Gaussian Process Kernels for Pattern Discovery and Extrapolation

Andrew Gordon Wilson   AGW38@CAM.AC.UK
Department of Engineering, University of Cambridge, Cambridge, UK

Ryan Prescott Adams   RPA@SEAS.HARVARD.EDU
School of Engineering and Applied Sciences, Harvard University, Cambridge, USA

Abstract

Gaussian processes are rich distributions over functions, which provide a Bayesian nonparametric approach to smoothing and interpolation. We introduce simple closed form kernels that can be used with Gaussian processes to discover patterns and enable extrapolation. These kernels are derived by modelling a spectral density – the Fourier transform of a kernel – with a Gaussian mixture. The proposed kernels support a broad class of stationary covariances, but Gaussian process inference remains simple and analytic. We demonstrate the proposed kernels by discovering patterns and performing long range extrapolation on synthetic examples, as well as atmospheric CO_2 trends and airline passenger data. We also show that it is possible to reconstruct several popular standard covariances within our framework.

1. Introduction

Machine learning is fundamentally about pattern discovery. The first machine learning models, such as the perceptron (Rosenblatt, 1962), were based on a simple model of a neuron (McCulloch & Pitts, 1943). Papers such as Rumelhart et al. (1986) inspired hope that it would be possible to develop intelligent agents with models like neural networks, which could automatically discover hidden representations in data. Indeed, machine learning aims not only to equip humans with tools to analyze data, but to fully automate the learning and decision making process.

Research on Gaussian processes (GPs) within the machine learning community developed out of neural net-

works research, triggered by Neal (1996), who observed that Bayesian neural networks became Gaussian processes as the number of hidden units approached infinity. Neal (1996) conjectured that "there may be simpler ways to do inference in this case".

These simple inference techniques became the cornerstone of subsequent Gaussian process models for machine learning (Rasmussen & Williams, 2006). These models construct a prior directly over functions, rather than parameters. Assuming Gaussian noise, one can analytically infer a posterior distribution over these functions, given data. Gaussian process models have become popular for non-linear regression and classification (Rasmussen & Williams, 2006), and often have impressive empirical performance (Rasmussen, 1996).

The properties of likely functions under a GP, e.g., smoothness, periodicity, etc., are controlled by a positive definite *covariance kernel*[1], an operator which determines the similarity between pairs of points in the domain of the random function. The choice of kernel profoundly affects the performance of a Gaussian process on a given task – as much as the choice of architecture, activation functions, learning rate, etc., can affect the performance of a neural network.

Gaussian processes are sometimes used as expressive statistical tools, where the pattern discovery is performed by a human, and then hard coded into parametric kernels. Often, however, the squared exponential (Gaussian) kernel is used by default. In either case, GPs are used as smoothing interpolators with a fixed (albeit infinite) set of basis functions. Such simple smoothing devices are not a realistic replacement for neural networks, which were envisaged as intelligent agents that could discover hidden features in data[2] via adaptive basis functions (MacKay, 1998).

---

[1] The terms *covariance kernel*, *covariance function*, *kernel function*, and *kernel* are used interchangeably.

[2] We refer to *representations*, *features* and *patterns* interchangeably. *Features* sometimes means low dimensional representations of data, like neurons in a neural network.

However, Bayesian nonparametrics can help build automated intelligent systems that reason and make decisions. It has been suggested that the human ability for inductive reasoning – concept generalization with remarkably few examples – could derive from a prior combined with Bayesian inference (Yuille & Kersten, 2006; Tenenbaum et al., 2011; Steyvers et al., 2006). Bayesian nonparametric models, and Gaussian processes in particular, are an expressive way to encode prior knowledge, and also reflect the belief that the real world is infinitely complex (Neal, 1996).

With more expressive kernels, one could use Gaussian processes to learn hidden representations in data. Expressive kernels have been developed by combining Gaussian processes in a type of Bayesian neural network structure (Salakhutdinov & Hinton, 2008; Wilson et al., 2012; Damianou & Lawrence, 2012). However, these approaches, while promising, typically 1) are designed to model specific types of structure (e.g., input-dependent correlations between different tasks); 2) make use of component GPs with simple interpolating kernels; 3) indirectly induce complicated kernels that do not have a closed form and are difficult to interpret; and 4) require sophisticated approximate inference techniques that are much more demanding than that required by simple analytic kernels.

Sophisticated kernels are most often achieved by composing together a few standard kernel functions (Archambeau & Bach, 2011; Durrande et al., 2011; Gönen & Alpaydın, 2011; Rasmussen & Williams, 2006). Tight restrictions are typically enforced on these compositions and they are hand-crafted for specialized applications. Without such restrictions, complicated compositions of kernels can lead to overfitting and unmanageable hyperparameter inference. Moreover, while some compositions (e.g., addition) have an interpretable effect, many other operations change the distribution over functions in ways that are difficult to identify. It is difficult, therefore, to construct an effective inductive bias for kernel composition that leads to automatic discovery of the appropriate statistical structure, without human intervention.

