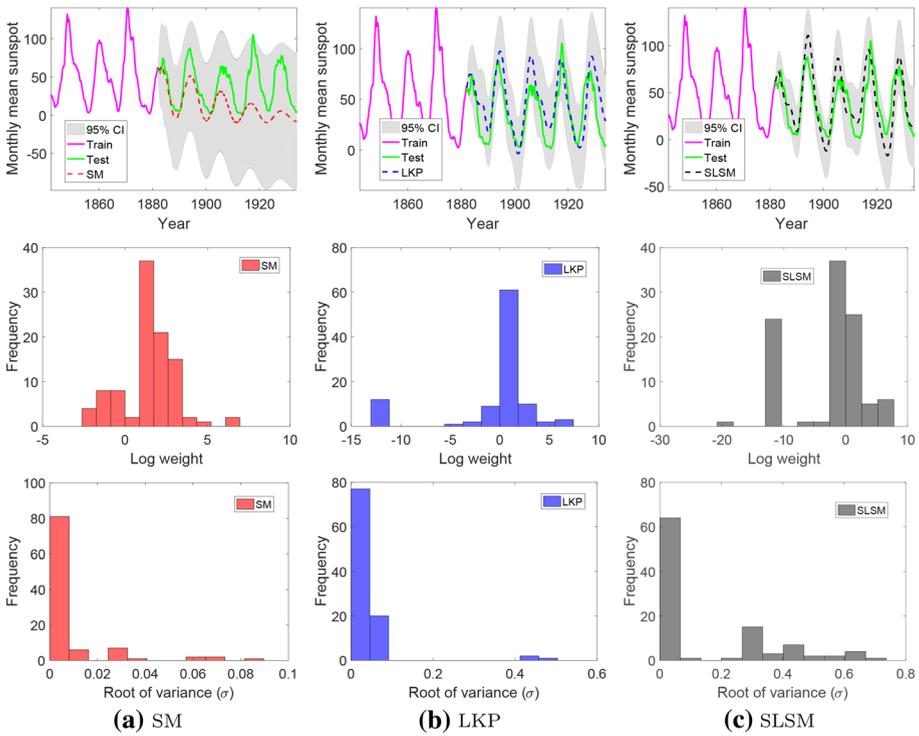
Fig. 6 Extrapolation and histogram of hyper-parameters of SM, LKP, and SLSM with 100 components performed on mean sunspot dataset. First row: extrapolation of SM, LKP, SLSM. Second row: histogram of log weight log(*w*); Third row: histogram of variance σ; For histogram of weights in LKP and SLSM, most of the components have very small weights close to zero, which means the LKP and SLSM are very sparse. While for SM a large number of components have weights bigger than 10, which means SM is very dense and these SM components usually have bigger magnitude in the frequency domain. On the other hand, the histogram of variance σ shows that SLSM components are much more heavy tailed with bigger variances than those of LKP and SM. SM components in the frequency domain are very short tailed, which exhibit fast decaying behaviors

training data so that the data is centered around the mean and has unit standard deviation. The mean and standard deviation of the training data are used to normalized test data.

Note that for SM, LKP, and SLSM on this unevenly discrete sampled multi-dimensional dataset, the spectral density structures are unclear and their hyper-parameter spaces are much larger than for time series. Input locations in this multi-dimensional dataset are not like those of time series: they are unevenly discretely distributed. Thus we cannot get the multidimensional Fourier transform of the data. So we do not know their spectral densities structure. The size of their hyper-parameter space is equal to their input dimension times the length of the time series. Therefore, on this small multi-dimensional dataset we did not apply our procedure for pruning components. Also, we could not use informed hyper-parameter initialization for SM, LKP, and SLSM, hence we just randomly initialize all kernels.

LKP with Lévy prior and RJ-MCMC inference was developed and tested on one-dimensional data. So, in order to compare LKP with SLSM we only consider SLSM with $\gamma = 0$, with standard exact inference instead of RJ-MCMC inference used in LKP.

