Contents

Introduction to GPs (Wilkinson)

Isotropic $\subset$ stationary processes

GPs inherit its properties primarily from the kernel:

- Smoothness

- Differentiability: relate to how often the function changes direction (up/down). A covariance function going up and down around zero will generate wiggly functions.
- Variance

Away from observed data, the GP reverts to the prior. Keep this in mind when choosing a prior.

RBF kernel amplitude

∞ differentiable

$k(x, x') = A^2 exp(-(x - x')^2/s^2)$

Thus oscillates $\pm 3A$: about three standard deviations

Matern family

Choose n times differentiable

Brownian motion

Brownian motion: integral of white noise

White noise

$k(x, x') = \delta_{xx'}$

Linear functions

Given that $y(x) = cx$ with $c \sim N(0, 1)$.

First, $p(y(x)|c) = \int dc \delta(y(x) - cx) N(c; 0, 1) = 1/x N(y(x)/x; 0, 1)$. This is indeed a normal distribution whose width increases as x.

Second, the expectation of the unknown function values (unknown since c is not given) given two coordinates x_1, x_2 is:

$$< y(x_1)y(x_2) > \equiv \int dy_1 dy_2 dc y_1 y_2 p(y_1, y_2, c|x_1, x_2)$$

The integrations over y_1 and y_2 are trivial due to the definite relation between y and x once c is known. For this model,

$$p(y(x), c|x) = p(y(x)|x, c)p(c|x) = \delta(y - cx)p(c).$$

Thus

$$< y(x_1)y(x_2) >= \int dccx_1cx_2p(c) = x_1x_2 < c^2 >= x_1x_2.$$

Why use GPs?

1. Mathematical niceties

Assume $f, g \sim GP$.

- Closed under addition: $(f + g) \sim GP$
- Closed under conditioning
- Closed under any linear operator L: $Lf \sim GP(Lm, L^2k)$. This includes differentiation and integration… and convolution!

2. Non-parametric extension of linear regression

Linear regression (least squares) and GP regression are equivalent when $k(x, x') = xx'$.

GP regression can be though of as linear regression with $\phi(x)$ where ϕ is a possibly infinite vector of features or basis functions.

Reproducing Kernel Hilbert spaces

This seems to be about fitting spans of basis functions to data and could be useful to the sflinear project, especially since convolution with a fixed function is a linear operator.

Kernel design (Durrande)

Positive semi-definiteness is necessary for the requirement that linear combinations of Gaussian variables must have variance ≥ 0.

Very nice 3D example on slide 7.

Very nice visualization of GPs: https://github.com/awav/interactive-gp

> Looking at samples from a kernel is one way of understanding how your choice of kernel is affecting the assumptions underlying the GP.

Testing the quality of fitted GP

Test on held-out set (new data):

- Accurate mean?
- % of points falling in confidence intervals?
- Correct empirical covariances?

In addition, check if the normalised residuals are independent N(0, 1).

Empirical distributions of ML parameters can be obtained through ad-hoc procedures such as leave-one-out cross validation (fit new ML parameters for each left out datapoint).

Maximum likelihood

Slide 24.

The $|k(X, X)|$ term in the likelihood $L(\sigma^2, \theta)$ acts as a regularization term, favoring models with more structure (simpler). This is probably why it is also called the marginal likelihood.

Composing kernels

Summing

Adding kernels and sampling $\equiv$ sampling from individual kernels and summing

Removing a fitted linear trend from data $\equiv$ using a linear trend mean function

Using a linear function in the mean is almost equivalent to using a linear trend in $k(x, x')$ if the variance of the xx' term is large. But he prefers to put it in the $k(x, x')$ because there is some regularization being done.

Additive models in high dimensions are one way to cover the space with only n data points scaling linearly with the amount of dimensions

Product

Function composition

This is like warping (non-linear transformation) of the input space.

Remarkable: $k(x, x') = xx'$ is a valid kernel $\rightarrow k(x, x') = f(x)f(x')$ is a valid kernel for any f

Round table with Durrante

Linear regression with basis functions

Q: Linear regression with a finite set of parametrized basis functions (the sflinear model): can it be generalized automatically to GPs?

A: Yes. The amplitudes and the noise are Gaussian, so you have a stochastic function which can be represented by a GP. The expression for the kernel will depend on the form of your basis functions.

And then he said something about defining an inner product for the basis functions which relates to RKHS.

Positivity and moment constraints of GPs

A GP f must have support on the whole real line because, for example, the marginal distribution at $f(x)$ must be Gaussian, and thus $f(x) \in R$.

Positivity can be enforced by taking the exponential, **or transforming the data**, for example taking the log. Another way can be to differentiate a monotonic GP, since differentiation is a linear operator. Search for "monotonic Gaussian Processes".

Finally, it is possible to construct GPs which integrate to zero by using a linear transformation. See `Durrande Kernel design.pdf` slide 72.

This and the previous question can be new directions to the sflinear model.

Sparse covariance matrices

Q: Short lengthscales will often generate matrices which are approximately zero far away from the diagonal. Can this be exploited?

A: First of all, a sparse covariance matrix does not imply a sparse precision matrix: the correlation between pairs of points can and does propagate (see MacKay's `humble.pdf`). But there exist banded

kernels for which the linear algebra does simplify. For example, the precision matrix of Matern kernels can be computed in closed form, and $O(N)$ complexity can be reached (`day2/Dai Variational Gaussian Processes.pdf` slide 58). Sparsity in the precision matrix can be exploited as well; there is a whole field for that called **Gaussian Markov fields**. The efficiency of Kalman filters which make prediction at linear cost is based on this sparsity.

The use of (non-conjugate) priors for GP hyperparameters

Q: How common is it? What is the computational cost?

A: It doesn't seem that common. You would need to do MCMC (HMC) or variational inference for non-conjugate likelihoods.

Representation learning with GPs (Wilson)

Very good slides!

The tension between flexibility and prior information: you need good inductive biases that are not too restrictive, but just gently nudge.

Neural nets can have useful inductive biases when used as kernels.

Spectral mixture kernels for stationary GPs

Express the spectrum as a mixture of Gaussians for automatic kernel learning. Here the inductive bias is the stationarity assumption.

Aside: Student-t processes

(Shah, Wilson, and Ghahramani 2014)

Interesting: the predictive variance depends on the values of the observations, unlike vanilla GPs.

Functional kernel learning

Ideal for Nick's thesis. See `edu/mscthesis/nick/notes.md`

Shah, Amar, Andrew Wilson, and Zoubin Ghahramani. 2014. "Student-T Processes as Alternatives to Gaussian Processes." In *Artificial Intelligence and Statistics*, 877–85.