This difficulty is exacerbated by the fact that it is challenging to say anything about the covariance function of a stochastic process from a single draw if *no* assumptions are made. If we allow the covariance between any two points in the input space to arise from *any* positive definite function, with equal probability, then we gain essentially no information from a single realization. Most commonly one assumes a restriction to *stationary* kernels, meaning that covariances are invariant to translations in the input space.

In this paper, we explore flexible classes of kernels that go beyond composition of simple analytic forms, while maintaining the useful inductive bias of stationarity. We propose new kernels which can be used to automatically discover patterns and extrapolate far beyond the available data. This class of kernels contains many stationary kernels, but has a simple closed form that leads to straightforward analytic inference. The simplicity of these kernels is one of their strongest qualities. In many cases, these kernels can be used as a drop in replacement for the popular squared exponential kernel, with benefits in performance and expressiveness. By learning features in data, we not only improve predictive performance, but we can more deeply understand the structure of the problem at hand – greenhouse gases, air travel, heart physiology, brain activity, etc.

After a brief review of Gaussian processes in Section 2, we derive the new kernels in Section 3 by modelling a spectral density with a mixture of Gaussians. We focus our experiments in Section 4 on elucidating the fundamental differences between the proposed kernels and the popular alternatives in Rasmussen & Williams (2006). In particular, we show how the proposed kernels can automatically discover patterns and extrapolate on the CO_2 dataset in Rasmussen & Williams (2006), on a synthetic dataset with strong negative covariances, on a difficult synthetic sinc pattern, and on airline passenger data. We also use our framework to reconstruct several popular standard kernels.

2. Gaussian Processes

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution. Using a Gaussian process, we can define a distribution over functions $f(x)$,

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

where $x \in \mathbb{R}^P$ is an arbitrary input variable, and the mean function $m(x)$ and covariance kernel $k(x, x')$ are defined as

$$m(x) = \mathbb{E}[f(x)], \qquad (2)$$
$$k(x, x') = \text{cov}(f(x), f(x')). \qquad (3)$$

Any collection of function values has a joint Gaussian distribution

$$[f(x_1), f(x_2), \ldots, f(x_N)]^\top \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}), \qquad (4)$$

where the $N \times N$ covariance matrix $\mathbf{K}$ has entries $\mathrm{K}_{ij} = k(x_i, x_j)$, and the mean $\boldsymbol{\mu}$ has entries $\mu_i = m(x_i)$. The properties of the functions – smoothness, periodicity, etc. – are determined by the kernel function.

The popular *squared exponential* (SE) kernel has the form

$$k_{\text{SE}}(x, x') = \exp(-0.5||x - x'||^2/\ell^2). \quad (5)$$

Functions drawn from a Gaussian process with this kernel function are infinitely differentiable, and can display long range trends. GPs with a squared exponential kernel are simply smoothing devices: the only covariance structure that can be learned from data is the length-scale ℓ, which determines how quickly a Gaussian process function varies with x.

Assuming Gaussian noise, one can analytically infer a posterior predictive distribution over Gaussian process functions, and analytically derive a marginal likelihood of the observed function values $\boldsymbol{y}$ given only hyperparameters θ, and the input locations $\{x_n\}_{n=1}^N$, $p(\boldsymbol{y}|\theta, \{x_n\}_{n=1}^N)$. This marginal likelihood can be optimised to estimate hyperparameters such as ℓ, or used to integrate out the hyperparameters via Markov chain Monte Carlo (Murray & Adams, 2010). Detailed Gaussian process references include Rasmussen & Williams (2006), Stein (1999), and Cressie (1993).

3. Kernels for Pattern Discovery

In this section we introduce a class of kernels that can discover patterns, extrapolate, and model negative covariances. This class contains a large set of stationary kernels. Roughly, a kernel measures the similarity between data points. As in Equation (3), the covariance kernel of a GP determines how the associated random functions will tend to vary with inputs (predictors) $x \in \mathbb{R}^P$. A *stationary kernel* is a function of $\tau = x - x'$, i.e., it is invariant to translation of the inputs.

Any stationary kernel (aka covariance function) can be expressed as an integral using Bochner's theorem (Bochner, 1959; Stein, 1999):

Theorem 3.1 *(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*

$$k(\tau) = \int_{\mathbb{R}^P} e^{2\pi i s^\top \tau} \psi(\mathrm{d}s), \quad (6)$$

where ψ is a positive finite measure.

If ψ has a density $S(s)$, then S is called the *spectral density* or *power spectrum* of k, and k and S are Fourier duals (Chatfield, 1989):

$$k(\tau) = \int S(s) e^{2\pi i s^\top \tau} ds, \quad (7)$$

$$S(s) = \int k(\tau) e^{-2\pi i s^\top \tau} d\tau. \quad (8)$$

In other words, a spectral density entirely determines the properties of a stationary kernel. Substituting the squared exponential kernel of (5) into (8), we find its spectral density is $S_{\text{SE}}(s) = (2\pi\ell^2)^{P/2} \exp(-2\pi^2\ell^2 s^2)$. Therefore SE kernels, and mixtures of SE kernels, are a very small corner of the set of possible stationary kernels, as they correspond only to Gaussian spectral densities centered on the origin.