Table 1 Performance of SLSM and other kernels in terms of average MSE (over 10 independent runs) $\pm$ standard deviation, and performance of traditional forecasting methods

Kernel	Rail miles	Electricity	Sunspot	Computer	Pole-telecom
LIN	6910 ± 871	9710 ± 616	4413 ± 1296	1850 ± 49	2397 ± 24
SE	$206{,}248 \pm 98{,}312$	$251{,}272 \pm 3600$	2125 ± 71	$36{,}156 \pm 4740$	58 ± 14
Poly	$304{,}892 \pm 4891$	$260{,}211 \pm 1832$	2179 ± 140	$41{,}472 \pm 2530$	108 ± 28
PER	9107 ± 581	$10{,}788 \pm 5386$	2266 ± 1153	$25{,}266 \pm 816$	1614 ± 1120
RQ	8433 ± 2542	$11{,}495 \pm 1688$	2175 ± 115	5402 ± 1174	51 ± 7
MA	$259{,}732 \pm 20{,}192$	$238{,}185 \pm 7417$	2258 ± 221	$34{,}239 \pm 5513$	49 ± 7
Gabor	$304{,}877 \pm 4884$	$260{,}535 \pm 2019$	2236 ± 148	$42{,}642 \pm 224$	2612 ± 0.9
NN	5159 ± 333	9652 ± 1783	1234 ± 108	$12{,}239 \pm 1147$	230 ± 4
ADD	$282{,}873 \pm 15{,}384$	$250{,}704 \pm 3325$	2265 ± 178	676 ± 85	57 ± 7
ARIMA	7622	163,950	25,031	–	–
Auto-ARIMA	290,116	216,497	1662	–	–
ETS	3753	22,721	514,272	–	–
SM	4527 ± 571	4049 ± 400	795 ± 31	559 ± 52	42 ± 6
LKP	5023 ± 581	4257 ± 424	613 ± 74	–	–
$\text{SLSM}_{\gamma=0}$	4225 ± 569	3476 ± 197	385 ± 32	538 ± 42	40 ± 1
SLSM	1466 ± 220	3167 ± 118	314 ± 22	$\mathbf{298} \pm 26$	$\mathbf{29} \pm 1$
SM (LTH)	3214 ± 583	3686 ± 214	772 ± 31	–	–
$\text{SLSM}_{\gamma=0}$ (LTH)	2552 ± 698	3253 ± 196	321 ± 19	–	–
SLSM (LTH)	$\mathbf{1205} \pm 146$	$\mathbf{3047} \pm 104$	$\mathbf{228} \pm 37$	–	–

Bold denotes the best performance result for each experiment

The other baseline kernels use automatic relevance determination (ARD) to scale each dimension independently. Results of this experiment, in terms of MSE, MAE and NLML are given in Tables 1, 2, and 3, respectively.

6.5 Long-term telecommunication forecasting: a large dataset

In order to investigate scalability of SLSM, we consider a large dataset: the pole telecom dataset (Nguyen and Bonilla 2014). The pole telecomm dataset describing a telecommunication problem has 26 non-constant input dimensions, 10,000/5000 samples for train/test, respectively. By using rBCM, we can reduce the computational complexity of GP inference to $\mathcal{O}(mn_m^2)$, where m and n_m denote the number of BCM's and the size of data used in each BCM, respectively.

For SLSM, SLSM ($\gamma = 0$), and SM, we use $Q = 5$ components and randomly hyperparameter initialization due to the unclear spectral density structure. ARIMA, ETS, and LKP using Lévy prior and RJ-MCMC inference cannot be (directly) applied to multi-dimensional data. In this experiment we do not perform automatic pruning of components with the LTH algorithm, due to the relatively low efficiency of its inference process. Results indicate good performance of SLSM when applied to a large dataset.

Table 2 Performance of SLSM and other kernels in terms of average MAE (over 10 independent runs) ± standard deviation, and performance of traditional forecasting methods

