A Process over all Stationary Covariance Kernels

Andrew Gordon Wilson

June 9, 2012

Abstract

I define a process over all stationary covariance kernels. I show how one might be able to perform inference that scales as $\mathcal{O}(nm^2)$ in a GP regression model using this process as a prior over the covariance kernel, with n datapoints and $m < n$. I also show how the stationarity assumption can be relaxed.

1 Introduction

Gaussian process regression models in machine learning (Rasmussen and Williams, 2006) are considered to be "Bayesian nonparametric models". However, at the heart of every Gaussian process regression model – *controlling all the modelling power* – is a parametrised covariance kernel, greatly restricting the flexibility of the corresponding Gaussian process. One would almost never believe that the true process underlying real data has a parametrised kernel used in Gaussian process regression.

A fully Bayesian nonparametric treatment of regression would place a nonparametric prior over the Gaussian process covariance kernel, to represent uncertainty over the values of the kernel function, and to reflect the belief that the kernel does not have a simple parametric form. However, typically we only have access to a single realisation of a stochastic process, and therefore it is difficult to extract useful information about the covariance structure of that process if we make *no* assumptions about the covariance kernel; for example, to estimate $\text{cov}(f(x_i), f(x_j))$ we could only use the single pair $(f(x_i), f(x_j))$. On the other hand, if we were to assume the process $f(x)$ is stationary, we could estimate $\text{cov}(f(x_i), f(x_j))$, by considering the function $f(x)$ evaluated at all pairs of points (x_a, x_w) such that $x_a - x_w = x_i - x_j$.

In this paper, I define a process over all stationary kernels, and use it as a prior over the covariance kernel in a Gaussian process regression (or classification) model. With this process, interesting covariance structures – periodicity, Markovian dynamics, etc. – and mixtures of covariance structures can be discovered without having to a priori "hard-code" these structures into a (sum of) parametric covariance kernels. Moreover, in typical Gaussian process regression or classification, with a parametrised kernel, inference is $\mathcal{O}(n^3)$, where n is the number of datapoints. In this process, inference could possibly be $\mathcal{O}(m^2 n)$, where $m < n$. I also show how the stationarity assumption can be relaxed.

2 Prior over Stationary Kernels

Bochner's theorem says that any stationary kernel can be expressed as

$$k(x_i, x_j) = \int \exp(2\pi i s \cdot (x_i - x_j)) d\mu(s)\,, \tag{1}$$

where μ is a positive finite measure, and x_i, x_j are $p \times 1$ vector valued inputs.

If μ has a density $S(s)$, we can write Bochner's theorem as

$$k(x_i, x_j) = \int \exp(2\pi i s \cdot (x_i - x_j)) S(s) ds\,, \tag{2}$$

$$S(s) = \int k(x_i, x_j) \exp(-2\pi i s \cdot (x_i - x_j)) d(x_i - x_j)\,. \tag{3}$$

Therefore there is some "spectral density" or "power spectrum" $S(s)$ corresponding to any popular parametrised stationary covariance kernel. For example, consider the popular squared exponential covariance kernel,

$$k_{SE}(x_i, x_j) = a_0 \exp(-0.5(x_i - x_j)^\top R^{-1}(x_i - x_j))\,, \tag{4}$$

where $R = \mathrm{diag}(l_1^2, \ldots, l_p^2)$. Then using (3),

$$S(s) = a_0 \sqrt{|2\pi R|} \exp(-2\pi^2 s^\top R s) \tag{5}$$

The normalised spectral density $p(s) = S(s)/k(0,0)$.

Since mixtures of Gaussian are dense in the set of distribution functions, Bochner's theorem says that if our spectral density is an infinite mixture of Gaussians, then the induced process will have support for any stationary kernel functions.