However, by using a mixture of Gaussians that have non-zero means, one can achieve a much wider range of spectral densities. Indeed, mixtures of Gaussians are dense in the set of all distribution functions (Kostantinos, 2000). Therefore, the dual of this set is also dense in stationary covariances. That is, we can approximate *any* stationary covariance kernel to arbitrary precision, given enough mixture components in the spectral representation. This observation motivates our approach, which is to model GP covariance functions via spectral densities that are scale-location mixtures of Gaussians.

We first consider a simple case, where

$$\phi(s\,;\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(s - \mu)^2\}, \quad \text{and} \quad (9)$$

$$S(s) = [\phi(s) + \phi(-s)]/2, \quad (10)$$

noting that spectral densities are symmetric (Rasmussen & Williams, 2006). Substituting $S(s)$ into equation (7), we find

$$k(\tau) = \exp\{-2\pi^2\tau^2\sigma^2\} \cos(2\pi\tau\mu). \quad (11)$$

If $\phi(s)$ is instead a mixture of Q Gaussians on $\mathbb{R}^P$, where the qth component has mean vector $\boldsymbol{\mu}_q = (\mu_q^{(1)}, \ldots, \mu_q^{(P)})$ and covariance matrix $\mathbf{M}_q = \text{diag}(v_q^{(1)}, \ldots, v_q^{(P)})$, and τ_p is the pth component of the P dimensional vector $\tau = x - x'$, then

$$k(\tau) = \sum_{q=1}^Q w_q \prod_{p=1}^P \exp\{-2\pi^2\tau_p^2 v_q^{(p)}\} \cos(2\pi\tau_p\mu_q^{(p)}). \quad (12)$$

The integral in (7) is tractable even when the spectral density is an arbitrary Gaussian mixture, allowing us to derive[3] the exact closed form expressions in Eqs. (11) and (12), and to perform analytic inference with Gaussian processes. Moreover, this class of kernels is expressive – containing many stationary kernels – but nevertheless has a simple form.

These kernels are easy to interpret, and provide drop-in replacements for kernels in Rasmussen & Williams (2006). The weights w_q specify the relative contribution of each mixture component. The inverse

---

[3]Detailed derivations of Eqs. (11) and (12), and code, are in the supplement (Wilson & Adams, 2013).

means $1/\mu_q$ are the component periods, and the inverse standard deviations $1/\sqrt{v_q}$ are length-scales, determining how quickly a component varies with the inputs x. The kernel in Eq. (12) can also be interpreted through its associated spectral density. In Section 4, we use the learned spectral density to interpret the number of discovered patterns in the data and how these patterns generalize. Henceforth, we refer to the kernel in Eq. (12) as a spectral mixture (SM) kernel.

4. Experiments

We show how the SM kernel in Eq. (12) can be used to discover patterns, extrapolate, and model negative covariances. We contrast the SM kernel with popular kernels in, e.g., Rasmussen & Williams (2006) and Abrahamsen (1997), which typically only provide smooth interpolation. Although the SM kernel generally improves predictive likelihood over popular alternatives, we focus on clearly visualizing the learned kernels and spectral densities, examining patterns and predictions, and discovering structure in the data. Our objective is to elucidate the fundamental differences between the proposed SM kernel and the alternatives.

In all experiments, Gaussian noise is assumed, so that marginalization over the unknown function can be performed in closed form. Kernel hyperparameters are trained using nonlinear conjugate gradients to optimize the marginal likelihood $p(\boldsymbol{y}|\theta, \{x_n\}_{n=1}^N)$ of the data $\boldsymbol{y}$ given hyperparameters θ, as described in Section 2, assuming a zero mean GP. A type of "automatic relevance determination" (MacKay et al., 1994; Tipping, 2004) takes place during training, minimizing the effect of extraneous components in the proposed model, through the complexity penalty in the marginal likelihood (Rasmussen & Williams, 2006). Moreover, the exponential terms in the SM kernel of Eq. (12) have an annealing effect on the marginal likelihood, reducing multimodality in the frequency parameters, making it easier to naively optimize the marginal likelihood without converging to undesirable local optima. For a fully Bayesian treatment, the spectral density could alternatively be integrated out using Markov chain Monte Carlo (Murray & Adams, 2010), rather than choosing a point estimate. However, we wish to emphasize that the SM kernel can be successfully used in the same way as other popular kernels, without additional inference efforts.

We compare with the popular squared exponential (SE), Matérn (MA), rational quadratic (RQ), and periodic (PE) kernels. In each comparison, we attempt to give these alternative kernels fair treatment: we initialise hyperparameters at values that give high marginal likelihoods and which are well suited to the datasets, based on features we can already see in the data. Conversely, we randomly initialise the parameters for the SM kernel. Training runtimes are on the order of minutes for all tested kernels. In these examples, comparing with multiple kernel learning (MKL) (Gönen & Alpaydın, 2011) has limited additional value. MKL is not typically intended for pattern discovery, and often uses mixtures of SE kernels. Mixtures of SE kernels correspond to scale-mixtures of Gaussian spectral densities, and do not perform well on these data, which are described by highly multimodal non-Gaussian spectral densities.

