Gaussian Processes

Robin Aldridge-Sutton

Definition

A Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution, e.g. for any $\mathbf{x} \in \mathbb{R}^k$,

$$f(\mathbf{x}) \sim N(\mu(\mathbf{x}) = \mathbf{0}, \Sigma = K(\mathbf{x}, \mathbf{x})),$$
$$K(\mathbf{x}, \mathbf{x}')_{i,j} = k(x_i, x'_j),$$

where $k(x, x')$ is called the covariance or kernel function, and can be any real positive-definite kernel, a symmetric function making Σ real and positive-definite, which means

$$\mathbf{x}^T \Sigma \mathbf{x} > 0,$$

implying that every linear combination of values of f has positive variance, and that Σ^{-1} exists, and that L exists, such that $LL^T = \Sigma$, e.g.

$$k(x_i, x'_j) = \sigma_f^2 \exp\left(\frac{(x_i - x'_j)^2}{2l^2}\right).$$

This is known as the squared exponential covariance function, and is an example of a positive-definite kernel that is defined by a function which is applied to the difference between the inputs, which is known as a positive-definite function (there is an alternative definition in dynamical systems).

Sampling

You can sample from a GP by taking the square-root of the covariance matrix and multiplying a standard normal vector $\mathbf{v} \sim N(\mathbf{0}, I)$ by it before adding the mean,

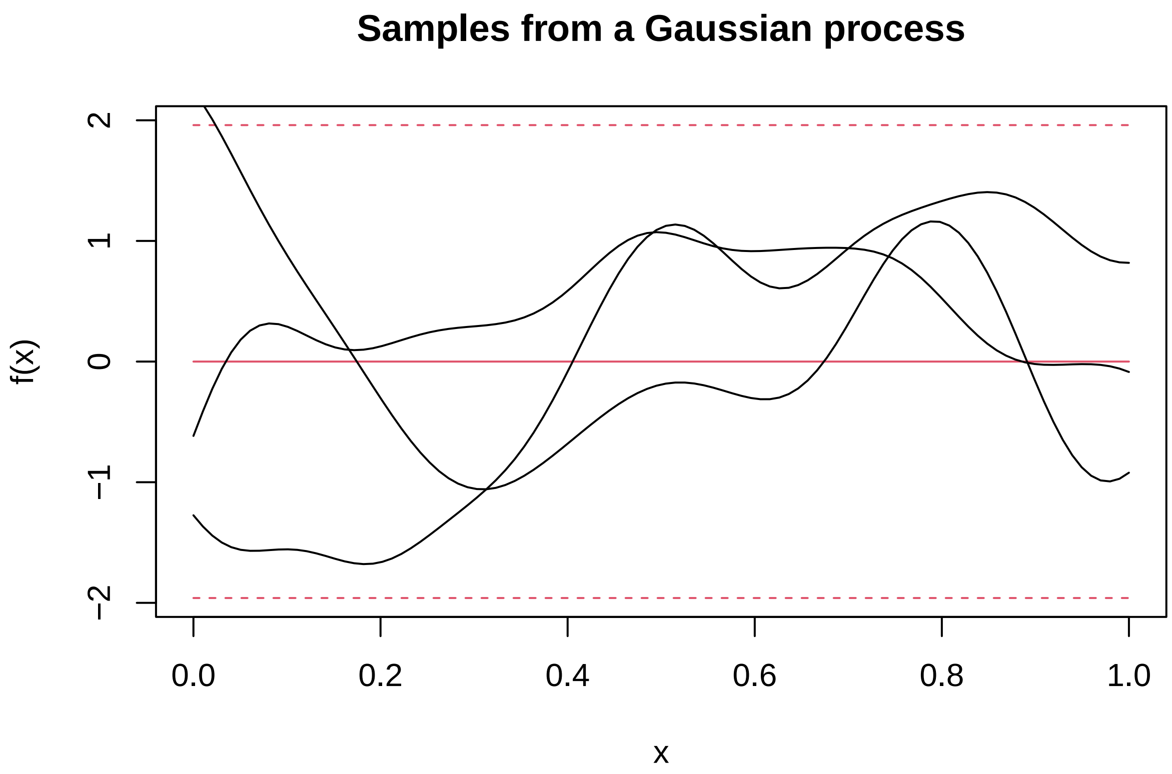
$$cov(L\mathbf{v} + \mu(\mathbf{x})) = Lcov(\mathbf{v})L^T = LL^T = \Sigma$$

$$E(L\mathbf{v} + \mu(\mathbf{x})) = LE(\mathbf{v}) + \mu(\mathbf{x}) = \mu(\mathbf{x})$$

$$\implies L\mathbf{v} + \mu(\mathbf{x}) \sim N(\mu(\mathbf{x}), \Sigma)$$

```
# Functions to sample from and predict values of a Gaussian process.
source("GP funcs.R")

# Plot samples from a GP
plot_GP_samps(
  l = 0.1, # Length scale
  sigma_f = 1, # Function standard deviation
  n_samps = 3 # Number of samples
)
```