Kernel	Rail miles	Electricity	Sunspot	Computer	Pole-telecom
LIN	65.8 ± 20.2	83.0 ± 2.9	37.5 ± 0.9	33.1 ± 0.1	41.0 ± 0.2
SE	530.6 ± 22.1	467.7 ± 60.3	27.4 ± 12.2	32.6 ± 13.6	3.8 ± 0.1
Poly	548.1 ± 4.8	502.1 ± 1.5	39.2 ± 1.1	119.0 ± 3.2	4.2 ± 0.4
PER	84.6 ± 0.9	72.4 ± 26.5	35.5 ± 2.0	82.9 ± 7.5	25.9 ± 13.1
RQ	77.8 ± 11.5	81.3 ± 3.2	36.9 ± 2.6	30.1 ± 12.3	3.7 ± 0.3
MA	495.9 ± 29.9	472.2 ± 10.8	37.4 ± 1.1	20.1 ± 4.1	3.4 ± 0.2
Gabor	548.1 ± 4.8	501.1 ± 4.1	37.8 ± 1.0	120.6 ± 1.8	29.3 ± 0.9
NN	67.9 ± 15.8	75.9 ± 3.1	29.0 ± 1.1	35.5 ± 0.6	11.6 ± 0.1
ADD	528.5 ± 17.5	489.7 ± 8.3	38.0 ± 1.4	13.2 ± 2.3	4.2 ± 0.3
ARIMA	70.5	394.7	103.4	–	–
Auto-ARIMA	534.5	456.6	31.7	–	–
ETS	49.9	111.9	581.2	–	–
SM	56.2 ± 11.8	50.5 ± 4.2	24.3 ± 4.1	17.2 ± 4.8	3.6 ± 0.2
LKP	64.3 ± 12.8	71.7 ± 9.2	24.6 ± 6.1	–	–
SLSM$_{\gamma=0}$	54.8 ± 5.2	64.5 ± 12.9	15.7 ± 3.1	12.0 ± 2.8	3.4 ± 0.1
SLSM	41.8 ± 6.6	41.7 ± 3.5	12.4 ± 1.8	**9.4** ± 1.5	**2.9** ± 0.1
SM (LTH)	47.0 ± 4.4	44.1 ± 1.9	23.9 ± 0.9	–	–
SLSM$_{\gamma=0}$ (LTH)	40.2 ± 5.6	41.6 ± 1.6	13.9 ± 0.5	–	–
SLSM (LTH)	**28.4** ± 3.2	**40.8** ± 0.9	**12.2** ± 1.3	–	–

Bold denotes the best performance result for each experiment

Table 3 Performance of SLSM and other kernels in terms of average NLML (over 10 independent runs) ± standard deviation, and performance of traditional forecasting methods

Kernel	Rail miles	Electricity	Sunspot	Computer	Pole-telecom
LIN	694.3 ± 101.0	375.1 ± 50.1	1336.7 ± 380.1	625.8 ± 14.5	49,926 ± 22.4
SE	671.6 ± 151.1	339.3 ± 49.5	1784.0 ± 540.2	548.4 ± 94.3	40,600 ± 1013.2
Poly	935.2 ± 135.2	379.4 ± 67.0	2620.2 ± 546.6	622.5 ± 114.7	42,670 ± 1351.9
PER	715.6 ± 64.4	359.6 ± 96.2	1702.4 ± 279.1	745.5 ± 3.4	48,543 ± 3287.2
RQ	663.9 ± 96.4	338.6 ± 46.0	1639.6 ± 576.9	257.6 ± 50.9	39,058 ± 175.4
MA	770.0 ± 78.7	339.5 ± 77.4	1898.7 ± 715.6	583.5 ± 61.7	39,359 ± 509.4
Gabor	936.7 ± 136.1	415.3 ± 79.0	2258.3 ± 603.3	824.8 ± 24.0	53,433 ± 0.1
NN	744.3 ± 13.2	358.9 ± 58.0	1043.5 ± 171.8	420.2 ± 17.4	45,119 ± 23.5
ADD	766.0 ± 22.1	335.6 ± 68.5	791.9 ± 223.8	303.0 ± 4.5	39,616 ± 182.6
SM	**595.7** ± 11.8	337.7 ± 32.1	623.9 ± 11.2	**180.2** ± 11.9	49,679 ± 16.3
LKP	699.3 ± 21.3	334.3 ± 5.0	612.1 ± 4.6	–	–
SLSM$_{\gamma=0}$	686.1 ± 23.0	339.5 ± 3.5	612.0 ± 3.0	217.1 ± 5.2	49,630 ± 135.6
SLSM	658.6 ± 14.5	342.5 ± 3.6	**611.4** ± 1.9	205.3 ± 5.2	**36,812** ± 997.2
SM (LTH)	613.3 ± 45.7	337.9 ± 4.4	616.3 ± 6.9	–	–
SLSM$_{\gamma=0}$ (LTH)	641.5 ± 44.2	**333.1** ± 4.3	617.5 ± 1.1	–	–
SLSM (LTH)	638.6 ± 23.4	340.0 ± 4.1	613.0 ± 2.2	–	–

Bold denotes the best performance result for each experiment

6.6 Long-term forecasting: large number of times series

We conclude our empirical investigation with experiments on the M3-Competition monthly data, which includes more than 1000 time series (Makridakis and Hibon 2000). The M3-Competitions data consists mainly of business and economic time series of five types: micro, industry, macro, finance, and demographic. Because most of these time series are short, in order to perform long-term forecasting, we use the first 80% time points of each time series as training data, and the remaining data of each time series as test.