4.1. Extrapolating Atmospheric CO_2

Keeling & Whorf (2004) recorded monthly average atmospheric CO_2 concentrations at the Mauna Loa Observatory, Hawaii. The data are shown in Figure 1. The first 200 months are used for training (in blue), and the remaining 301 months (≈ 25 years) are used for testing (in green).

This dataset was used in Rasmussen & Williams (2006), and is frequently used in Gaussian process tutorials, to show how GPs are flexible statistical tools: a human can look at the data, recognize patterns, and then hard code those patterns into covariance kernels. Rasmussen & Williams (2006) identify, by looking at the blue and green curves in Figure 1a, a long term rising trend, seasonal variation with possible decay away from periodicity, medium term irregularities, and noise, and hard code a stationary covariance kernel to represent each of these features.

However, in this view of GP modelling, all of the interesting pattern discovery is done by the human user, and the GP is used simply as a smoothing device, with the flexibility to incorporate human inferences in the prior. Our contention is that such pattern recognition can also be performed algorithmically. To discover these patterns without encoding them *a priori* into the GP, we use the spectral mixture kernel in Eq. (12), with $Q = 10$. The results are shown in Figure 1a.

In the training region, predictions using each kernel are essentially equivalent, and entirely overlap with the training data. However, unlike the other kernels, the SM kernel (in black) is able to discover patterns in the training data and accurately extrapolate over a long range. The 95% high predictive density (HPD) region contains the true CO_2 measurements for the duration of the measurements.

We can see the structure discovered by the SM kernel in the learned log spectral density in Figure 1b. Of the $Q = 10$ components, only seven were used. There is a peak at a frequency near 0.07, which corresponds
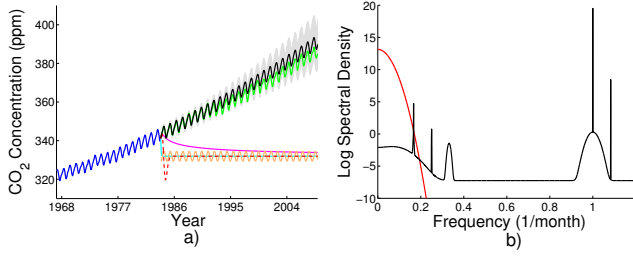
Figure 1. Mauna Loa CO_2 Concentrations. a) Forecasting CO_2. The training data are in blue, the testing data in green. Mean forecasts made using the SM kernel are in black, with 2 stdev about the mean (95% of the predictive mass) in gray shade. Predictions using the Matérn (MA), squared exponential (SE), rational quadratic (RQ), and periodic kernels (PE) are in cyan, dashed red, magenta, and orange, respectively. b) The log spectral densities of the learned SM and squared exponential kernels are in black and red, respectively.

to a period of about 14 months, roughly a year, which is in line with seasonal variations. However, there are departures away from periodicity, and accordingly, there is a large variance about this peak, reflecting the uncertainty in this learned feature. We see other, sharper peaks corresponding to periods of 6 months, 4 months, 3 months and 1 month.

In red, we show the spectral density for the learned SE kernel, which misses many of the frequencies identified as important using the SM kernel. The peak at a frequency of 1 is an artifact caused by aliasing. While artifacts caused by aliasing (noise, mean function, etc.) into the kernel do not affect the performance of the model, for interpretability it can be useful to restrict the model from learning frequencies greater than the sampling rate of the data.

All SE kernels have a Gaussian spectral density centred on zero. Since a mixture of Gaussians centred on the origin is a poor approximation to many density functions, combinations of SE kernels have limited expressive power. Indeed the spectral density learned by the SM kernel in Figure 1b is highly non-Gaussian. The test predictive performance using the SM, SE, MA, RQ, and PE kernels is given in Table 1, under the heading CO_2.

4.2. Recovering Popular Kernels

The SM class of kernels contains many stationary kernels, since mixtures of Gaussians can be used to construct a wide range of spectral densities. Even with a small number of mixture components, e.g., $Q \leq 10$, the SM kernel can closely recover popular stationary kernels catalogued in Rasmussen & Williams (2006).
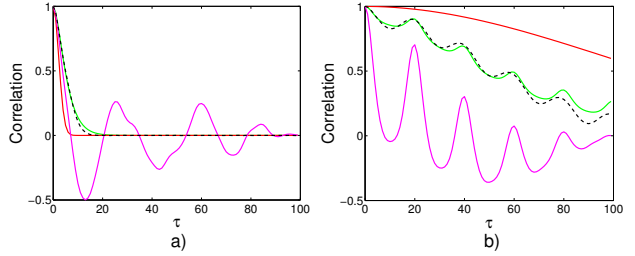

Figure 2. Recovering popular correlation functions (normalised kernels), with $\tau = x - x'$. The true correlation function underlying the data is in green, and SM, SE, and empirical correlation functions are in dashed black, red, and magenta, respectively. Data are generated from a) a Matérn kernel, and b) a sum of RQ and periodic kernels.