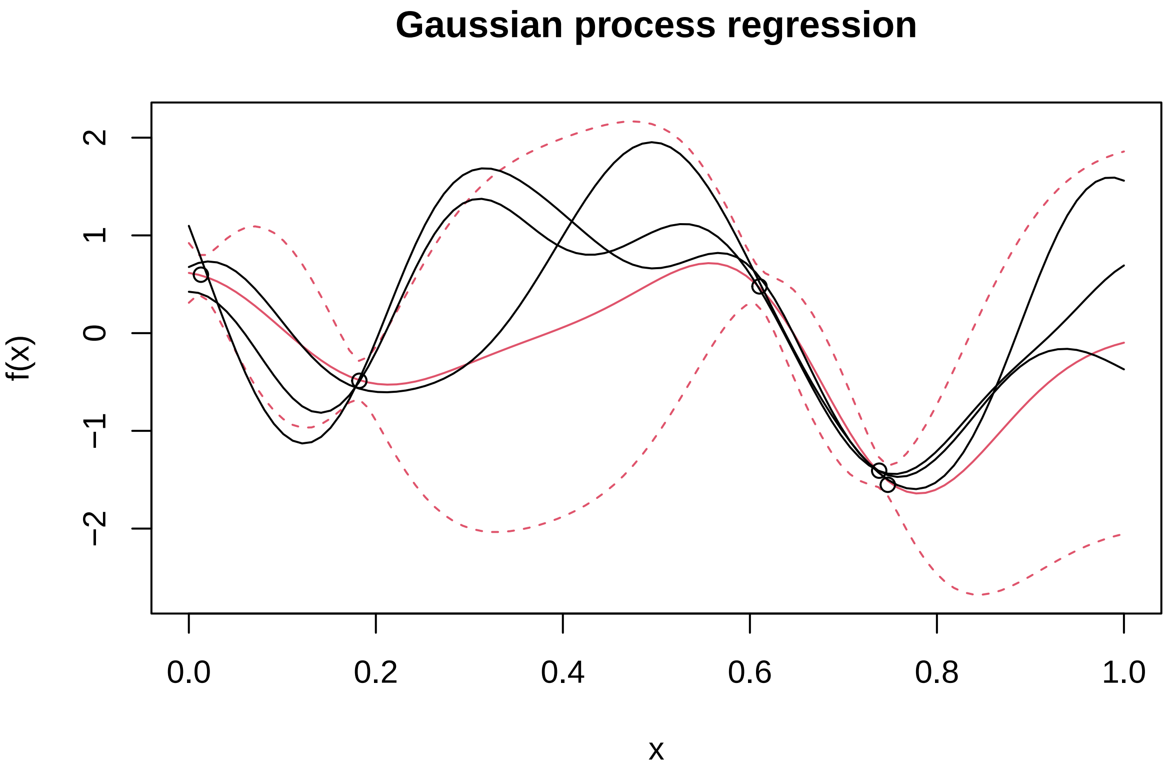
Gaussian process regression

A GP can be used as a functional prior. The posterior is then the conditional distribution of functions given observations at a set of input values. If the observations are assumed to include some Gaussian noise this is another GP, e.g.

$$
\begin{gathered}
\mathbf{y} = f(\mathbf{x}) + \epsilon, \\
f(\mathbf{x}) \sim N(\mu(\mathbf{x}) = \mathbf{0}, \Sigma = K(\mathbf{x}, \mathbf{x})), \\
\epsilon \sim N(0, \sigma_n^2 I_d), \\
\implies f(\mathbf{x}')|\mathbf{y}(\mathbf{x}) \sim N(\mu(\mathbf{x}'), \Sigma), \\
\mu(\mathbf{x}') = K(\mathbf{x}', \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I_d]^{-1}\mathbf{y}(\mathbf{x}), \\
\Sigma = K(\mathbf{x}', \mathbf{x}') - K(\mathbf{x}', \mathbf{x})[K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I_d]^{-1} K(\mathbf{x}, \mathbf{x}').
\end{gathered}
$$

The posterior mean function can be seen to be a linear function, either of the observed values, or of the covariances of the input values of the predictions with those of each of the observations.

```
plot_GP_regression(
  n_obs = 5, # Number of points to observe
  l = 0.1, # Length scale
  sigma_f = 1, # Function standard deviation
  sigma_n = 0.1 # Noise standard deviation
)
```