Time series in M3-Competition monthly data have different overall scales. The MSE is sensitive to the overall scale of the test data. Therefore, here we use the Standardized Mean Squared Error (SMSE): $\text{SMSE}(\mathbf{y}_i^*, \tilde{\mathbf{y}}_i^*) = \sum_{i=1}^n \left(y_i^* - \tilde{y}_i^* \right)^2 / (n \cdot \mathbb{V}[\mathbf{y}^*])$ by normalizing with the variance of the target values of the test (Rasmussen and Williams 2006), then average the predictive performances of all M3-Competition monthly time series.

Table 5 contains SMSE results obtained by GP with spectral kernels, and by the ARIMA, and ETS methods. Results indicate superior performance of GP with spectral kernels over traditional forecasting methods. Note, however, that we did not tune the parameters of ARIMA, and ETS methods, but used their default values given in ECOTOOL toolbox.

6.7 Discussion

Results of our experiments indicate that a GP with the SLSM kernel consistently outperforms other baselines in terms of MSE and MAE, but its NLML is not always the lowest. A reason for this phenomenon could be that the marginal likelihood surfaces for the GPs with SM, LKP, and SLSM kernels usually have many local optima. Also, computation of NLML [Eq .(3)] uses the training data, which may show a good fit, but this does not guarantee a better generalization performance. Analogously, our experiments indicate that a better fitting of the empirical spectral density does not necessarily yield the best extrapolation performance (see Figs. 5 and 3). Based on these considerations, MSE and MAE are better evaluation metrics, because they are computed on the test data, with a slight preference for MAE, which is less sensitive to outliers.

Unsatisfactory performance of methods like SE, Poly, MA, Gabor, ADD, and Auto-ARIMA, is due to the fact that these methods are designed to perform interpolation, while we consider extrapolation tasks. Better performance of GPs with SLSM over other spectral kernels like SM and LKP, can be explained by the fact that SLSM extends the covariance range in the GP model, which can correlate test points far away from training and thus be beneficial for long range extrapolation. Specifically, compared to other kernels, the spectral density of the SLSM kernel tends to have very sharp peaks. In addition, we can observe that the spectral density SLSM has much bigger variance than SM and LKP, which further extends the tails of components.

Since the variance in the frequency domain is inversely related to the length-scale in the time domain, these components will lead to a kernel with long-range covariances in the time domain. This can explain the good extrapolation performance of the kernel. Also, compared to the SM and LKP kernels, the SLSM kernel can discover more short period patterns, corresponding to spectral density components with large means. In general the long-term variation is more important for extrapolation because short period patterns mixed with noise and specific local factors are difficult to capture.

Table 4 Effect of the LTH pruning algorithm on GP models, in terms of percentage of kernel components pruned and increase in MAE and MSE performance

Dataset	Kernel	Pruned nr.	Reduced hyps	Improved MSE (%)	Improved MAE (%)
Rail miles	SM (LTH)	2.4	7.2	29.01	16.27
Rail miles	$SLSM_{\gamma=0}$ (LTH)	3.5	10.5	49.19	37.55
Rail miles	SLSM (LTH)	4.1	16.4	17.77	32.15
Sunspot	SM (LTH)	4.7	14.1	2.84	1.56
Sunspot	$SLSM_{\gamma=0}$ (LTH)	5.1	15.3	47.71	43.52
Sunspot	SLSM (LTH)	5.3	21.2	27.41	2.24
Electricity	SM (LTH)	2.6	7.8	8.95	12.65
Electricity	$SLSM_{\gamma=0}$ (LTH)	3.0	9.0	23.58	41.94
Electricity	SLSM (LTH)	2.8	11.2	3.78	2.23