As an example, we start by sampling 100 points from a one-dimensional GP with a Matérn kernel with degrees of freedom $\nu = 3/2$:

$$k_{\mathrm{MA}}(\tau) = a(1 + \frac{\sqrt{3}\tau}{\ell}) \exp(-\frac{\sqrt{3}\tau}{\ell}), \qquad (13)$$

where $\ell = 5$ and $a = 4$. Sample functions from a Gaussian process with this Matérn kernel are far less smooth (only once-differentiable) than Gaussian process functions with a squared exponential kernel.

We attempt to reconstruct the kernel underlying the data by training an SM kernel with $Q = 10$. After training, only $Q = 4$ components are used. The log marginal likelihood of the data – having integrated away the Gaussian process – using the trained SM kernel is -133, compared to -138 for the Matérn kernel that generated the data. Training the SE kernel in (5) gives a log marginal likelihood of -140.

Figure 2a shows the learned SM correlation function[4], compared to the generating Matérn correlation function, the empirical autocorrelation function, and learned squared exponential correlation function. Although often used in geostatistics to guide choices of GP kernels (and parameters) (Cressie, 1993), the empirical autocorrelation function tends to be unreliable, particularly with a small amount of data (e.g., $N < 1000$), and at high lags (for $\tau > 10$). In Figure 2a, the empirical autocorrelation function is erratic and does not resemble the Matérn kernel for $\tau > 10$. Moreover, the squared exponential kernel cannot capture the heavy tails of the Matérn kernel, no matter what length-scale it has. Even though the SM kernel is infinitely differentiable, it can closely approximate processes which are finitely differentiable, be-

---

[4]A *correlation function* $c(x, x')$ is a normalised covariance kernel $k(x, x')$, such that $c(x, x') = k(x, x')/\sqrt{k(x, x)k(x', x')}$ and $c(x, x) = 1$.

cause mixtures of Gaussians can closely approximate the spectral densities of these processes, even with a small number of components, as in Figure 2a.

Next, we reconstruct a mixture of the rational quadratic (RQ) and periodic kernels (PE) in Rasmussen & Williams (2006):

$$k_{\mathrm{RQ}}(\tau) = (1 + \frac{\tau^2}{2\,\alpha\,\ell_{RQ}^2})^{-\alpha}, \qquad (14)$$

$$k_{\mathrm{PE}}(\tau) = \exp(-2\sin^2(\pi\,\tau\,\omega)/\ell_{PE}^2). \qquad (15)$$

The rational quadratic kernel in (14) is derived as a scale mixture of squared exponential kernels with different length-scales. The standard periodic kernel in (15) is derived by mapping the two dimensional variable $u(x) = (\cos(x), \sin(x))$ through the squared exponential kernel in (5). Derivations for both the RQ and PE kernels in (14) and (15) are in Rasmussen & Williams (2006). Rational quadratic and Matérn kernels are also discussed in Abrahamsen (1997). We sample 100 points from a Gaussian process with kernel $10k_{\mathrm{RQ}} + 4k_{\mathrm{PE}}$, with $\alpha = 2$, $\omega = 1/20$, $\ell_{RQ} = 40$, $\ell_{PE} = 1$.

We reconstruct the kernel function of the sampled process using an SM kernel with $Q = 4$, with the results shown in Figure 2b. The heavy tails of the RQ kernel are modelled by two components with large periods (essentially aperiodic), and one short length-scale and one large length-scale. The third component has a relatively short length-scale, and a period of 20. There is not enough information in the 100 sample points to justify using more than $Q = 3$ components, and so the fourth component in the SM kernel has no effect, through the complexity penalty in the marginal likelihood. The empirical autocorrelation function somewhat captures the periodicity in the data, but significantly underestimates the correlations. The squared exponential kernel learns a long length-scale: since the SE kernel is highly misspecified with the true generating kernel, the data are explained as noise.

4.3. Negative Covariances

All of the stationary covariance functions in the standard machine learning Gaussian process reference Rasmussen & Williams (2006) are everywhere positive, including the periodic kernel, $k(\tau) = \exp(-2\sin^2(\pi\,\tau\,\omega)/\ell^2)$. While positive covariances are often suitable for interpolation, capturing negative covariances can be essential for extrapolating patterns: for example, linear trends have long-range negative covariances. We test the ability of the SM kernel to learn negative covariances, by sampling
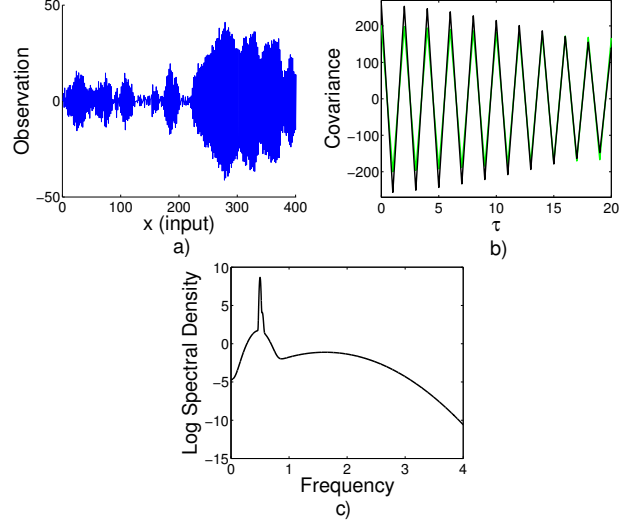

Figure 3. Negative Covariances. a) Observations of a discrete time autoregressive (AR) series with negative covariances. b) The SM learned kernel is shown in black, while the true kernel of the AR series is in green, with $\tau = x - x'$. c) The spectral density of the learned SM kernel is in black.