First, let us consider $S(s)$ when it is a single Gaussian, and integrate (2), assuming x_i, x_j, s are scalars:

$$k(x_i, x_j) = \int \exp(2\pi i s(x_i - x_j)) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(s - \mu)^2) ds \tag{6}$$

let $m = x_i - x_j$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp[2\pi i m - \frac{1}{2\sigma^2}(s^2 - 2\mu s + \mu^2)] ds \tag{7}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp[-\frac{1}{2\sigma^2} s^2 + (2\pi i m + \frac{\mu}{\sigma^2}) s - \frac{\mu^2}{\sigma^2}] ds \tag{8}$$

$$\text{let } a = \frac{1}{2\sigma^2}\,, b = 2\pi i m + \frac{\mu}{\sigma^2}\,, c = -\frac{\mu^2}{2\sigma^2}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int \exp(-a(s - \frac{b}{2a})^2) \exp(\frac{b^2}{4a} + c) ds \tag{9}$$

$$= \exp[(2\pi i m + \frac{\mu}{\sigma^2})^2 \frac{\sigma^2}{2} - \frac{\mu^2}{2\sigma^2}] \tag{10}$$

$$= \exp[(-4\pi^2 m^2 + 4\pi i m \frac{\mu}{\sigma^2} + \frac{\mu^2}{\sigma^4}) \frac{\sigma^2}{2} - \frac{\mu^2}{2\sigma^2}] \tag{11}$$

$$= \exp[-2\pi^2 (x_i - x_j)^2 \sigma^2][\cos(2\pi(x_i - x_j)\mu) + i \sin(2\pi(x_i - x_j)\mu))]\,. \tag{12}$$

Therefore if $S(s) = \sum_{q=1}^{Q} w_q \exp(-\frac{1}{2\sigma_q^2}(s-\mu_q)^2)$, then the integral (2) is analytic, and will be a sum of Q terms of the form (12). This result easily generalises to the case where x_i, x_j, s are p-dimensional vectors, and $S(s)$ is a sum of p variate Gaussians with axis aligned covariance matrices; the integral in (2) becomes a product of integrals like (6).

Notice that when $\mu = 0$ the imaginary part of (12) disappears. Indeed, the spectral density $S(s)$ should be symmetric about the origin $s = 0$. Therefore we can let $S(s) = [p(s) + p(-s)]/2$, where $p(s)$ is an infinite mixture of Gaussians, and still have coverage of all stationary kernels. To derive (12) we let $S(s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(s-\mu)^2)$. If we instead follow the same derivation, but using $S(-s)$ as the spectral density, the final result is

$$k(x_i, x_j) = \exp[-2\pi^2(x_i - x_j)^2\sigma^2][\cos(2\pi(x_i - x_j)\mu) - i\sin(2\pi(x_i - x_j)\mu))]\,. \quad (13)$$

Therefore if $p(s)$ is a Gaussian, then a spectral density $S(s) = p(s) + p(-s)$ will be real with kernel

$$k(x_i, x_j) = \exp[-2\pi^2(x_i - x_j)^2\sigma^2][\cos(2\pi(x_i - x_j)\mu)]\,. \quad (14)$$

So if $p(s)$ is an infinite mixture of univariate Gaussians, then the kernel for $S(s) = p(s) + p(-s)$ is

$$k(x_i, x_j) = \lim_{Q\to\infty} \sum_{q=1}^{Q} w_q \exp[-2\pi^2(x_i - x_j)^2\sigma_q^2][\cos(2\pi(x_i - x_j)\mu_q)] \quad (15)$$

Supposing $p(s)$ is an infinite mixture of p-dimensional axis aligned Gaussians, where the q^{th} component has mean vector $\boldsymbol{\mu}_q = (\mu_q^{(1)}, \ldots, \mu_q^{(p)})$ and covariance matrix $M = \text{diag}(v_q^{(1)}, \ldots, v_q^{(p)})$, and m_j is the jth component of the p dimensional vector $x_i - x_j$, and $S(s) = p(s) + p(-s)$,

$$k(x_i, x_j) = \lim_{Q\to\infty} \sum_{q=1}^{Q} w_q \prod_{a=1}^{p} \exp[-2\pi^2 m_a^2 v_q^{(a)}][\cos(\pi m_a \mu_q^{(a)})] \quad (16)$$