Table 5 Performance in terms of SMSE on the M3-Competition monthly data

Kernel	Micro	Industry	Macro	Finance	Demographic
ARIMA	4.71	2.86	10.86	8.53	13.60
Auto-ARIMA	41.43	13.58	130.06	33.86	72.05
ETS	3.05	4.72	14.87	6.60	18.16
SM	1.92	2.29	7.03	4.58	13.46
LKP	3.71	4.34	7.75	6.57	13.98
$SLSM_{\gamma=0}$	1.76	2.14	4.41	3.46	9.23
SLSM	1.70	1.97	3.81	3.04	8.47
SM (LTH)	1.71	2.06	6.98	4.26	13.31
$SLSM_{\gamma=0}$ (LTH)	1.39	1.87	4.37	3.39	8.33
SLSM (LTH)	**1.23**	**1.63**	**3.76**	**2.93**	**7.76**

Bold denotes the best performance result for each experiment

Note that SLSM ($\gamma = 0$) has the same kernel structure and implementation as LKP, but uses LBFGS implemented in GPML toolbox to perform inference instead of Reversed-Jump MCMC (RJ-MCMC) in LKP (Jang et al. 2017). Results of our experiments indicate that SLSM is also superior to SLSM ($\gamma = 0$), showing that the kernel used in the LKP method (that is, SLSM ($\gamma = 0$)) can be improved by incorporating skewness. Also, results show superior performance of SLSM ($\gamma = 0$) over LKP, indicating that for this type of kernel, gradient descent optimization usually infers a GP model that is better than that obtained using sampling based optimization. In general, we could not compare results achieved independently by the authors of the LKP method, because no quantitative results are reported in their paper. For SM, In general, we could not compare results achieved independently by the authors of the LKP method (Jang et al. 2017), because no quantitative results are reported in their paper. For SM, we run experiments also on the monthly airline passenger with same training and test splitting as in Wilson and Adams (2013). The monthly airline passenger time series has length 144, with measurements from 1949 to 1961. It uses 96 records for training and the rest 48 records as test data. For SM we achieved the following results: average NLML, MAE, and MSE equal to 353.99, 16.87, and 460.04, respectively, close to those achieved independently in Wilson and Adams (2013).

Thus our implementation of SM is in agreement with the original method described in Wilson and Adams (2013). On the same dataset, SMSL achieved the following results: average NLML, MAE, and MSE equal to 354.71, 16.78, 456.24, respectively.

We performed experiments with varying number of components, $Q = 10$, Q between 2 and 10, $Q = 100$, and investigated the effect of pruning components using the proposed LTH algorithm. Results indicate that automatic pruning using our LTH algorithm can drastically reduce the number of kernel components, which reduces the size of the hyperparameter space, and hence improves performance (see Table 4).

7 Conclusion

We have proposed a new kernel for long-term forecasting with GPs, derived in a principled way by modeling the empirical spectral densities using a skewed Laplace spectral mixture, whose characteristics are more beneficial for long-range forecasting than those of the Gaussian mixture used in SM kernels.

Experiments on real-world datasets demonstrated that by using the SLSM kernel dense structures as well as sparse structures in the data can be captured. In addition, the SLSM kernel was shown to be able to learn periodical patterns, especially short-term patterns in the signal.

Overall, the SLSM kernel appears to be especially beneficial for time series containing patterns with long-range dependencies, as indicated, for example, by experiments with the electricity and rail miles datasets. Other real-world tasks (Yan et al. 2009) suited for SLSM include forecasting sea level (Ercan et al. 2013) to estimate the climate change trends, and long-term forecasting of remaining useful battery life (Richardson et al. 2017) to ensure reliable system operation and to minimise maintenance costs.

The SLSM kernel also has a connection with RQ kernel, which not only provides a new interpretation of the RQ kernel, but also demonstrates the usefulness of the SLSM kernel as a principled alternative to the exponential and RQ kernels. We might further extend SLSM to increase its long-range modeling capability by using a RQ kernel with $\alpha < 1$ instead of the Cauchy function term of the SLSM kernel.