400 points from a simple AR(1) discrete time GP:

$$y(x+1) = -e^{-0.01}y(x) + \sigma\epsilon(x), \qquad (16)$$

$$\epsilon(x) \sim \mathcal{N}(0, 1), \qquad (17)$$

which has kernel

$$k(x, x') = \sigma^2(-e^{-.01})^{|x-x'|}/(1 - e^{-.02}). \qquad (18)$$

The process in Eq. (16) is shown in Figure 3a. This process follows an oscillatory pattern, systematically switching states every $x = 1$ unit, but is not periodic and has long range covariances: if we were to only view every second data point, the resulting process would vary rather slowly and smoothly.

We see in Figure 3b that the learned SM covariance function accurately reconstructs the true covariance function. The spectral density in Figure 3c) shows a sharp peak at a frequency of 0.5, or a period of 2. This feature represents the tendency for the process to oscillate from positive to negative every $x = 1$ unit. In this case $Q = 4$, but after automatic relevance determination in training, only 3 components were used. Using the SM kernel, we forecast 20 units ahead and compare to other kernels in Table 1 (`NEG COV`).

4.4. Discovering the Sinc Pattern

The sinc function is defined as $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$. We created a pattern combining three sinc functions:

$$y(x) = \mathrm{sinc}(x+10) + \mathrm{sinc}(x) + \mathrm{sinc}(x-10). \qquad (19)$$
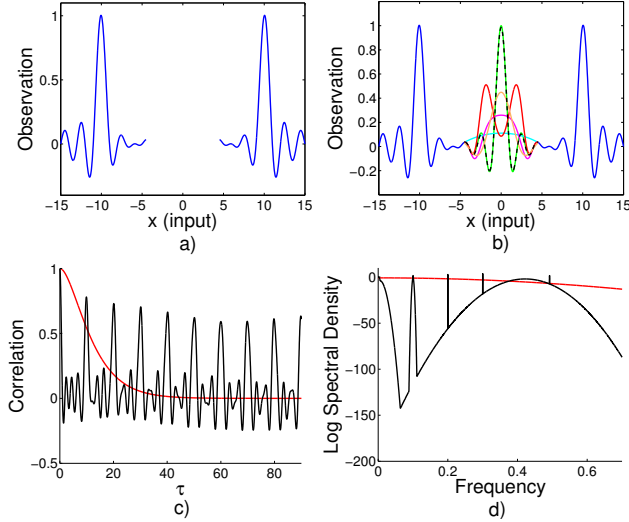
Figure 4. Discovering a sinc pattern. a) The training data are shown in blue. The goal is to fill in the missing region $x \in [-4.5, 4.5]$. b) The training data are in blue, the testing data in green. The mean of the predictive distribution using the SM kernel is shown in dashed black. The mean of the predictive distributions using the squared exponential, Matérn, rational quadratic, and periodic kernels, are in red, cyan, magenta, and orange. c) The learned SM correlation function (normalised kernel) is shown in black, and the learned Matérn correlation function is in red, with $\tau = x - x'$. d) The log spectral densities of the SM and SE kernels are in black and red, respectively.

This is a complex oscillatory pattern. Given only the points shown in Figure 4a, we wish to complete the pattern for $x \in [-4.5, 4.5]$. Unlike the CO_2 example in Section 4.1, it is perhaps even difficult for a human to extrapolate the missing pattern from the training data. It is an interesting exercise to focus on this figure, identify features, and fill in the missing part.

Notice that there is complete symmetry about the origin $x = 0$, peaks at $x = -10$ and $x = 10$, and destructive interference on each side of the peaks facing the origin. We therefore might expect a peak at $x = 0$ and a symmetric pattern around $x = 0$.

As shown in Figure 4b, the SM kernel with $Q = 10$ reconstructs the pattern in the region $x \in [-4.5, 4.5]$ almost perfectly from the 700 training points in blue. Moreover, using the SM kernel, 95% of the posterior predictive mass entirely contains the true pattern in $x \in [-4.5, 4.5]$. GPs using Matérn, SE, RQ, and periodic kernels are able to predict reasonably within $x = 0.5$ units of the training data, but entirely miss the pattern in $x \in [-4.5, 4.5]$.