3 Inference

We let $S(s) = [p(s) + p(-s)]/2$, where

$$p(s) = \lim_{q\to\infty} \sum_{q=1}^{Q} w_q \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp[\frac{(s-\mu_q)^2}{2\sigma_q^2}]\,. \quad (17)$$

$w_q \sim \text{GEM}(\alpha)$, and we can put Gaussian and inverse Gamma priors on μ_q and σ_q^2 respectively.

The observations $y(x) \sim \mathcal{N}(f(x),)$, where $f(x) \sim \mathcal{GP}(0, k)$. For any set of observations $\boldsymbol{y} = (y_1(x_1), \ldots, y_n(x_n))^\top$ we wish to perform inference over $\boldsymbol{f} = (f_1(x_1), \ldots, f_n(x_n))^\top$. Every pair of function values $f(x_i), f(x_j)$ is assigned to a cluster with mean and variance $\mu_{z_{ij}}$ and $\sigma_{z_{ij}}^2$ where z_{ij} is a cluster assignment variable. Given cluster assignments means, and variances, $\gamma = \{z_{ij}\}, \{\mu_q\}, \{\sigma_q^2\}$, $p(\boldsymbol{f}|\boldsymbol{y}, \gamma)$ is Gaussian. We can infer posteriors over γ and $\boldsymbol{f}$ using a CRP Gibbs

sampling procedure, with slice sampling or HMC for non-conjugate updates. However, based on my experience Gibbs sampling hyperparameters of GP regression models, I think this would mix very poorly. It would be preferable to do VB inference, and consider the GP function values and DPM components jointly, using the stick breaking construction? Although bookkeeping would be more painful, it would probably also be better to immediately work with multivariate axis aligned Gaussians. Or it may be *a lot* better to work with isotropic covariance functions, and only use 1D Gaussians in our DPM mixture (see the next section).

By calculating the spectral density of a particular covariance function (e.g. see (5)) we can set the means on our prior mixture components so that the expected kernel for our process is that covariance function. This way our process can use that parametrized kernel as a "base kernel". I suspect this could be valuable for good performance and reasonable inference – and it will let one incorporate more intuition into our process.

4 Isotropic Covariance Function

It may be advantageous to further restrict our covariance functions to be *isotropic*, and not just stationary. I think the covariance function for most stationary processes will be isotropic to a good approximation, and this assumption will allow us to extract yet more signal from the data for learning the covariance function. The other *major benefit* is we can use *univariate* Gaussians in our DPM for the kernel function, regardless of the dimensionality of the input space. The drawback is we have to evaluate a Bessel function.

If the covariance function is isotropic it can be proven that the spectral density S(s) is a function of $|s|$ only – e.g. we can henceforth treat s as one dimensional. Supposing the input space is p dimensional, and we let $r = ||x_i - x_j||$, then switching to spherical coordinates,

$$k(r) = \frac{2\pi}{r^{p/2-1}} \int_0^\infty S(s) J_{p/2-1}(2\pi r s) s^{p/2} ds \,, \tag{18}$$

$$S(s) = \frac{2\pi}{s^{p/2} - 1} \int_0^\infty k(r) J_{p/2-1}(2\pi r s) r^{p/2} dr \,, \tag{19}$$

where $J_{p/2-1}$ is a Bessel function of order $p/2-1$. A Bessel function of order α is

$$J_\alpha(x) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\,\Gamma(m+\alpha+1)} \left(\tfrac{1}{2}x\right)^{2m+\alpha} \tag{20}$$

We can analytically integrate a product of a Gaussian $S(s)$ and a polynomial $J_{p/2-1}(2\pi r s) s^{p/2}$ quite easily. The spectral density $S(s)$ for a Matern covariance function is given in equation 4.15 of Rasmussen and Williams (2006).

