CSC 2541: Bayesian Methods for Machine Learning

Radford M. Neal, University of Toronto, 2011

Lecture 7

From Linear Basis Function Models to Gaussian Processes

We saw previously that a linear basis function model with a Gaussian prior for the coefficients defines a Gaussian prior distribution for any set of observed or unobserved responses.

If we fix all means to zero, this Gaussian prior distribution is determined by the covariances between responses, which we saw could sometimes be computed even when the number of basis functions is infinite.

But why bother?

We can just *start* with a function that defines the covariances between responses. As long this function always produces positive-definite covariance matrices, we can use it to infer unobserved responses from observed responses.

[Actually, we might bother with the original basis functions approach if that was the simplest way of expressing our prior beliefs, but often the covariances themselves have more intuitive meaning.]

Defining a Gaussian Process Model

We'll model the response y_i associated with covariate vector x_i as being equal to some function, f, of x_i plus Gaussian noise: $y_i \;=\; f(x_i) + e_i$, with $e_i \sim N(0, \sigma^2)$.

We can define the "noise-free" covariances in terms of a function $K(x, x')$, as:

$$\text{Cov}(f(x_i), f(x_{i'})) \;=\; K(x_i, x_{i'})$$

I'll always assume that the prior means of all the y_i are zero, so this is enough to specify a multivariate Gaussian distribution for the value of the function at any set of x_i's.

Since this gives us a prior for $f(x)$ at an arbitrarily large set of x's, it effectively gives us a prior for the function f itself.

The covariances for actual observations will be

$$\text{Cov}(y_i, y_{i'}) \;=\; \sigma^2 \delta_{i,i'} \;+\; K(x_i, x_{i'})$$

The Covariance Function

We can choose the noise-free covariance function, $K(x, x')$, to be anything we want, **provided** that it produces positive definite (or at least positive semi-definite) covariance matrices for the y_i's with any allowed set of x_i's.

It's not easy to determine whether some arbitrary $K(x, x')$ will produce positive definite covariance matrices. But there are some well-known classes of valid covariance functions.

Furthermore, the *sum* of two valid covariance functions, K_1 and K_2, is also a valid covariance function. It can be seen as the covariance function for the sum of two functions, one drawn from the Gaussian process with covariance K_1 and the other drawn independently from the Gaussian process with covariance K_2.

The *product* of two covariance functions is also a valid covariance function (though this isn't so obvious).

Constant and Linear Covariance Functions

The *constant* covariance function:

$$K(x, x') \quad = \quad \sigma_0^2$$

can be derived from a model in which the function is an unknown constant: $f(x) = \mu$, with $\mu \sim N(0, \sigma_0^2)$.

The *linear* covariance function:

$$K(x, x') \quad = \quad \sigma_1^2 \, x \, x'$$

comes from a simple linear regression model, $f(x) \; = \; \beta x$ with $\beta \sim N(0, \sigma_1^2)$.

If we have two covariates, we can add linear covariances based on each, plus a constant covariance, to get

$$K(x, x') \quad = \quad \sigma_0^2 \, + \, \sigma_1^2 \, x_1 \, x_1' \, + \, \sigma_2^2 \, x_2 \, x_2'$$

(Note that subscrips on x here select covariates, not cases.)

Stationary Covariance Functions

A *stationary* covariance function can be written as $K(x, x') = K(x-x')$. It is translationally invariant, since only the difference $x - x'$ matters.

Typically, the covariance goes down with increasing distance of x and x'. One class of valid covariance functions of this form is

$$K(x, x') \;=\; \gamma^2 \exp(-\rho^2 ||x - x'||^r)$$

where $||x||$ is the Euclidean norm. This is valid for any $r \in (0, 2]$. It's clearly rotationally symmetric.

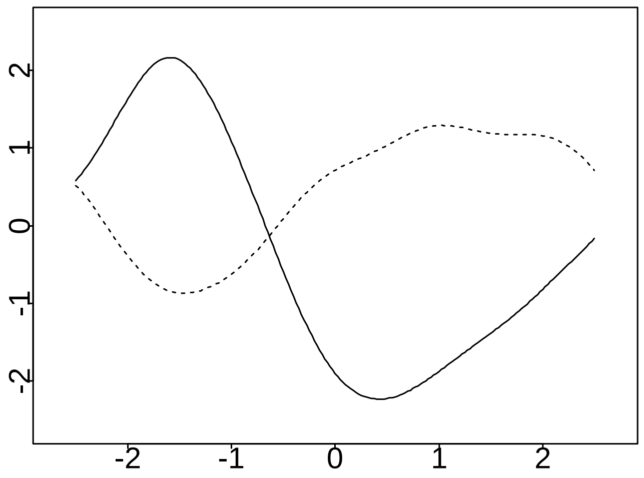
One can compare the above to the following:

$$K(x, x') \;=\; \gamma^2 \exp\Big(-\rho^2 \sum_{j=1}^{p} |x_j - x_j'|^r\Big)$$

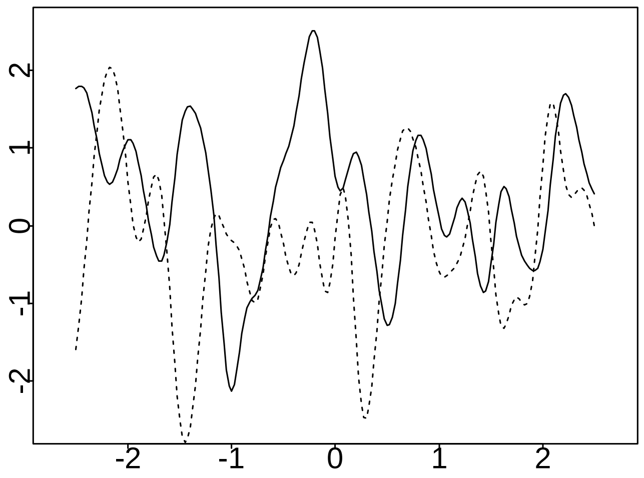
They're the same when $p = 1$ or $r = 2$, but they otherwise are different.

We can see that the second form is valid since it's a product of covariance functions of the first form.

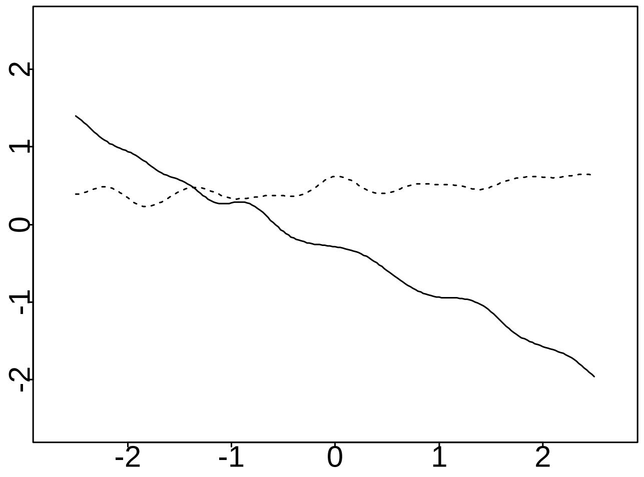
Functions From Some Gaussian Process Priors



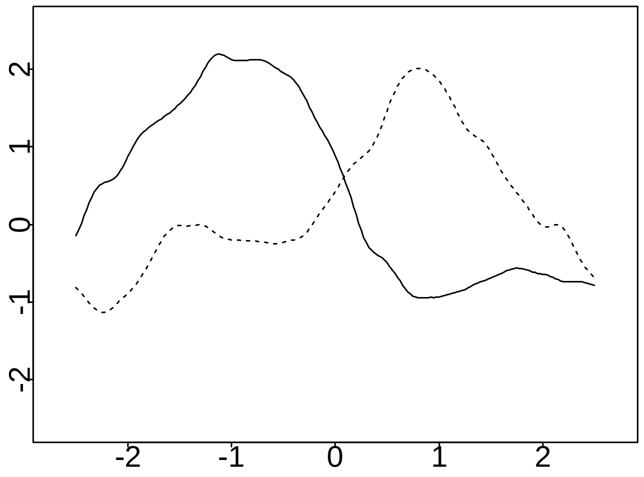
$\exp(-(x-x')^2)$



$\exp(-5^2(x-x')^2)$



$$1 + xx' + 0.1^2 \exp(-3^2(x-x')^2)$$



$$\exp(-(x-x')^2) + 0.1^2 \exp(-5^2(x-x')^2)$$

Predictions with Gaussian Process Models

If we know the covariance function, and the noise variance, predictions for test cases with a Gaussian process model can be done with straightforward matrix operations.

As we saw before, if y is the vector of responses in the n training cases, and y_* is the response for a test case, the conditional distribution of y_* given y will be Gaussian, with mean and variance given by

$$E(y_* \,|\, y_1, \ldots, y_n) \;=\; k^T C^{-1} y, \qquad \text{Var}(y_* \,|\, y_1, \ldots, y_n) \;=\; v \;-\; k^T C^{-1} k$$

Here C is the $n \times n$ covariance matrix of the responses, in the training set, k is the vector of covariances of the response for the test case with the responses for training cases, and v is the variance of the test response (covariance of y_* with itself).

If $K(x, x')$ takes $O(p)$ time to compute, then C^{-1} will take $O(pn^2 + n^3)$ time to compute, after which a prediction for each test case takes $O(pn)$ time for just the mean, plus $O(n^2)$ time if the variance is also required.

When the Covariance Function Isn't Known

In practice, the covariance function usually has some unknown parameters — such as the scale parameters γ and ρ in the exponential covariance function. The noise variance is also typically not known.

The covariance matrix of responses, C, needed for prediction, will depend on these unknown parameters.

One could find the maximum likeihood estimates for the unknown parameters, and then use these single values for prediction.

The full Bayesian approach is to average predictions over the posterior distribution for the unknown parameters, probably using MCMC.