The SLSM kernel can also be generalized to scalable inference like Cartesian structured multidimensional datasets like images following Gilboa et al. (2015), Wilson (2014), Wilson et al. (2014) and Wilson and Nickisch (2015). The multidimensional SLSM kernel becomes:

$$k_{\text{SLSM}}(\tau) = \prod_{p=1}^{P} \sum_{i=1}^{Q} k_{\text{SLSM},i}^{(p)}(\tau) \tag{18}$$

where $k_{\text{SLSM},i}^{(p)}(\tau)$ is the i-th SLSM component on the p-th dimension. Equation (18) can be used for fast inference by decomposing K_{SLSM} as the Kronecker product of matrices over each input dimension $K_{\text{SLSM}} = K_{\text{SLSM}}^1 \otimes \cdots \otimes K_{\text{SLSM}}^P$. It is interesting to investigate applications involving such data.

The main limitation of the proposed kernel is that it is not suited for modeling nonstationary phenomena (Remes et al. 2017). An interesting open problem is exploring how to overcome this limitation without loss of interpretation.

References

Bhowmick, D., Davison, A., Goldstein, D. R., & Ruffieux, Y. (2006). A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics, 7*(4), 630–641.

Bochner, S. (2016). *Lectures on Fourier Integrals. (AM-42)* (Vol. 42). Princeton University Press.

Chalupka, K., Williams, C. .K., & Murray, I. (2013). A framework for evaluating approximation methods for Gaussian process regression. *Journal of Machine Learning Research, 14*(Feb), 333–350.

Chen, K., van Laarhoven, T., Chen, J., & Marchiori, E. (2019). Incorporating dependencies in spectral kernels for Gaussian processes. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 565–581). Springer.

Deisenroth, M., & Ng, J. W. (2015). Distributed Gaussian processes. In *Proceedings of The 32nd international conference on machine learning* (pp. 1481–1490).

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., & Zoubin, G. (2013). Structure discovery in nonparametric regression through compositional kernel search, pp. 1166–1174. PMLR, Atlanta, Georgia, USA. http://proceedings.mlr.press/v28/duvenaud13.html.

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

Eltoft, T., Kim, T., & Lee, T. (2006). On the multivariate Laplace distribution. *IEEE Signal Processing Letters, 13*(5), 300–303. https://doi.org/10.1109/LSP.2006.870353.

Ercan, A., Kavvas, M. L., & Abbasov, R. K. (2013). *Long-range dependence and sea level forecasting*. Springer.

Fragiadakis, K., & Meintanis, S. G. (2011). Goodness-of-fit tests for multivariate Laplace distributions. *Mathematical and Computer Modelling, 53*(5–6), 769–779.

Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *7th international conference on learning representations, ICLR 2019*, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net https://openreview.net/forum?id=rJl-b3RcF7.

Gazor, S., & Zhang, W. (2003). Speech probability distribution. *Signal Processing Letters IEEE, 10*(7), 204–207. https://doi.org/10.1109/LSP.2003.813679.

Gilboa, E., Saatci, Y., & Cunningham, J. P. (2015). Scaling multidimensional inference for structured Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(2), 424–436. https://doi.org/10.1109/TPAMI.2013.192.

Herlands, W., Wilson, A., Nickisch, H., Flaxman, S., Neill, D., Van Panhuis, W., & Xing, E. (2016). Scalable Gaussian processes for characterizing multidimensional change surfaces. In *Artificial intelligence and statistics* (pp. 1013–1021).

Hoffman, F. (1997). An introduction to Fourier theory. *Extraído el, 2*.

Hyun Jin Park, T. W. L. (2004). Modeling nonlinear dependencies in natural images using mixture of Laplacian distribution.

Jang, P. A., Loeb, A., Davidow, M., & Wilson, A. G. (2017). Scalable Lévy process priors for spectral kernel learning. In *Advances in neural information processing systems* (pp. 3943–3952).

Kostantinos, N. (2000). Gaussian mixtures and their applications to signal processing. In *Advanced signal processing handbook: theory and implementation for radar, sonar, and medical imaging real time systems* (pp. 3-1).

Kotz, S., Kozubowski, T.J., & Podgórski, K. (2001). Asymmetric multivariate Laplace distribution. In *The Laplace distribution and generalizations* (pp. 239–272). Springer.

Kozubowski, T. J., & Podgórski, K. (2000). A multivariate and asymmetric generalization of Laplace distribution. *Computational Statistics, 15*(4), 531–540. https://doi.org/10.1007/PL00022717.