5 Efficient Inference

Lázaro-Gredilla et al. (2010) evaluate (2) by re-writing it as

$$k(x_i, x_j) = \int \exp(2\pi i s \cdot (x_i - x_j)) S(s) ds \tag{21}$$

$$= \int \exp(2\pi i s \cdot x_i) \exp(2\pi i s \cdot x_j)^* S(s) ds \,. \tag{22}$$

(22) is just

$$\mathbb{E}_{S(s)}[\exp(2\pi i s \cdot x_i) \exp(2\pi i s \cdot x_j)^*] \,, \tag{23}$$

where * is a complex conjugate operator. They estimate (23) using a simple Monte Carlo sum, noting that it is valid to sample pairs $s_r, -s_r$, since the spectral density is symmetric about the origin. By sampling in this way, the complex part of the simple Monte carlo sum is eliminated:

$$\mathbb{E}_{S(s)}[\exp(2\pi i s \cdot x_i) \exp(2\pi i s \cdot x_j)^*] \approx \frac{1}{m} \sum_{r=1}^{m} \cos(2\pi s_r \cdot (x_i - x_j)) \; = k(x_i, x_j). \tag{24}$$

This is *exactly* the same covariance function as in Bayesian linear regression with trigonometric basis functions. The marginal likelihood is in equation 8 of Lázaro-Gredilla et al. (2010) and can be evaluated in $\mathcal{O}(nm^2)$ operations, where m is the number of basis functions, and n is the number of observations. In their paper they determine the values of the points $\{s_1, \ldots, s_m\}$ (called spectral frequencies in the Bayesian linear regression model) by maximizing the Bayesian linear regression marginal likelihood. Sometimes this model works well, but sometimes it severely overfits: see figures 7 and 8 in Lázaro-Gredilla et al. (2010). They are basically just doing Bayesian linear regression with trigonometric basis functions and maximum likelihood for setting the parameters of the basis functions. The only relation it has to typical "Bayesian nonparametric" Gaussian process regression is that as the number of trinogometric basis functions $\to \infty$, they can, in principle, approximate any GP with any stationary kernel arbitrarily well; however, in practice, they won't be able to do this too well (at all?) with maximum likelihood estimation of the s_r. They would need to be able to estimate the real spectral density of whatever covariance function they want arbitrarily well (given the data), and to sample the spectral frequencies from that spectral density to do their simple Monte Carlo sum. But to do that, they would need to use our DPM mixture model to estimate the spectral density of the true process.

We could possibly exploit the connection between our kernel function in (15) and the kernel function for a Bayesian linear regression model with a finite number of basis functions (for a choice of basis functions that gives us the same form for the covariance kernel as in (15)). Although we are using an infinite mixture of Gaussians, we can "truncate" the mixture at e.g. $m = 100$ Gaussians. Our model would be more flexible and would not suffer from the overfitting and performance problems in Lázaro-Gredilla et al. (2010). For instance, we can naturally incorporate a base kernel, and will not overfit. Our process could also be used as a general prior over covariance kernels in any GP regression model.

6 Non-stationarity

We can allow for some nonstationarity by instead using a dependent Dirichlet process mixture. We can let the mixing weights $\{w_q\}$ be input dependent, such that $e.g. w_q = w_q(a, b)$. If $w_i(a, b)$ has support for any continuous function of two inputs, then we can essentially have support for any covariance function. We can construct a dependent dirichlet process with such support by using the stick breaking construction, and generating $w_i(a, b)$ by transforming a Gaussian process (with two inputs) such that it has beta marginals (e.g. (Wilson and Ghahramani, 2010)).

References

Lázaro-Gredilla, M., Quiñonero-Candela, J., Rasmussen, C., and Figueiras-Vidal, A. (2010). Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for Machine Learning*. The MIT Press.

Wilson, A. G. and Ghahramani, Z. (2010). Copula processes. In *NIPS*.