Figure 4c shows the learned SM correlation function (normalised kernel). For $\tau \in [0, 10]$ there is a local

pattern, roughly representing the behaviour of a single sinc function. For $\tau > 10$ there is a repetitive pattern representing a new sinc function every 10 units – an extrapolation a human might make. With a sinc functions centred at $x = -10, 0, 10$, we might expect more sinc patterns every 10 units. The learned Matérn correlation function is shown in red in Figure 4c – unable to discover complex patterns in the data, it simply assigns high correlation to nearby points.

Figure 4d shows the (highly non-Gaussian) spectral density of the SM kernel, with peaks at $0.003, 0.1, 0.2, 0.3, 0.415, 0.424, 0.492$. In this case, only 7 of the $Q = 10$ components are used. The peak at 0.1 represents a period of 10: every 10 units, a sinc function is repeated. The variance of this peak is small, meaning the method will extrapolate this structure over long distances. By contrast, the squared exponential spectral density simply has a broad peak, centred at the origin. The predictive performance for recovering the missing 300 test points (in green) is given in Table 1 (SINC).

4.5. Airline Passenger Data

Figure 5a shows airline passenger numbers, recorded monthly, from 1949 to 1961 (Hyndman, 2005). Based on only the first 96 monthly measurements, in blue, we wish to forecast airline passenger numbers for the next 48 months (4 years); the corresponding 48 test measurements are in green. Companies wish to make such long range forecasts to decide upon expensive long term investments, such as large passenger jets, which can cost hundreds of millions of dollars.

There are several features apparent in these data: short seasonal variations, a long term rising trend, and an absence of white noise artifacts. Many popular kernels are forced to make one of two choices: 1) Model the short term variations and ignore the long term trend, at the expense of extrapolation. 2) Model the long term trend and treat the shorter variations as noise, at the expense of interpolation.

As seen in Figure 5a, the Matérn kernel is more inclined to model the short term trends than the smoother SE or RQ kernels, resulting in sensible interpolation (predicting almost identical values to the training data in the training region), but poor extrapolation – moving quickly to the prior mean, having learned no structure to generalise beyond the data. The SE kernel interpolates somewhat sensibly, but appears to underestimate the magnitudes of peaks and troughs, and treats repetitive patterns in the data as noise. Extrapolation using the SE kernel is poor. The RQ kernel, which is a scale mixture of SE kernels, is more able to manage different length-scales in the
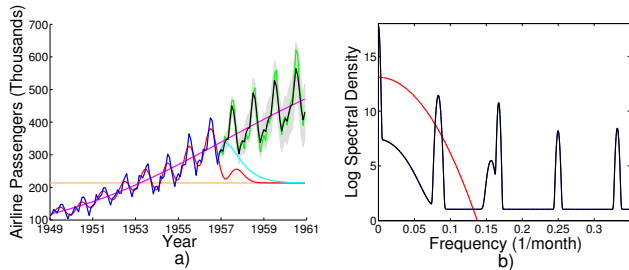
Figure 5. Predicting airline passenger numbers. a) The training and testing data are in blue and green, respectively. The mean of the predictive distribution using the SM kernel is shown in black, with 95% of the predictive probability mass shown in gray shade. The mean of the predictive distribution using the SM kernel is shown in black. The mean of the predictive distributions using the SE, Matérn, RQ, and periodic kernels, are in red, cyan, magenta, and orange, respectively. In the training region, the SM and Matérn kernels are not shown, since their predictions essentially overlap with the training data. b) The log spectral densities of the SM and squared exponential kernels are in black and red, respectively.

Table 1. We compare the test performance of the proposed spectral mixture (SM) kernel with squared exponential (SE), Matérn (MA), rational quadratic (RQ), and periodic (PE) kernels. The SM kernel consistently has the lowest mean squared error (MSE) and highest log likelihood ($\mathcal{L}$).

	SM	SE	MA	RQ	PE
CO$_2$					
MSE	**9.5**	1200	1200	980	1200
$\mathcal{L}$	**170**	−320	−240	−100	−1800
NEG COV					
MSE	**62**	210	210	210	210
$\mathcal{L}$	**−25**	−70	−70	−70	−70
SINC					
MSE	**0.000045**	0.16	0.10	0.11	0.05
$\mathcal{L}$	**3900**	2000	1600	2000	600
AIRLINE					
MSE	**460**	43000	37000	4200	46000
$\mathcal{L}$	**−190**	−260	−240	−280	−370

data, and generalizes the long term trend better than the SE kernel, but interpolates poorly.

By contrast, the SM kernel interpolates nicely (overlapping with the data in the training region), and is able to extrapolate complex patterns far beyond the data, capturing the true airline passenger numbers for years after the data ends, within a small band containing 95% of the predictive probability mass.