Makridakis, S., & Hibon, M. (2000). The m3-competition: Results, conclusions and implications. *International Journal of Forecasting, 16*(4), 451–476.

Minguillón, J., & Pujol, J. (2001). JPEG standard uniform quantization error modeling with applications to sequential and progressive operation modes. *Journal of Electronic Imaging, 10*(2), 475–485. https://doi.org/10.1117/1.1344592.

Moon, T. K. (1997). The expectation-maximization algorithm. *IEEE Signal Processing Magazine, 13*(6), 47–60.

Nguyen, T., & Bonilla, E. (2014). Fast allocation of gaussian process experts. In *Proceedings of The 31st international conference on machine learning* (pp. 145–153).

Oliva, J. B., Dubey, A., Wilson, A. G., Póczos, B., Schneider, J., & Xing, E. P. (2016). Bayesian nonparametric kernel-learning. In *Artificial intelligence and statistics* (pp. 1078–1086).

Pearce, T., Tsuchida, R., Zaki, M., Brintrup, A., & Neely, A. (2020). Expressive priors in Bayesian neural networks: Kernel combinations and periodic functions. In *Uncertainty in artificial intelligence* (pp. 134–144). PMLR.

Pedregal, D. J. (2019). Time series analysis and forecasting with ECOTOOL. *PLoS One, 14*(10), e0221238.

Pessanha, J. F. M., & Leon, N. (2015). Forecasting long-term electricity demand in the residential sector. *Procedia Computer Science, 55,* 529–538.

Quiñonero-Candela, J., & Rasmussen, C .E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research, 6*(Dec), 1939–1959.

Rasmussen, C .E., & Nickisch, H. (2010). Gaussian processes for machine learning (GPML) toolbox. *Journal of Machine Learning Research, 11*(Nov), 3011–3015.

Rasmussen, C .E., & Williams, C .K. (2006). *Gaussian process for machine learning*. MIT press.

Remes, S., Heinonen, M., & Kaski, S. (2017). Non-stationary spectral kernels. In *Advances in neural information processing systems* (pp. 4645–4654).

Richardson, R. R., Osborne, M. A., & Howey, D. A. (2017). Gaussian process regression for forecasting battery state of health. *Journal of Power Sources, 357,* 209–219.

Samo, Y. L. K., & Roberts, S. (2015). Generalized spectral kernels. arXiv preprint arXiv:1506.02236.

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. 3273–3282). PMLR.

Stein, M. (1999). Interpolation of spatial data: Some theory for kriging.

Sun, S., Zhang, G., Wang, C., Zeng, W., Li, J., & Grosse, R. (2018). Differentiable compositional kernel learning for Gaussian processes. arXiv preprint arXiv:1806.04326.

Visk, H. (2009). On the parameter estimation of the asymmetric multivariate Laplace distribution. *Communications in Statistics-Theory and Methods, 38*(4), 461–470.

Wilson, A., & Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th international conference on machine learning (ICML-13)* (pp. 1067–1075).

Wilson, A. .G. (2014). *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. University of Cambridge.

Wilson, A. G., Gilboa, E., Nehorai, A., & Cunningham, J. P. (2014). Fast kernel learning for multidimensional pattern extrapolation. In *Advances in neural information processing systems* (pp. 3626–3634).

Wilson, A. G., & Nickisch, H. (2015). Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In: F. R. Bach & D. M. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning, ICML 2015*, Lille, France, 6–11 July 2015, *JMLR Workshop and Conference Proceedings*, vol. 37, pp. 1775–1784. JMLR.org . http://jmlr.org/proceedings/papers/v37/wilson15.html.

Yan, W., Qiu, H., & Xue, Y. (2009). Gaussian process for long-term time-series forecasting. In *2009 international joint conference on neural networks* (pp. 3420–3427) . https://doi.org/10.1109/IJCNN.2009.5178729.

Authors and Affiliations

Kai Chen[1,2] (ID) **· Twan van Laarhoven[1] · Elena Marchiori[1]**

Twan van Laarhoven
tvanlaarhoven@cs.ru.nl

Elena Marchiori
elenam@cs.ru.nl

[1]   Institute for Computing and Information Sciences (iCIS), Radboud University, Nijmegen,
      The Netherlands

[2]   Future Network of Intelligence Institute (FNii), The Chinese University of Hong Kong, Shenzhen,
      China