Connection to Bayesian linear models with infinitely many basis functions

Placing a Gaussian prior on the weights of a Bayesian linear model that projects the inputs into a feature space with infinitely many basis functions $\phi(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$, defines a GP prior on the model function. Taking observed input values as column vectors we have

$$f(\mathbf{X}) = \phi(\mathbf{X})^T \mathbf{w},$$

$$\mathbf{w} \sim N(\mathbf{0}, \Sigma_p),$$

$$\implies f(\mathbf{X}) \sim N(\mathbf{0}, \phi(\mathbf{X})^T \Sigma_p \phi(\mathbf{X})).$$

For a finite set of basis functions the covariance matrix for a larger set of inputs would not have full rank, so the inverse would not exist to define a Gaussian distribution for them.

The covariance function multiplies the projection of the inputs in the feature space by the square root of the prior covariance matrix and takes the inner product.

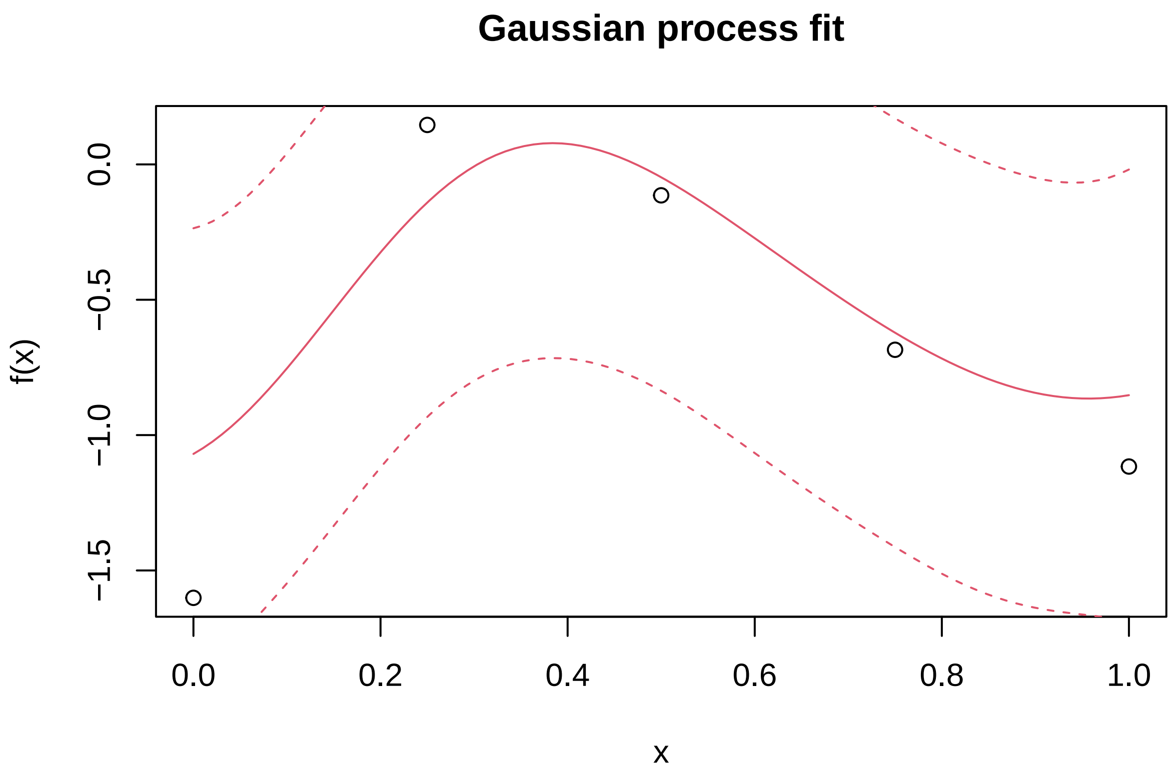
$$\phi(\mathbf{X})^T \Sigma_p \phi(\mathbf{X}') = \phi(\mathbf{X})^T L_p L_p^T \phi(\mathbf{X}') = (L_p^T \phi(\mathbf{X})).(L_p^T \phi(\mathbf{X}'))$$

Any infinite-dimensional inner product defines a covariance function, and any covariance function can be expressed as such an inner product.

Hyper-parameter tuning

There are various methods of model selection and hyper-parameter tuning. One is maximizing the posterior probability of the observations. The partial derivatives are known for some covariance functions making the numerical process tractable.

```r
plot_GP_fit(
  n_obs = 5, # Number of points to observe
  l = 0.1, # Length scale
  sigma_f = 1, # Function standard deviation
  sigma_n = 0.1 # Noise standard deviation
)
```



Notes taken mainly from Rasmussen & Williams, Gaussian Processes for Machine Learning, 2006.