Of the $Q = 10$ initial components specified for the SM kernel, 7 were used after training. The learned spectral density in Figure 5b shows a large sharp low frequency peak (at about 0.00148). This peak corresponds to the rising trend, which generalises well beyond the data (small variance peak), is smooth (low frequency), and is important for describing the data (large relative weighting). The next largest peak corresponds to the yearly trend of 12 months, which again generalises, but not to the extent of the smooth rising trend, since the variance of this peak is larger than for the peak near the origin. The higher frequency peak at $x = 0.34$ (period of 3 months) corresponds to the beginnings of new seasons, which can explain the effect of seasonal holidays on air traffic. A more detailed study of these features and their properties – frequencies, variances, etc. – could isolate less obvious factors affecting airline passenger numbers. Table 1 (`AIRLINE`) shows predictive performance for forecasting 48 months of airline passenger numbers.

5. Discussion

We have derived expressive closed form kernels and we have shown that these kernels, when used with Gaussian processes, can discover patterns in data and extrapolate over long ranges. The simplicity of these kernels is one of their strongest properties: they can be used as drop-in replacements for popular kernels such as the squared exponential kernel, with major benefits in expressiveness and performance, while retaining simple training and inference procedures.

Gaussian processes have proven themselves as smoothing interpolators. Pattern discovery and extrapolation is an exciting new direction for Bayesian nonparametric approaches, which can capture rich variations in data. We have shown how Bayesian nonparametric models can naturally be used to generalise a pattern from a small number of examples.

We have only begun to explore what could be done with such pattern discovery methods. In future work, one could integrate away a spectral density, using recently developed efficient MCMC for GP hyperparameters (Murray & Adams, 2010). Moreover, recent Toeplitz methods (Saatchi, 2011) could be applied to the SM kernel to significantly speed up inference.

References

Abrahamsen, P. A review of Gaussian random fields and correlation functions. *Norweigan Computing Center Technical report*, 1997.

Archambeau, C. and Bach, F. Multiple Gaussian process models. *arXiv preprint arXiv:1110.5238*, 2011.

Bochner, Salomon. *Lectures on Fourier Integrals.(AM-42)*, volume 42. Princeton University Press, 1959.

Chatfield, C. *Time Series Analysis: An Introduction.* London: Chapman and Hall, 1989.

Cressie, N.A.C. *Statistics for Spatial Data (Wiley Series in Probability and Statistics).* Wiley-Interscience, 1993.

Damianou, A.C. and Lawrence, N.D. Deep Gaussian processes. *arXiv preprint arXiv:1211.0358*, 2012.

Durrande, N., Ginsbourger, D., and Roustant, O. Additive kernels for Gaussian process modeling. *arXiv preprint arXiv:1103.4023*, 2011.

Gönen, M. and Alpaydın, E. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

Hyndman, R.J. Time series data library. 2005. http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/.

Keeling, C. D. and Whorf, T. P. Atmospheric CO2 records from sites in the SIO air sampling network. *Trends: A Compendium of Data on Global Change. Carbon Dioxide Information Analysis Center*, 2004.

Kostantinos, N. Gaussian mixtures and their applications to signal processing. *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*, 2000.

MacKay, David J.C. Introduction to Gaussian processes. In Christopher M. Bishop, editor (ed.), *Neural Networks and Machine Learning*, chapter 11, pp. 133–165. Springer-Verlag, 1998.

MacKay, D.J.C. et al. Bayesian nonlinear modeling for the prediction competition. *Ashrae Transactions*, 100(2):1053–1062, 1994.

McCulloch, W.S. and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biology*, 5(4):115–133, 1943.

Murray, Iain and Adams, Ryan P. Slice sampling covariance hyperparameters in latent Gaussian models. In *Advances in Neural Information Processing Systems 23*, 2010.

Neal, R.M. *Bayesian Learning for Neural Networks.* Springer Verlag, 1996. ISBN 0387947248.

Rasmussen, Carl Edward. *Evaluation of Gaussian Processes and Other Methods for Non-linear Regression.* PhD thesis, University of Toronto, 1996.

Rasmussen, Carl Edward and Williams, Christopher K.I. *Gaussian processes for Machine Learning.* The MIT Press, 2006.

Rosenblatt, F. *Principles of Neurodynamics.* Spartan Book, 1962.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Saatchi, Yunus. *Scalable Inference for Structured Gaussian Process Models.* PhD thesis, University of Cambridge, 2011.

Salakhutdinov, R. and Hinton, G. Using deep belief nets to learn covariance kernels for Gaussian processes. *Advances in Neural Information Processing Systems*, 20:1249–1256, 2008.

Stein, M.L. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Verlag, 1999.

Steyvers, M., Griffiths, T.L., and Dennis, S. Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, 10(7):327–334, 2006.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022):1279–1285, 2011.

Tipping, M. Bayesian inference: An introduction to principles and practice in machine learning. *Advanced Lectures on Machine Learning*, pp. 41–62, 2004.

Wilson, A.G. and Adams, R.P. Gaussian process kernels for pattern discovery and extrapolation supplementary material and code. 2013. http://mlg.eng.cam.ac.uk/andrew/smkernelsupp.pdf.

Wilson, Andrew G., Knowles, David A., and Ghahramani, Zoubin. Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Wilson, Andrew Gordon, Knowles, David A, and Ghahramani, Zoubin. Gaussian process regression networks. *arXiv preprint arXiv:1110.4411*, 2011.

Yuille, A. and Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301–308, 2006.